FALSIFYVC: ERROR-CONTROLLED RELIABILITY TESTING OF VIRTUAL CELL MODELS ACROSS COMPARATOR AUTHORITIES

Anonymous authors
Paper under double-blind review

ABSTRACT

Virtual cell predictors are judged by held-out accuracy on unseen perturbations. Their added value depends on the comparator's selection information and rule, which existing reliability tests take as given. FalsifyVC indexes verdicts by this comparator authority, separating the reliability threshold, margin and run budget. Exact binomial tests count training runs whose excess over the comparator clears a registered margin. On two Replogle screens, no GEARS run clears the margin against any comparator across twelve fresh target splits with its reference preprocessing, or against fixed pseudobulk and a held-out per-target envelope on the registered split. STATE and CPA are refuted outside their supported regime, STATE also with training-data gene features. A training-only selector yields positive GEARS mean excess on the registered split but is weaker than training-selected ridge, and the excess vanishes on fresh splits. On an independent screen GEARS clears the margin in none of thirty runs, on one split or across thirty. A registered positive control clears the margin in all thirty runs and is supported, then refuted once an additive baseline is registered. The results refute the specified added-value claims under these conditions and show that the rule can support a large advantage over registered comparators.

1 INTRODUCTION

Virtual cell models predict transcriptomic responses to unseen genetic perturbations (Callahan et al., 2025). Prominent predictors include GEARS (Roohani et al., 2024), STATE (Adduri et al., 2026) and CPA (Lotfollahi et al., 2023). Genome-scale Perturb-seq screens provide held-out responses against which these predictions can be tested (Replogle et al., 2022). The practical question is not accuracy alone, but improvement over a comparator that could be used in the same setting.

Evidence for that improvement remains contested. Deep predictors may fail against linear or simple control baselines yet beat uninformative ones under other metrics (Ahlmann-Eltze et al., 2025; Csendes et al., 2025; Wong et al., 2025; Miller et al., 2025). Systematic variation and a control reference shared by predictions and truth can also inflate performance (Viñas Torné et al., 2026; Nicol et al., 2026). Large benchmarks broaden model and dataset coverage (Wu et al., 2025; Wei et al., 2026; Mao et al., 2026). Recent evaluation work examines metric calibration and baseline choice (Vollenweider & Bühlmann, 2026; Schäfer et al., 2026).

However, the run-level tools that quantify such evidence take the comparator as given. Probability-of-improvement estimands and run-budget analyses address variation across repeated runs (Agarwal et al., 2021; Bouthillier et al., 2021; Dror et al., 2019), and tests for data snooping and superior predictive ability adjust for the search among the predictors being compared (White, 2000; Hansen, 2005). The test alone does not specify the information the comparator's selection could use or the rule it followed, although weak or post-hoc baselines are a known source of overoptimism (Ferrari Dacrema et al., 2019; Leech et al., 2024; McGreivy & Hakim, 2024). The same predictions can earn different verdicts under different comparator rules without the report showing why.

To address this limitation, we propose FalsifyVC, a decision layer indexing run-level verdicts by comparator authority, the selection information and rule (Figure 1). The reliability threshold, margin and run budget remain separate parameters. It counts complete training runs whose worst-

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.

environment added value clears a registered margin and applies exact binomial tests to return supported, refuted or inconclusive. Frozen predictions permit comparing a fixed comparator, a training-only selector and a per-target envelope chosen with held-out truth without retraining. This specifies comparators rather than adding a new test.

Our contributions are as follows:

- We propose a comparator-indexed run-level decision procedure that separates information access, selection rule and decision parameters. Error control is exact for fixed scoring choices, with a union-bound extension for refutations after the one outcome-exposed choice. An exact design analysis gives the run budget at which support becomes attainable, and a registered positive control clears the margin in every run and is supported until an explanatory baseline is registered (Sections 3, 4 and 6.5, Appendices A and R).
- We run a registered campaign of sixty training runs and follow-up tests registered before scoring. GEARS, the one model given unseen-target representations in that campaign, is refuted against every comparator on twelve fresh splits of each Replogle screen with its reference preprocessing and on an independent screen. On the registered raw-count split it is refuted against the fixed comparator, although on K562 a target-resampled interval includes zero. STATE and CPA are refuted outside their supported regime, STATE also inside it when training-data gene features encode held-out targets (Sections 6.1 and 6.4, Appendices L, N, O and P).
- We trace the campaign's only positive mean excess, GEARS against a training-only selector on both screens, to a selector that scores below the fixed comparator. A stronger selector with the same features does not materially outscore training-selected ridge. The excess is fragile. Only six of ten RPE1 runs clear the margin, support requires a secondary threshold before family correction, target-resampled intervals include zero, and fresh-split excess with the reference preprocessing is negative on both screens (Sections 6.2 and 6.3, Appendices G, I and L).

2 Related work

FalsifyVC uses the probability of improvement of Agarwal et al. (2021), the reliability threshold and run budget of Bouthillier et al. (2021), exact one-sided tests, and ordered error control for nested families (Goeman & Solari, 2010). Data-snooping and superior-predictive-ability tests adjust a comparison for the search over candidate models (White, 2000; Hansen, 2005). Discriminant-validity tests instead compare loss on intended and unintended outcomes (Coston, 2026). Algorithm selection places a per-instance selector between the single best solver, chosen on training instances, and a per-instance oracle, the virtual best solver (Rice, 1976; Xu et al., 2008; Bischl et al., 2016; Kerschke et al., 2019). Domain generalization reports each method with the data its model selection may use, up to a test-domain oracle (Gulrajani & Lopez-Paz, 2021). Tuning budgets determine many comparisons (Dodge et al., 2019; Melis et al., 2018; Lucic et al., 2018). A_1, A_2 and A_3 occupy these positions for the comparator, not the method under test. Most of the envelope's advantage reflects selection bias addressed by nested evaluation (Cawley & Talbot, 2010). Two one-sided tests and regions of practical equivalence already make margin-based decisions with an inconclusive outcome (Schuirmann, 1987; Lakens, 2017; Benavoli et al., 2017). Baseline construction, baseline strength and evaluation choices are studied by Viñas Torné et al. (2026), Vollenweider & Bühlmann (2026) and Schäfer et al. (2026) (Appendix V).

3 Comparator authority

Fix a predictor f, environments $\mathcal{E}$, targets $\mathcal{T}$, and a score s computed against held-out truth (Section 4). A *comparator* is a rule c that supplies a predicted response for every target t and environment e. The added value of f over c in e is

$$V(f, c, e) \;=\; s\big(f, e\big) \;-\; s\big(c, e\big). \tag{1}$$

Equation 1 is standard, but the comparator c is selected. Its selection information and rule belong in the verdict specification alongside the reliability threshold, margin, run budget and multiplicity correction (Section 4).

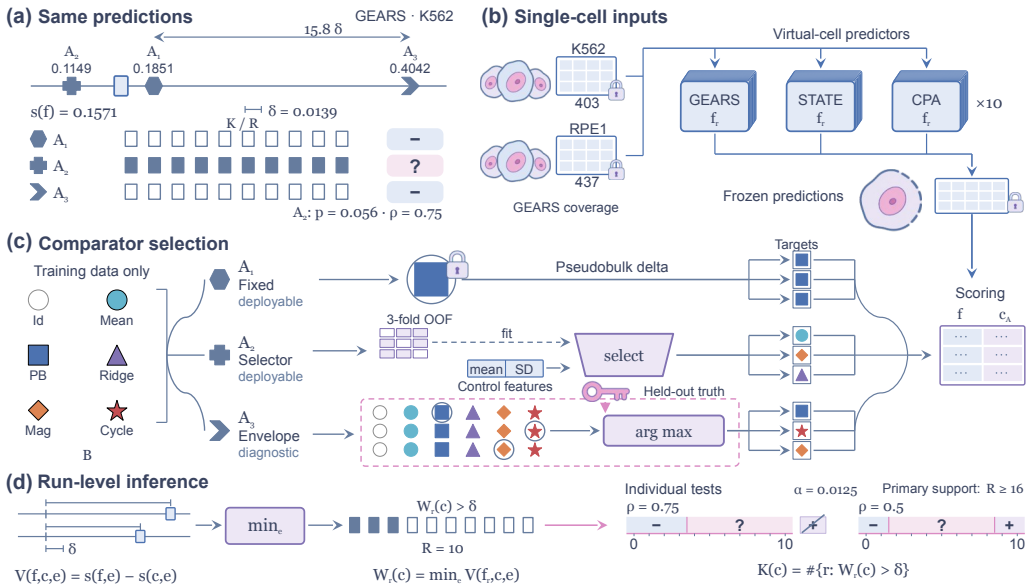


Figure 1: Comparator authority and run-level evidence. **(a)** The same frozen GEARS predictions on K562 in raw counts receive the primary verdicts refuted, inconclusive and refuted under A_1, A_2 and A_3. A_2 scores below A_1, so its inconclusive verdict reflects a weaker comparator. Ten successes in ten runs against A_2 still leave the primary test inconclusive at $p = 0.056$. The score axis and δ bracket are to scale. **(b)** Inputs, predictors and frozen predictions, with coverage counts for GEARS. Cell illustrations are schematic. **(c)** Three unordered authorities select from one family $\mathcal{B}$. A_1 and A_2 use training data only and differ in selection rule. Only A_3 uses held-out truth for selection. **(d)** Runs with $W_r(c) > \delta$ contribute to K. Primary support is unreachable at $R = 10$, and the $\rho = 0.5$ row is secondary. Panels **(c)** and **(d)** are schematic rather than further results.

Definition 1 (Comparator authority)**.** The *authority* of a comparator is the pair formed by the information its *selection* may use and the rule that selects it. We study three.

A_1**, fixed.** One member is chosen from training data alone before any test data are seen and is used unchanged for every target. The registered instance is pseudobulk-delta, named in advance. A post-hoc training-selected instance takes the member with the best out-of-fold training score and serves as the strong deployable control (Section 6.3).

A_2**, cross-fit selector.** A nearest-neighbor rule on training-control features of the target gene chooses a member per target, with labels from a cross-fit over training targets. A_2 uses the same information as A_1 and differs only in selection rule, so weak A_2 performance is a selector weakness, not an information constraint.

A_3**, oracle-max.** The comparator is chosen per target as the member of a family $\mathcal{B}$ that scores best on that target's held-out truth,

$$c_{A_3}(t, e) \;=\; \arg\max_{b \in \mathcal{B}} \; s\big(b, t, e\big). \tag{2}$$

Selection in Equation 2 requires held-out truth, so this rule is not deployable on an unseen response.

A_3 differs from A_1 in both information and selection rule. A post-hoc crossing attributes at least 95% of their gap to per-target selection with held-out truth. A single member chosen with held-out truth is the training-selected one on every screen (Appendix I). Appendix U explains why the authority must accompany a verdict.

Decision boundaries. $s(c_{A_3}, e) \geq s(c_{A_1}, e)$ holds by construction, since A_3 maximizes over a family that contains A_1's member, so $V(f, c_{A_3}, e) \leq V(f, c_{A_1}, e)$ and a refutation under A_1 implies

one under A_3. Their verdicts can differ only for a run whose excess over A_1 lies between δ and δ plus the score gap. Because A_3 is a maximum, such a difference may reflect selection, not information (Appendix J).

4 COMPARATOR-INDEXED RELIABILITY

4.1 SETUP AND NOTATION

A perturbation predictor f maps control cells and a perturbation label p in an environment e (e.g. a cell line) to a predicted post-perturbation profile. For any predictor g we summarize its effect by the predicted mean shift $\widehat{\Delta}_g(p,e) = \overline{g}(p,e) - \overline{C}_e \in \mathbb{R}^G$ ($\overline{C}_e$ the control mean, G genes). The held-out truth is $\Delta(p,e) = \overline{X}(p,e) - \overline{C}_e$ from Perturb-seq. The *skill* of g is the delta correlation

$$\sigma(g;p,e) = \mathrm{corr}\big(\widehat{\Delta}_g(p,e),\, \Delta(p,e)\big), \tag{3}$$

and the score $s(g,e)$ averages $\sigma(g;p,e)$ over the scored targets. We fix a *comparator family* $\mathcal{B}$ using only observational or marginal structure. Its six members were frozen before any confirmatory run. They are identity, mean-shift, pseudobulk-delta, ridge, and two shortcut predictors, one proportional to baseline expression and one supported on the S-phase and G2/M marker genes of Tirosh et al. (2016) (Appendix B). On an unseen target, mean-shift, pseudobulk-delta and ridge each return an average training effect and differ only in its weighting. Identity predicts the control profile, so its shift is exactly zero and its correlation is scored as zero (Appendix C).

The family is *registered*, not exhaustive. A comparator absent from $\mathcal{B}$ is not tested, so a verdict of supported means only that the model survives the comparators registered in advance. Under the per-target maximum, enlarging $\mathcal{B}$ can revoke such a verdict but never create one (Proposition 2).

The three authorities of Section 3 instantiate c from this family. The registered c_{A_1} is pseudobulk-delta, c_{A_2} is chosen per target by a rule fitted out of fold on training data alone, and $\sigma(c_{A_3};p,e) = \max_{b\in\mathcal{B}} \sigma(b;p,e)$, so the excess against the family maximum is the special case $V(f,c_{A_3},e)$.

Scores are averaged over targets rather than computed once on a pooled vector, whose seed-to-seed standard deviation is 24 times larger on the same predictions (Appendix D).

4.2 RUN-LEVEL STATISTIC

A run is one complete training run of a predictor. We test a property of that run, not an averaged score. For run r and comparator c,

$$W_r(c) \;=\; \min_{e\in\mathcal{E}}\; V\big(f_r,c,e\big), \tag{4}$$

the added value in the environment where the run does worst, so a run counts only if it is ahead in every environment. When environments are trained separately, a run pairs one run per environment. If outcomes are independent, the pair's success probability is the product of per-environment success probabilities.

Monotonicity of nested envelopes. Under a per-target maximum, W_r is non-increasing as the family grows, but a comparator selected on one fold and evaluated on another need not be (Appendix A), so the three authorities form an unordered axis rather than a ladder.

Run-level decisions. With fixed scoring choices and margin δ, count $K(c) = \#\{r : W_r(c) > \delta\}$ over R independent complete training runs. For $q(c) = \Pr\{W_r(c) > \delta\}$, support rejects the null $q(c) \le \rho$ using the upper binomial tail at ρ, and refutation rejects $q(c) \ge \rho$ using the lower tail. A fixed nested maximum can use ordered testing under the assumptions of Goeman & Solari (2010), but the tested authorities are unordered and have no such exemption. We report registered individual tests separately from a post-hoc Holm analysis. Exact binomial error control requires independent run outcomes and fixed scoring choices, and it does not cover choices made after viewing test outcomes (Appendix T). Proposition 1 states these conditions.

5 EXPERIMENTAL PROTOCOL AND DIAGNOSTICS

The datasets, models, comparator family, thresholds and decision rule were registered before the confirmatory runs, and later tests before scoring. Preliminary runs had already used the confirmatory test split. The per-model expression scale of Section 5.2 was fixed after 17 of the sixty training runs had completed, and A_2 was interpreted after its result. The scale, chosen between two versions of each dataset, is the only primary-verdict scoring choice made after test outcomes. Every refutation has $K = 0$ of 10, so by a union bound each keeps its level α however the scale was chosen (Corollary 1). Table T.1 lists what was known when each choice was fixed, and numerical results are derived from the retained predictions under the registered splits.

5.1 DATASETS AND ENVIRONMENTS

The confirmatory claim rests on two environments, the K562 and RPE1 essential-gene Perturb-seq screens of Replogle et al. (2022), obtained from scPerturb (Peidli et al., 2024). Both use the same processing pipeline and perturbation-level split procedure, so all cells of a held-out target are held out together.

K562 retains 2021 of 2057 targets (1617 train, 404 test) on 8563 genes and RPE1 retains 2204 of 2393 (1763 / 441) on 8749, after dropping targets with fewer than twenty cells. Each is scored on its own gene panel. Rebuilding either split from source reproduces the registered partition exactly (Table C.1).

Norman's genetic-interaction screen (Norman et al., 2019) was excluded from the confirmatory claim by a registered rule before any result was seen. The K562-versus-RPE1 contrast breaks expression-prior shortcuts and partly the cell-cycle axis but preserves library and guide-frequency artifacts, and no within-dataset batch contrast was run.

5.2 MODELS UNDER TEST AND REGISTERED BASELINE CLASS

The models under test are GEARS (Roohani et al., 2024), STATE (Adduri et al., 2026) and CPA (Lotfollahi et al., 2023). Each is scored in its output scale, raw counts for GEARS and log-transformed expression for STATE and CPA. CPA is transductive, so on these 0%-overlap splits its out-of-distribution perturbation embeddings are untrained. STATE encodes a genetic perturbation by a one-hot index unless gene features are supplied. None were supplied, so untrained weights also encode its held-out targets. We keep both for architectural diversity, treat their refutations as observations outside the supported regime, and do not rest the negative conclusion on them (Appendix B). GEARS predicts only perturbations in its gene graph, 403 of 404 K562 test targets and 437 of 441 on RPE1. Each comparison scores the model and every comparator on the same covered targets. A robustness arm fills uncovered targets with the control profile. The earlier study's cluster-level analyses and frozen scGPT head (Cui et al., 2024) are summarized in Appendix W.

5.3 PASS/FAIL CRITERIA

Each model has $R = 10$ complete runs per environment. The registered primary threshold is $\rho = 0.75$, attributed to Bouthillier et al. (2021), and $\rho = 0.5$ is secondary. The excess threshold $\delta = 0.0139$ is GEARS's seed-to-seed standard deviation on a K562 training-target development split, which shares no target with the confirmatory test split (Appendix D). Its range, $\delta = 0.0346$, is registered for sensitivity. Neither is a universal biological minimum effect. A specified claim is the registered hypothesis that a model's runs clear the margin with probability above ρ, not a claim made by the model's authors. For $K = \#\{r : W_r > \delta\}$ and $X \sim \mathrm{Bin}(R, \rho)$, support uses $\Pr(X \geq K)$ and refutation uses $\Pr(X \leq K)$, each at the individual level 0.0125. Ties do not count as successes.

At primary $\rho = 0.75$, $R = 10$ can refute at $K \leq 3$ but cannot support even $K = 10$. At secondary $\rho = 0.5$, support requires $K \geq 9$ and refutation $K \leq 1$. Appendix G applies the registered primary threshold to an analysis that first used only the secondary threshold.

Single-environment counts use W_r over one environment, which is $V(f_r, c, e)$. Joint counts use the minimum over the same ten seed-index pairs, with incomplete pairs reported as not assessed. In every case one screen has $K = 0$ or $K = R$, so each joint count is the same under any one-to-

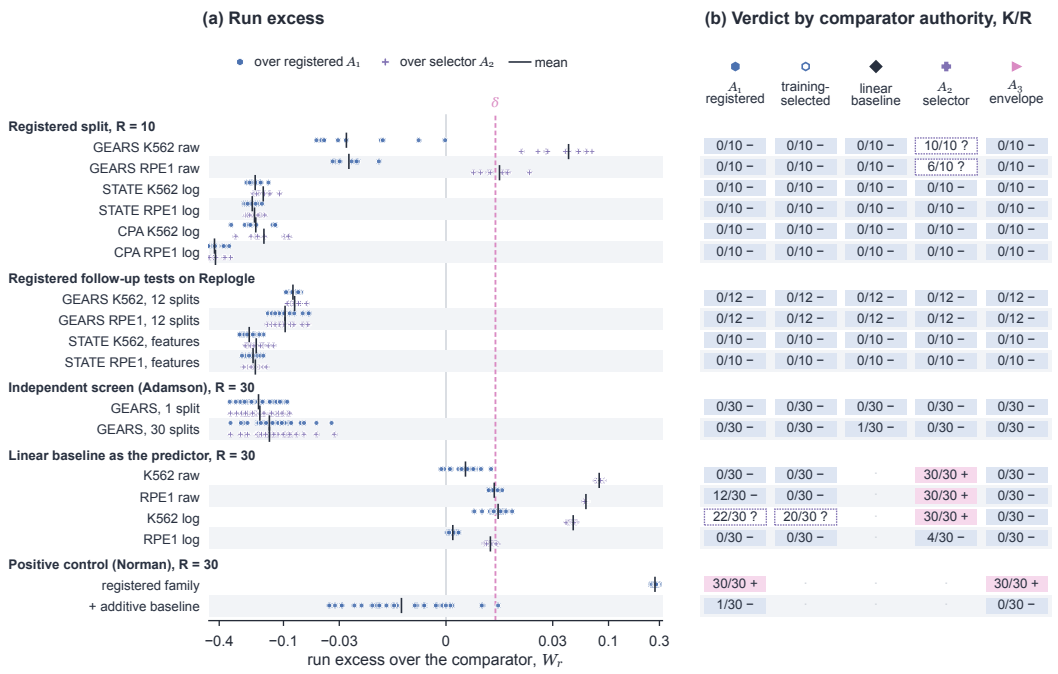


Figure 2: Main comparisons and follow-up run-level tests. In these comparisons, deep predictors are refuted against every comparator at least as strong as the fixed one. Initial STATE and CPA tests run outside their supported regime. Support occurs for the positive control and for the linear baseline against the weak selector. **(a)** Excess W_r of each run over the registered fixed comparator A_1 (hexagons) and the training-only selector A_2 (plus signs), with run means as bars and $\delta = 0.0139$ dashed, on a symmetric-log axis that is linear within ± 0.03. **(b)** Verdict under each comparator authority, with K/R the covered-arm runs clearing δ out of R and $+$ supported, $-$ refuted or ? inconclusive from both coverage arms at $\rho = 0.75$ and individual $\alpha = 0.0125$. The second positive-control row replaces A_1 by the additive baseline and adds it to the envelope. Appendices F and Q list the values.

one pairing of the two screens' runs (Appendix L). A post-hoc Holm analysis (Holm, 1979) at 0.05 spans 432 directional hypotheses for every combination of model, authority, environment endpoint, coverage arm, reliability threshold, excess threshold and direction. It excludes later controls.

Each model–screen verdict is computed twice, once on the targets the model can address and once on the complete frozen split with the remainder filled by the control profile. A comparison whose two coverage arms disagree is reported as inconclusive with both verdicts stated, and the more favorable arm is never adopted (Appendix C).

6 Results

6.1 Fixed-comparator and envelope results

The envelope inherits the fixed comparator's refutations. Figure 2 summarizes the main comparisons and follow-ups under each authority. All six registered cells give $K = 0$ of 10 under both A_1 and A_3, as do joint counts pairing seed IDs across screens (Appendix H). All six are primary refutations at $p = 0.25^{10} = 9.54 \times 10^{-7}$, the value of any zero count at $R = 10$. Because the fixed comparator belongs to the family, a refutation under A_1 implies one under A_3 (Proposition 2). Only a run with $\delta < V(f_r, c_{A_1}, e) \leq \delta + [s(c_{A_3}, e) - s(c_{A_1}, e)]$ could change verdict between them, and no evaluated run clears δ against A_1. GEARS comes closest, with mean excess -0.0280 on K562 and -0.0273 on RPE1 and a best K562 run at -2.43×10^{-4}.

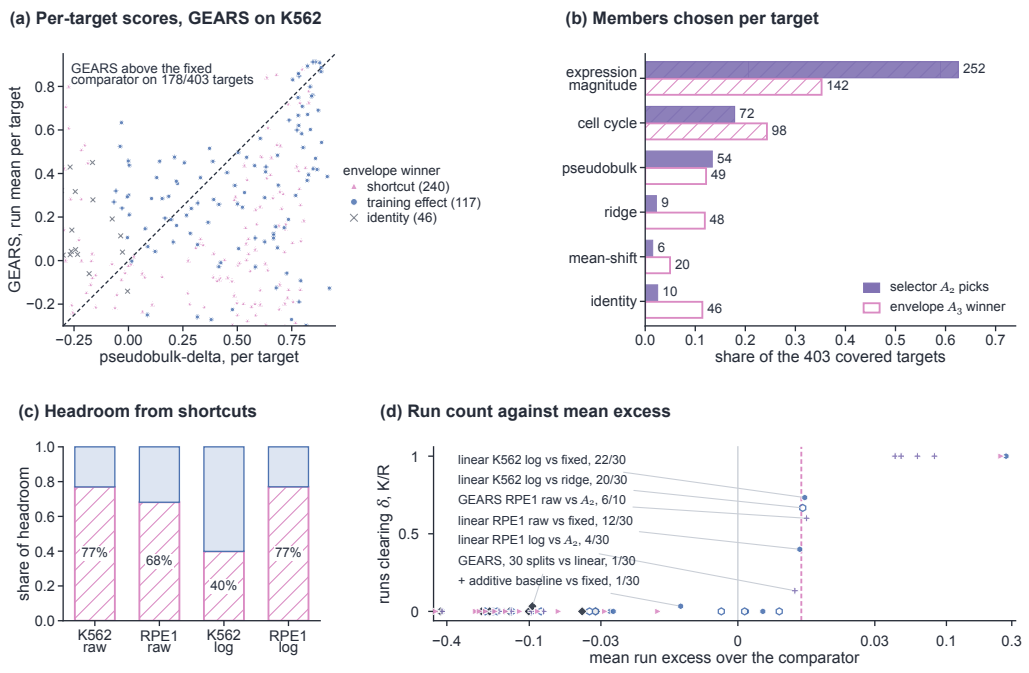


Figure 3: Sources of comparator strength. Per-target selection with held-out truth favors two shortcut members, and the training-only selector favors the same two. **(a)** Per-target score of GEARS, averaged over ten runs, against the registered fixed comparator on the 403 covered K562 targets in raw counts, marked by the member that wins each target in the envelope. **(b)** Targets assigned to each member by the selector A_2 (filled) and won by it in the envelope A_3 (outlined), with the shortcut members hatched. **(c)** Share of the envelope's gain over A_1 that comes from targets a shortcut member wins, by screen and expression scale. **(d)** Fraction of runs clearing δ against the mean run excess for every test and authority in Figure 2, with its glyphs and δ dashed. Near δ the mean does not determine the count.

Run-level stability, target resampling and coverage. The run-to-run standard deviation is below δ only for GEARS on RPE1, at 4.6×10^{-3}, and ranges from 0.017 to 0.085 in the other cells, yet no run in any cell clears δ against A_1. Resampling targets and runs leaves every A_1 and A_3 mean-excess interval below zero except GEARS/K562 against A_1, $[-0.099, +0.043]$. That raw-count refutation holds for this split, not the screen (Appendix L), whereas GEARS on its reference preprocessing is refuted on twelve fresh splits (Section 6.4). Referencing truth and every prediction to disjoint, batch-stratified control-cell halves removes the shared reference that Nicol et al. (2026) argue inflates the metric. The run-level rule retains 178 of 180 refutations over five partitions, three comparators and both arms. The two exceptions are GEARS/K562 against A_1 in one partition, with $K = 5$ and 6 of 10, where training-selected ridge still refutes (Appendix K).

6.2 Mean excess and run-level reliability

The registered A_2 selector (Appendix E) is scored for all six model–screen cells, and it is the only comparator against which any model has a positive mean excess. GEARS exceeds A_2 by $+0.0422$ on K562 and $+0.0150$ on RPE1. Paired intervals over the ten runs, $[+0.0327, +0.0523]$ and $[+0.0123, +0.0177]$, exclude zero on both screens (Table I.3), so a mean comparison against zero would report an advantage on both. Intervals that also resample targets, $[-0.036, +0.120]$ and $[-0.002, +0.033]$, include zero (Appendix L). Mean excess and success count describe different properties of the run distribution (Figure 3(d)). On RPE1 the mean exceeds δ although its interval contains δ. Only six of ten runs succeed. The exact 95% interval for q is 0.26 to 0.88, and the verdict is inconclusive at both reliability thresholds (Appendix G). On K562 all ten runs clear δ, but at the primary $\rho = 0.75$ the support tail is $0.75^{10} = 0.056$, so the run budget rather than any failed run

leaves the verdict inconclusive. The success probability is the relevant quantity when one complete run is trained and deployed without selection among runs, because the deployed run must then clear the margin on its own.

Only at the secondary $\rho = 0.5$, before family correction, does the K562 tail reach 0.5^{10} and the verdict become supported. Among the 216 outcome coordinates of the post-hoc family only this setting is supported, in both coverage arms (Appendix G). Its Holm-adjusted tail is 0.324, and at $\delta = 0.0346$ the count falls to $6/10$, so neither threshold supports added value once either choice is varied. The selector assigns 80.4% of targets to the two shortcut members (Figure 3(b)), and its score, 0.1149, lies well below the fixed comparator at 0.1851. The secondary, uncorrected support therefore reflects a weaker comparator, not a deployable advantage for GEARS (Appendix H).

6.3 Stronger comparators and implementation checks

A deterministic-comparator probability of improvement (Agarwal et al., 2021) reproduces all 54 point estimates from the reference rliable implementation. An independent exact-binomial implementation matches all 216 decisions before and after Holm correction. The A_1/A_3 refutations survive Holm, and at $R = 10$ the secondary A_2 support cannot (Appendix G).

The training-selected instance of A_1 picks ridge by out-of-fold training score in all four screen and scale settings, and nested training-only selection between it and A_2 agrees. It beats A_2 throughout, by as little as 0.009623 on RPE1 in log space (Appendix I). Against ridge, a post-hoc authority outside the registered analysis, every model–screen comparison gives $K = 0/10$. The GEARS paired intervals are negative on both screens (Table I.3). A stronger selector, registered before its test scoring, keeps A_2's features and averages each member's out-of-fold scores over neighbors. It recovers 79% to 101% of A_2's test deficit to ridge but exceeds ridge by at most 6.6×10^{-5} on test targets and never in nested training evaluation. Nested training-only selection keeps ridge in all four settings, and every model remains refuted (Table I.2). The linear gene-embedding baseline of Ahlmann-Eltze et al. (2025), fitted post hoc on each registered training set, refutes every model at $K = 0/10$ as well, and GEARS scores 0.033 and 0.041 below it (Table Q.2).

6.4 Follow-up and independent-screen tests

A registered Replogle test gives GEARS its reference preprocessing and a fresh target split for each of twelve runs per screen, with GEARS and every comparator fitted on the same cells. No run clears the margin against any comparator on either screen, $K = 0/12$, with mean excess -0.082 and -0.097 over the training-selected member and the best run at -0.058 (Appendix O). A registered follow-up gives STATE training-data gene features, which place held-out targets inside its supported regime. The run-level test still refutes STATE against every comparator on both screens, $K = 0/10$, with mean excess -0.209 and -0.193 over A_1. Relative to the same-seed one-hot runs, the features change the score by -0.025 and $+0.003$ (Appendix P). On the unfolded-protein-response Perturb-seq screen of Adamson et al. (2016), control labels were rebuilt after a published correction (Wong et al., 2025). GEARS used the main campaign's training recipe on the preprocessed values distributed with GEARS (Roohani et al., 2024), so the result does not rest on Replogle's raw-count scale. The primary comparator is the training-selected instance of A_1, and at $R = 30$ support needs $K \geq 28$ and refutation $K \leq 16$. On one registered split, fixed before any test score, none of thirty GEARS runs clears the margin. Mean excess is -0.171 on 15 covered targets, with a best run at -0.092, yielding primary refutation at $p = 0.25^{30} = 8.67 \times 10^{-19}$ in both coverage arms. A second test, registered before its first run, draws a target split for each of thirty runs by the same procedure and refits every comparator on that split's training targets. Again no run clears the margin against any registered comparator in either arm. Mean excess against the primary comparator is -0.135, with a best run at -0.035. Against the linear baseline one run clears it, also a refutation, so the result holds across target draws and training seeds (Appendix N). On the registered split the authorities differ by at most 0.011 in score (Table N.1). Across all screens and authorities only the cross-fit selector, on the registered raw-count split, leaves any of the three models inconclusive, and none is supported (Figure 2(b)). The only verdict changes, GEARS against A_2 on both screens, trace to a comparator weaker than the fixed one. Reporting only the comparator family would not reveal this weakness. Here, authority indexing documents the weakness without changing decisions from strong baselines with fixed training-only selection.

6.5 POSITIVE CONTROL AND OPERATING CHARACTERISTICS

A run-level positive control, registered before its runs (Appendix R), uses the Norman genetic-interaction screen (Norman et al., 2019). A ridge head on random gene embeddings predicts a pair from its genes' learned marginal effects, with a new embedding in each of thirty runs. Every run clears the margin against both the registered fixed comparator and registered envelope, giving a supported verdict at $p = 1.79 \times 10^{-4}$. Counts fall to $1/30$ against a registered additive-identity baseline and $0/30$ against the enlarged envelope, both refuted. Treated as the predictor under test, with thirty bootstrap runs per registered dataset, the linear gene-embedding baseline is supported against the training-only selector in three of four settings, yet refuted against the fixed comparator, or inconclusive where its mean excess sits at δ (Appendix Q). Norman remains outside the confirmatory claim. The head's mean excess over the registered envelope, $+0.2514$, is eighteen times δ, so the control says nothing about power near the margin. A synthetic latent oracle is supported in only $80/256$ studies at $R = 30$ and none at $R = 10$ (Appendix S.1). Noise and stable-confounder predictors are refuted in all 24 combinations of predictor, screen, scale and authority (Appendix S). Boundary-null error rates remain unmeasured.

For the registered rule, support first becomes attainable at $R = 16$, with power 0.4401 at true success probability $q = 0.95$, and $R = 30$ is the smallest budget with power above 0.8 at $q = 0.95$ (0.8122) but gives only 0.4114 at $q = 0.9$ (Appendix A). These are design calculations, and the runs sample training randomness on fixed screens rather than new environments. Two shortcuts supply 40% to 77% of envelope headroom across screens and scales (Figure 3(c)). Choosing members on disjoint cells removes 81% of the envelope's advantage over A_1 on K562 in raw counts and all of it elsewhere, so the in-sample envelope is diagnostic (Appendix J).

7 LIMITATIONS

Three screens, two sharing the K562 line, do not establish generality. The independent screen tests only GEARS. The registered campaign has one target split per screen. For GEARS/K562 in raw counts against the fixed comparator, a target-resampled interval includes zero and one of five disjoint control references leaves the verdict inconclusive (Appendices K and L), although GEARS on its reference preprocessing is refuted on twelve fresh splits. CPA runs outside its supported regime, as does STATE except in one follow-up. GEARS is trained on raw counts in the registered campaign, and no model was tuned. Ten runs cannot return primary support. The margin δ comes from GEARS seed variation on a K562 development split and is carried over unchanged, not calibrated to biological utility (Appendix D). The linear gene-embedding baseline of Ahlmann-Eltze et al. (2025) is a post-hoc comparator on the registered splits. The post-hoc crossing of information access and selection rule attributes at least 95% of the envelope gap to per-target selection with held-out truth (Appendix I). The expression scale was chosen after test outcomes were seen, and only the refutations carry a guarantee adjusted for that choice (Section 5). The score is contested and may be inflated by a shared control reference (Miller et al., 2025; Nicol et al., 2026), and the GEARS mean excess against A_2 turns negative under Spearman (Appendix M). The positive control is a known additive effect rather than a biological finding, and the rule's boundary-null error on real expression, selector uncertainty and the power of ordered testing are unmeasured.

8 CONCLUSION

FalsifyVC specifies comparator information access and selection rule alongside separate run-level decision parameters. GEARS is refuted against every comparator on fresh splits of both Replogle screens with its reference preprocessing and on an independent screen, and against the fixed comparator on the registered raw-count split. STATE and CPA are refuted outside their supported regime and STATE also inside it. A registered positive control clears the margin in every run and is supported until an additive baseline is registered. Neither the registered selector nor the stronger one adds a deployable increment over training-selected ridge, and indexing changes no strong-comparator verdict here. A claimed advantage should therefore be reported with its comparator's authority and training-side strength.

STATEMENT ON THE USE OF AI

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.

REPRODUCIBILITY STATEMENT

Section 5 and Table C.1 describe the frozen splits and the rule that produces them, and Section 5.3 states the decision rule, thresholds, comparator family and inference unit. Deviations from the registration are described in Sections 5 and 6.2. One analysis scores the sixty retained prediction profiles, and an independent implementation reproduces its comparator, envelope and model scores to within 10^{-6} (Appendix K). Per-run results, both coverage arms and the sensitivity analyses are in Appendices F–S, and the earlier five-context study is a separate report in the supplementary archive, which also holds the analysis code and retained per-target outputs. The registration records identify the authors' compute environment and will be released after review with their timestamp proofs.

REFERENCES

Britt Adamson, Thomas M. Norman, Marco Jost, et al. 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.

Rishabh Agarwal, Max Schwarzer, Pablo Samuel Castro, Aaron C. Courville, and Marc G. Bellemare. Deep reinforcement learning at the edge of the statistical precipice. In *Advances in Neural Information Processing Systems*, volume 34, pp. 29304–29320, 2021.

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.

Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019. doi: 10.48550/arXiv.1907.02893. URL https://arxiv.org/abs/1907.02893.

Alessio Benavoli, Giorgio Corani, Janez Demšar, and Marco Zaffalon. Time for a change: a tutorial for comparing multiple classifiers through Bayesian analysis. *Journal of Machine Learning Research*, 18(77):1–36, 2017.

Bernd Bischl, Pascal Kerschke, Lars Kotthoff, et al. ASlib: A benchmark library for algorithm selection. *Artificial Intelligence*, 237:41–58, 2016. doi: 10.1016/j.artint.2016.04.003.

Xavier Bouthillier, Pierre Delaunay, Mirko Bronzi, et al. Accounting for variance in machine learning benchmarks. In *Proceedings of Machine Learning and Systems*, volume 3, pp. 747–769, 2021.

Tiffany J. Callahan, Zane Beckwith, Thomas Merth, Constantijn van der Poel, Adam Lewis, and Pablo Lemos. Virtual cells as causal world models: A perspective on evaluation. In *NeurIPS 2025 Workshop on AI Virtual Cells and Instruments: A New Era in Drug Discovery and Development (AI4D3 2025)*, 2025. URL https://ai4d3.github.io/2025/papers/25_Virtual_Cells_as_Causal_Wor.pdf.

Gavin C. Cawley and Nicola L. C. Talbot. On over-fitting in model selection and subsequent selection bias in performance evaluation. *Journal of Machine Learning Research*, 11(70):2079–2107, 2010.

Cédric Colas, Olivier Sigaud, and Pierre-Yves Oudeyer. How many random seeds? statistical power analysis in deep reinforcement learning experiments. *arXiv preprint arXiv:1806.08295*, 2018. doi: 10.48550/arXiv.1806.08295. URL https://arxiv.org/abs/1806.08295.

Amanda Lee Coston. Falsifying discriminant validity of predictive algorithms. In *Proceedings of the 2026 ACM Conference on Fairness, Accountability, and Transparency*, pp. 3105–3128, 2026. doi: 10.1145/3805689.3812310.

Rafael M. O. Cruz, Robert Sabourin, and George D. C. Cavalcanti. Dynamic classifier selection: Recent advances and perspectives. *Information Fusion*, 41:195–216, 2018. doi: 10.1016/j.inffus.2017.09.010.

Gerold Csendes, Gema Sanz, Kristóf Z. Szalay, and Bence Szalai. Benchmarking foundation cell models for post-perturbation RNA-seq prediction. *BMC Genomics*, 26(1):393, 2025. doi: 10.1186/s12864-025-11600-2.

Haotian Cui, Chloe Wang, Hassaan Maan, Kuan Pang, Fengning Luo, Nan Duan, and Bo Wang. scGPT: toward building a foundation model for single-cell multi-omics using generative AI. *Nature Methods*, 21(8):1470–1480, 2024. doi: 10.1038/s41592-024-02201-0.

Janez Demšar. Statistical comparisons of classifiers over multiple data sets. *Journal of Machine Learning Research*, 7(1):1–30, 2006.

Jesse Dodge, Suchin Gururangan, Dallas Card, Roy Schwartz, and Noah A. Smith. Show your work: Improved reporting of experimental results. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 2185–2194, 2019. doi: 10.18653/v1/D19-1224.

Rotem Dror, Segev Shlomov, and Roi Reichart. Deep dominance - how to properly compare deep neural models. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 2773–2785, 2019. doi: 10.18653/v1/P19-1266.

Maurizio Ferrari Dacrema, Paolo Cremonesi, and Dietmar Jannach. Are we really making much progress? A worrying analysis of recent neural recommendation approaches. In *Proceedings of the 13th ACM Conference on Recommender Systems*, pp. 101–109, 2019. doi: 10.1145/3298689.3347058.

Raffaella Giacomini and Halbert White. Tests of conditional predictive ability. *Econometrica*, 74 (6):1545–1578, 2006. doi: 10.1111/j.1468-0262.2006.00718.x.

Jelle J. Goeman and Aldo Solari. The sequential rejection principle of familywise error control. *The Annals of Statistics*, 38(6):3782–3810, 2010. doi: 10.1214/10-AOS829.

Ishaan Gulrajani and David Lopez-Paz. In search of lost domain generalization. In *International Conference on Learning Representations*, 2021.

Peter Reinhard Hansen. A test for superior predictive ability. *Journal of Business & Economic Statistics*, 23(4):365–380, 2005. doi: 10.1198/073500105000000063.

Sture Holm. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2):65–70, 1979. URL https://www.jstor.org/stable/4615733.

Pascal Kerschke, Holger H. Hoos, Frank Neumann, and Heike Trautmann. Automated algorithm selection: Survey and perspectives. *Evolutionary Computation*, 27(1):3–45, 2019. doi: 10.1162/evco_a_00242.

Daniël Lakens. Equivalence tests: A practical primer for t tests, correlations, and meta-analyses. *Social Psychological and Personality Science*, 8(4):355–362, 2017. doi: 10.1177/1948550617697177.

Gavin Leech, Juan J. Vazquez, Niclas Kupper, Misha Yagudin, and Laurence Aitchison. Questionable practices in machine learning. *arXiv preprint arXiv:2407.12220*, 2024. doi: 10.48550/arXiv.2407.12220. URL https://arxiv.org/abs/2407.12220.

Marius Lindauer, Jan N. van Rijn, and Lars Kotthoff. The algorithm selection competitions 2015 and 2017. *Artificial Intelligence*, 272:86–100, 2019. doi: 10.1016/j.artint.2018.10.004.

Marc Lipsitch, Eric Tchetgen Tchetgen, and Ted Cohen. Negative controls: a tool for detecting confounding and bias in observational studies. *Epidemiology*, 21(3):383–388, 2010. doi: 10.1097/EDE.0b013e3181d61eeb.

Mohammad Lotfollahi, Anna Klimovskaia Susmelj, Carlo De Donno, et al. Predicting cellular responses to complex perturbations in high-throughput screens. *Molecular Systems Biology*, 19 (6):e11517, 2023. doi: 10.15252/msb.202211517.

Mario Lucic, Karol Kurach, Marcin Michalski, Sylvain Gelly, and Olivier Bousquet. Are GANs created equal? A large-scale study. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

Xinjie Mao, Songming Zhang, Qianhong Wen, et al. Benchmarking virtual cell models for in-the-wild perturbation response. *arXiv preprint arXiv:2604.27646*, 2026. doi: 10.48550/arXiv.2604.27646. URL https://arxiv.org/abs/2604.27646.

Nick McGreivy and Ammar Hakim. Weak baselines and reporting biases lead to overoptimism in machine learning for fluid-related partial differential equations. *Nature Machine Intelligence*, 6 (10):1256–1269, 2024. doi: 10.1038/s42256-024-00897-5.

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

Marija Milacic, Deidre Beavers, Patrick Conley, et al. The Reactome pathway knowledgebase 2024. *Nucleic Acids Research*, 52(D1):D672–D678, 2024. doi: 10.1093/nar/gkad1025.

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

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

Stefan Peidli, Tessa D. Green, Ciyue Shen, et al. scPerturb: harmonized single-cell perturbation data. *Nature Methods*, 21(3):531–540, 2024. doi: 10.1038/s41592-023-02144-y.

Jonas Peters, Peter Bühlmann, and Nicolai Meinshausen. Causal inference by using invariant prediction: identification and confidence intervals. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 78(5):947–1012, 2016. doi: 10.1111/rssb.12167.

Joseph M. Replogle, Reuben A. Saunders, Angela N. Pogson, et al. 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.

John R. Rice. The algorithm selection problem. In *Advances in Computers*, volume 15, pp. 65–118. Academic Press, 1976. doi: 10.1016/S0065-2458(08)60520-3.

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.

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 preprint*, 2026. doi: 10.64898/2026.07.23.740433. URL https://doi.org/10.64898/2026.07.23.740433.

Donald J. Schuirmann. A comparison of the two one-sided tests procedure and the power approach for assessing the equivalence of average bioavailability. *Journal of Pharmacokinetics and Biopharmaceutics*, 15(6):657–680, 1987. doi: 10.1007/BF01068419.

Uri Simonsohn, Joseph P. Simmons, and Leif D. Nelson. Specification curve analysis. *Nature Human Behaviour*, 4(11):1208–1214, 2020. doi: 10.1038/s41562-020-0912-z.

Sara Steegen, Francis Tuerlinckx, Andrew Gelman, and Wolf Vanpaemel. Increasing transparency through a multiverse analysis. *Perspectives on Psychological Science*, 11(5):702–712, 2016. doi: 10.1177/1745691616658637.

R. E. Tarone. A modified Bonferroni method for discrete data. *Biometrics*, 46(2):515–522, 1990. doi: 10.2307/2531456.

Itay Tirosh, Benjamin Izar, Sanjay M. Prakadan, et al. Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq. *Science*, 352(6282):189–196, 2016. doi: 10.1126/science.aad0501.

Sudhir Varma and Richard Simon. Bias in error estimation when using cross-validation for model selection. *BMC Bioinformatics*, 7(1):91, 2006. doi: 10.1186/1471-2105-7-91.

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 preprint*, 2026. doi: 10.64898/2026.04.20.719650. URL https://doi.org/10.64898/2026.04.20.719650.

Zhiting Wei, Yiheng Wang, Yicheng Gao, et al. Benchmarking algorithms for generalizable single-cell perturbation response prediction. *Nature Methods*, 23(2):451–464, 2026. doi: 10.1038/s41592-025-02980-0.

Halbert White. A reality check for data snooping. *Econometrica*, 68(5):1097–1126, 2000. doi: 10.1111/1468-0262.00152.

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 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.

Lin Xu, Frank Hutter, Holger H. Hoos, and Kevin Leyton-Brown. SATzilla: Portfolio-based algorithm selection for SAT. *Journal of Artificial Intelligence Research*, 32:565–606, 2008. doi: 10.1613/jair.2490.

APPENDIX

Appendices A–E give the assumptions behind the run-level decisions of Sections 3–5, the models and comparator family, target coverage, the calibration of δ and the registered selector. Appendices F–M support the Replogle results of Sections 6.1–6.3 with the values behind the main figures, the simultaneous reanalysis, the authority factorial, envelope diagnostics, and the control reference, target resampling and rank-metric sensitivities. Appendices N–S report the independent-screen test, the Replogle multi-split test, STATE with gene features, the linear baseline as a predictor, the positive control and the operating characteristics of Sections 6.4–6.5. Appendices T–W provide literature context and relate the work to the earlier five-context study, retained in full in a separate supplementary report.

CONTENTS

A Validity of the run-level decisions

B Models under test and the comparator family

C Target coverage and the scoring set

D Noise floors and threshold calibration

E The cross-fit selector

F Values behind the main figures

G Run-level reanalysis at both reliability thresholds

H Completed comparator analyses and mechanism control

I Mechanism controls across screens and expression scales

J Envelope selection diagnostics

K Disjoint control-reference sensitivity

L Target resampling and seed pairing

M Complementary rank-metric sensitivity

N Independent-screen test on the Adamson screen

O Replogle multi-split test on the reference preprocessing

P STATE with training-data gene features

Q The linear baseline as the predictor under test

R Run-level positive control on the Norman screen

S Calibration and power diagnostics

T Relation to multiverse and specification-curve analysis

U Additional context for the comparator authorities

V Extended related work

W Relation to the earlier five-context study

A VALIDITY OF THE RUN-LEVEL DECISIONS

This appendix gives the conditions under which the run-level verdicts of Section 4 control error. It then proves that enlarging a family aggregated by a per-target maximum can only remove support, and shows that a comparator chosen by a fitted rule has no such monotonicity.

Proposition 1 (Exact run-level tests)**.** *Fix the scoring choices, the comparator rule c and the margin δ. Let the indicators $Y_r = \mathbf{1}\{W_r(c) > \delta\}$, $r = 1, \ldots, R$, be independent with success probabilities at most ρ, let $K = \sum_r Y_r$ and let $X \sim \mathrm{Bin}(R, \rho)$. Then declaring support when $\Pr(X \geq K) \leq \alpha$ has probability at most α. If instead every success probability is at least ρ, declaring refutation when $\Pr(X \leq K) \leq \alpha$ has probability at most α.*

Proof. The upper tail $G(k) = \Pr(X \geq k)$ is non-increasing in k, so the support event is $\{K \geq k^*\}$ with $k^* = \min\{k : G(k) \leq \alpha\}$, or $k^* = R + 1$ when no count qualifies. A sum of independent Bernoulli variables whose success probabilities are at most ρ is stochastically dominated by X, so $\Pr(K \geq k^*) \leq \Pr(X \geq k^*) = G(k^*) \leq \alpha$. For refutation the lower tail $F(k) = \Pr(X \leq k)$ is non-decreasing, the event is $\{K \leq k_*\}$ with $k_* = \max\{k : F(k) \leq \alpha\}$, or $k_* = -1$ when no count qualifies, and K now stochastically dominates X, so $\Pr(K \leq k_*) \leq F(k_*) \leq \alpha$. □

The statement covers single-environment counts based on $V(f_r, c, e)$ and joint counts based on $W_r(c) = \min_e V(f_r, c, e)$ alike, because each is a fixed function of the run. It does not cover runs that share randomness beyond the fixed screens, or scoring choices adapted after test outcomes were seen. Section 5 records where the campaign departs from this ideal.

Corollary 1 (Selection among fixed specifications)**.** *Let a scoring specification be chosen from m fixed candidates by any rule, including one that reads the test outcomes, and let each candidate's test satisfy Proposition 1 at level α/m in a fixed direction. Then the test of the chosen specification falsely rejects its null hypothesis with probability at most α.*

Proof. A false rejection by the chosen test is a false rejection by one of the m candidate tests, so its probability is at most the sum of theirs, $m \cdot \alpha/m$. □

The per-model expression scale is the only primary-verdict scoring choice made after confirmatory outcomes were seen. The choice was between the raw-count and log-scale versions of each registered dataset (Table T.1). The three control arms registered on the same date are secondary analyses that leave every primary verdict unchanged. The interpretation of A_2 is not a scoring choice. At the primary $\rho = 0.75$, every refuted verdict of the sixty-run campaign, in both coverage arms, has $K = 0$ of 10 and tail $0.25^{10} = 9.54 \times 10^{-7}$, below α/m for every m up to 13,107, so each refutation keeps its individual level α whichever scale was chosen and however it was chosen. The two GEARS verdicts against A_2 are inconclusive and support no claim.

Proposition 2 (Envelope monotonicity)**.** *Let $\mathcal{B} \subseteq \mathcal{B}'$ be registered families whose per-target envelopes $\sigma(c_{\mathcal{B}}; p, e) = \max_{b \in \mathcal{B}} \sigma(b; p, e)$ and $c_{\mathcal{B}'}$ are evaluated on the same targets and the same realization. Then $W_r(c_{\mathcal{B}'}) \leq W_r(c_{\mathcal{B}})$ for every run r, so $K(c_{\mathcal{B}'}) \leq K(c_{\mathcal{B}})$. Enlarging the family can revoke support but never create it, and can create a refutation but never revoke one.*

Proof. For every target the maximum over $\mathcal{B}'$ is at least the maximum over $\mathcal{B}$, and averaging over targets gives $s(c_{\mathcal{B}'}, e) \geq s(c_{\mathcal{B}}, e)$. Hence $V(f_r, c_{\mathcal{B}'}, e) \leq V(f_r, c_{\mathcal{B}}, e)$ in every environment, and the minimum over environments preserves the inequality. No indicator $\mathbf{1}\{W_r > \delta\}$ can rise, so K cannot rise. The support tail $\Pr(X \geq K)$ is non-increasing in K and the refutation tail $\Pr(X \leq K)$ is non-decreasing, which gives the two conclusions. □

Monotonicity holds for a fixed target set and aggregation unit. It does not transfer a verdict between units, because a per-target statistic and an average over target groups are different quantities. The unit is therefore fixed in advance and used throughout.

Maximum aggregation versus selection. The proof uses the fact that a maximum over a larger family is at least the maximum over a smaller one on the same realization. Comparator selection does not have this property. If a rule fitted on one fold selects the comparator for another, the inequality can reverse. The rule may select a member that scores poorly on the evaluation fold, so enlarging $\mathcal{B}$ raises the measured excess rather than lowering it. Four targets in two folds make this concrete. Let f score 0.5 on every target and b_1 score 0.4, so the excess against $\{b_1\}$ is 0.10. Add b_2 scoring 0.9 on the first fold and -0.4 on the second. The rule fitted on the first fold picks b_2, giving excess 0.9 on the second. The rule fitted on the second fold picks b_1, giving 0.1 on the first. The mean excess against $\{b_1, b_2\}$ is therefore 0.50. The per-target maximum over the same family gives -0.15, as Proposition 2 requires.

The registered selector gives a real-data example. The cross-fit selector in Section 6.2 uses truth-free features but is weaker than the fixed comparator it was intended to improve on. Measured added value against it is therefore *higher*, not lower. Proposition 2 therefore governs nested envelopes aggregated by a maximum and does not govern the three comparator authorities, which form an unordered axis.

Scope. The tests concern a run's added value over a registered comparator under a fixed score, not causal validity. A comparator absent from the registered family is not tested, and a stable confounder can pass until its axis is registered. The proofs and scope of the earlier five-context battery, whose unit was a target or a Reactome pathway cluster (Milacic et al., 2024), are in the earlier-study report in the supplementary archive.

Design calculations. For the registered $\rho = 0.75$ and $\alpha = 0.0125$, support is first attainable at $R = 16$, where all sixteen runs must succeed. Power is not monotone in R, because the critical count moves in integer steps. At a true success probability $q = 0.95$ it is 0.4401 at $R = 16$ and 0.3585 at $R = 20$. $R = 30$, requiring 28 successes, is the smallest budget with power above 0.8 at

Table B.1: Training configurations of the models under test. Every run uses the configuration shown and differs only in its seed. Cell caps apply to the training and evaluation files.

Model	Release	Input	Unseen target	Training
GEARS	cell-gears 0.1.2	raw counts	gene-ontology graph	20 epochs, width 64, batch 32
STATE	arc-state 0.10.5	$\log(1+x)$	one-hot index	4000 steps, width 672, 4 layers, sets of 32
CPA	cpa-tools 0.8.8	$\log(1+x)$	learned embedding	50 epochs, latent 128, batch 512

All three hold out 10% of the training data for validation and use at most 80,000 training cells. GEARS and STATE predict at most 20,000 evaluation cells.

$q = 0.95$ (0.8122). At that budget, power is 0.4114 at $q = 0.9$, 0.1514 at $q = 0.85$ and 0.0442 at $q = 0.8$. Power 0.8 at $q = 0.9$ needs $R = 64$.

B MODELS UNDER TEST AND THE COMPARATOR FAMILY

Models under test. Each model uses its released implementation with one configuration, fixed before the confirmatory split was scored and not tuned on it. Runs differ only in training seed (Table B.1). No model received a hyperparameter search. The comparators received none either, and choosing among members is the only selection in the family. GEARS represents an unseen target gene through its gene-ontology graph, builds its co-expression graph from the training perturbations, and cannot address a target absent from that graph. STATE and CPA assign each perturbation its own input index or embedding, so a held-out target reaches them through weights that received no gradient during training. On Replogle, GEARS is trained and scored on raw counts, whereas its reference pipeline uses log-normalized expression. On Adamson it uses the preprocessed values distributed with GEARS. STATE and CPA use $\log(1+x)$ of the raw counts, an elementwise transform without library-size normalization. Both screens are split by perturbation with test fraction 0.2 and seed 7, after excluding perturbations with fewer than twenty cells. The rule uses no expression data. The development split applies the same rule to the training targets. Each model run draws at most 80,000 training cells with its own seed, whereas the comparators are fitted once on the frozen dataset's 80,000-cell training sample. Both use the same training targets. Each verdict therefore compares a run with one fitted comparator, and the Replogle multi-split test fits both on the same cells (Appendix O). In preliminary measurements a run took about 4.3 GPU hours for GEARS, 0.5 for STATE and 1.4 for CPA on a Replogle screen.

Comparator family. Every member is fitted on a screen's training targets in the model's expression scale, without using any test cell. The control reference is the mean training control profile $\bar{C}$, and $\bar{X}_p$ is the mean training profile of perturbation p.

- *Identity* predicts $\bar{C}$, so its shift is zero and its score is set to zero.
- *Pseudobulk-delta* adds $\bar{X}_p - \bar{C}$ for a training perturbation and, for an unseen target, the mean of all perturbed training cells minus $\bar{C}$.
- *Mean-shift* adds the average of the cell-level shifts from $\bar{C}$, per training perturbation and, for an unseen target, over all perturbed training cells.
- *Ridge* regresses cell-level shifts on an unpenalized intercept and one-hot indicators of the training perturbations with penalty 1. An unseen target has no indicator, so ridge returns the intercept. It weights each training perturbation's mean shift by $n_p/(n_p+1)$ for n_p cells, close to equally, whereas pseudobulk-delta and mean-shift weight by cell count.
- *Expression magnitude* predicts the shift $a\bar{C}$, with a the least-squares projection of the mean training shift onto $\bar{C}$.
- *Cell cycle* predicts a constant shift a on the S-phase and G2/M marker genes of Tirosh et al. (2016) distributed with Seurat and zero elsewhere, with a the mean training shift over those genes.

Each member therefore predicts one fixed shift across held-out targets. The family spans five fixed directions besides identity. The per-target envelope picks, target by target, the direction best aligned

Table C.1: The two frozen environments. Dropping targets with fewer than twenty cells removes 36 targets in K562 and 189 in RPE1. The screens are therefore not equally well populated and are reported separately rather than pooled. The gene panels also differ, so the worst case over environments is taken with each environment scored on its own panel. Rebuilding either split from the source data reproduces the registered partition exactly.

Environment	Targets kept	Train / test targets	Genes
K562	2021 of 2057	1617 / 404	8563
RPE1	2204 of 2393	1763 / 441	8749

with the held-out truth, and the selector A_2 picks one from two training-only features of the target gene. Mean-shift and pseudobulk-delta coincide in exact arithmetic. In the scored predictions they differ only through single-precision accumulation of the pseudobulk mean, which lowers the pseudobulk score by at most 0.00039 (K562, raw counts). Every run's excess over mean-shift is lower still, so no verdict depends on this numerical difference.

C Target coverage and the scoring set

Section 5.2 scores each model–screen cell on the targets the model can address and adds a control-filled robustness arm. This appendix gives the frozen splits and the coverage rule behind both arms.

A predictor may be structurally unable to address some targets. GEARS predicts only perturbations that appear in its perturbation graph. The campaign covers 403 of 404 in K562 (excluding OR4F4), and 437 of 441 in RPE1 (excluding FAM102B, FAM229A, OR4F4 and SEM1). All comparators are rescored on the covered targets of each model–screen cell. Because the scoring set changes the measured excess, it is fixed in advance rather than chosen afterwards. The rule was registered before any confirmatory run. Identity predicts the control profile, so its predicted shift is exactly zero and its delta correlation has a zero denominator. The scoring routine assigns zero to such a correlation and -1 only to values that remain non-finite. Identity therefore remains among the six members of the evaluated envelope.

R1. The primary scoring set for a model–screen cell is the intersection of that model's covered targets with the frozen test targets. The model and every comparator, under all three authorities, are scored on that same set. Matching target sets makes the within-cell comparison fair, but a different set can still change the numerical difference. The robustness arm measures this sensitivity.

R2. Each cell reports the number of targets scored, the number excluded, and the names of the excluded targets.

R3. A robustness arm scores the same cell on the complete frozen test set, with targets the model cannot address filled by the control profile. This arm includes unsupported targets in the model evaluation.

R4. If the two arms give different verdicts for a cell, that cell is reported as *inconclusive* and both verdicts are stated. The more favorable one is not adopted. Without this clause the gap between the two arms would itself be an adjustable quantity.

R5. Coverage differs between models, so added value is not compared across rows of a results table. Each model receives its own verdict, and every table states the number of targets behind each row.

D Noise floors and threshold calibration

The margin δ is set from measured run-to-run variation rather than by convention. Two sources of variation are measured separately, GPU nondeterminism at a fixed seed and variation across training seeds. Both use GEARS in raw counts on a development split of the K562 screen, which holds out 323 of the 1617 confirmatory training targets and trains on the remaining 1294. It therefore shares no target with the confirmatory test split, and all its runs preceded the registration.

Table D.1: Seed-to-seed variability of six candidate metrics over six GEARS seeds at the confirmatory configuration on the K562 development split. *Pooled* correlates one concatenated vector across all targets. *Per-target* averages the target-specific statistics over the 320 of the 323 held-out development targets that GEARS covers. Both columns score the same predictions with different aggregations.

Metric (standard deviation over 6 seeds)	pooled	per-target	ratio
Delta correlation	3.339×10^{-1}	1.387×10^{-2}	$24\times$
Delta correlation (range, not s.d.)	7.898×10^{-1}	3.455×10^{-2}	$23\times$
DE direction accuracy	2.458×10^{-1}	1.876×10^{-2}	$13\times$
DE overlap	1.291×10^{-1}	1.168×10^{-2}	$11\times$
Mean absolute error	7.624×10^{-3}	7.973×10^{-3}	$\approx 1\times$
Flat correlation on expression	4.162×10^{-4}	1.667×10^{-4}	$2.5\times$

The nondeterminism floor. Eight three-epoch repetitions with the same seed, split and configuration differ only through GPU nondeterminism. Over those eight, the pooled delta correlation has standard deviation 4.31×10^{-2} and a range of 1.15×10^{-1}, taking values between 0.689 and 0.805. A flat correlation on absolute expression has standard deviation 3.11×10^{-4} over the same eight runs, which is $1/140$ of the delta correlation's.

Training-seed variability. Training seeds also vary in the protocol. Table D.1 reports six GEARS seeds at the twenty-epoch confirmatory configuration and shows whether the primary metric must be this variable. Final training losses range from 1.3545 to 1.4375, and GEARS selects the scored checkpoint on its validation targets. The six runs therefore reach similar losses although their pooled scores differ widely. The same δ is applied to STATE, CPA and the other screens without recalibration.

Threshold selection. The primary metric is the per-target delta correlation, and δ is set at its measured seed-to-seed standard deviation, $\delta = 0.0139$. The conservative alternative based on the observed range, $\delta = 0.0346$, is also reported wherever the two thresholds yield different verdicts. Both were fixed before any confirmatory run.

The flat correlation on absolute expression is more stable than either and is not used. It is stable because it is nearly insensitive to the perturbation effect, which is the quantity the models are built to predict. Choosing it for its stability would be the metric analogue of choosing a weak comparator.

Aggregation sensitivity. The same predictions score 0.432 pooled and 0.172 per-target, and the pooled version has a 24 times larger seed-to-seed standard deviation. Metric-dependent rankings have been reported in this field (Mao et al., 2026). Table D.1 reports the variance ratio on one set of predictions. At three runs per model, the seed-to-seed range of the pooled form covers most of its own scale.

E THE CROSS-FIT SELECTOR

Registered selector. The cross-fit selector A_2 is one simple instance, frozen before the confirmatory runs. Its features are two truth-free properties of the target gene, its mean and its standard deviation in the training control cells. Labels come from a three-fold cross-fit over training targets, where the comparators are fitted on two folds and the winning member on the held-out fold becomes that target's label. Test-time selection takes a majority vote over the fifteen nearest labeled targets in standardized feature space. A target absent from the panel receives the fixed comparator. On the development split it scores 0.0502 against the fixed comparator's 0.1482, so it is *worse* by 0.0980. It was nevertheless registered. Because the registered expectations covered only A_1 and A_3, the interpretation of its confirmatory result in Section 6.2 is exploratory.

Selector discrimination. The out-of-fold labels assign a shortcut comparator to 738 of 1294 training targets, and the selector learns that majority. It does not learn *where* the shortcut wins, so at test

Table F.1: Six model–screen cells, $R = 10$ frozen runs each, under the fixed comparator A_1 and the envelope A_3. $s(f)$ is the run-averaged model score and $\bar{W}$ its mean excess on the named screen. The gap $s(c_{A_3}) - s(c_{A_1})$ is shared by cells on the same screen and scale. The last column divides the model's absolute mean excess over the fixed comparator by that gap. Every cell gives $K = 0/10$ under both, a primary refutation in both coverage arms. GEARS (†) is scored in raw counts, the others in log space.

Model	Env	$s(f)$	$s(c_{A_1})$	$\bar{W}^{A_1}$	$s(c_{A_3})$	$\bar{W}^{A_3}$	gap/δ	$\lvert\bar{W}^{A_1}\rvert$/gap
GEARS	K562†	0.1571	0.1851	−0.0280	0.4042	−0.2471	15.8	0.13
GEARS	RPE1†	0.3658	0.3931	−0.0273	0.5093	−0.1435	8.4	0.23
STATE	K562	0.0450	0.2291	−0.1841	0.2639	−0.2190	2.5	5.29
STATE	RPE1	0.1978	0.3939	−0.1961	0.4313	−0.2335	2.7	5.24
CPA	K562	0.0472	0.2291	−0.1818	0.2639	−0.2167	2.5	5.22
CPA	RPE1	−0.0490	0.3939	−0.4429	0.4313	−0.4804	2.7	11.83

time it routes 260 of 323 targets to a shortcut, including the targets where the shortcut loses. The failure is one of discrimination, not of the label distribution.

Scope. Two richer designs were examined on the development split and not registered, a larger k-NN and a regression on the comparators' predicted deltas. Their features, standardized on training targets, fell far outside the training range on test targets, so these designs say nothing about selectors in general. The regression, for instance, had an out-of-fold mean of $+0.17$ but a predicted test mean of -7.40 on a metric bounded in $[-1, 1]$, and the fault was not resolved. The registered selector uses features that transfer, with test drifts of 0.108 and 0.040 training standard deviations. The registered claim is narrow. One simple deployable selector is worse than the fixed comparator, and no claim is made that no better selector exists. Section 6.2 interprets its result within this scope, and Appendix I tests a stronger selector with the same information.

F VALUES BEHIND THE MAIN FIGURES

Table F.1 reports scores and mean excesses for the six registered model–screen cells in Figure 2. Table F.2 reports the primary verdict for every test and authority except the linear baseline as the predictor under test, which Table Q.1 reports.

G RUN-LEVEL REANALYSIS AT BOTH RELIABILITY THRESHOLDS

The registered primary reliability threshold is $\rho = 0.75$. The threshold $\rho = 0.5$ is secondary. An earlier analysis applied only the secondary threshold. The reanalysis retains all scores and evaluates both thresholds with the appropriate upper and lower binomial tails. The registration is unchanged. The level $\alpha = 0.0125$ is registered as an individual-test level. It is not a familywise allocation over the three models and the expanded results table, so familywise control is reported separately by Holm.

The post-hoc Holm family has $3 \times 3 \times 3 \times 2 \times 2 \times 2 \times 2 = 432$ directional slots for model, authority, endpoint, coverage, reliability threshold, excess threshold and direction. Every slot is assessed (Appendix H). The family size is fixed, so a missing measurement would have remained unassessed rather than reducing the family. Holm refutation and support are interpreted only conditional on valid individual tails and independent training-run indicators within each frozen screen and recipe. Because the tests are discrete, not every slot can be rejected at $R = 10$. A $\rho = 0.5$ tail cannot fall below $0.5^{10} \approx 9.8 \times 10^{-4}$ and a $\rho = 0.75$ support tail below $0.75^{10} \approx 0.056$, while only the 108 $\rho = 0.75$ refutation slots can reach 0.25^{10}. After at most those 108 rejections the next Holm threshold is $0.05/324 \approx 1.5 \times 10^{-4}$, and a restriction to testable hypotheses (Tarone, 1990) leaves only the same 108 slots. Neither Holm's procedure nor its restriction to testable hypotheses could therefore confirm the $\rho = 0.5$ support of GEARS against A_2. A weighted procedure fixed in advance could have, but none was registered. At $R = 10$ the familywise procedure cannot reject that cell's support null, so its simultaneous verdict is inconclusive. Table G.1 reports the joint endpoint under

Table F.2: Primary verdicts ($\rho = 0.75$, $\delta = 0.0139$, $\alpha = 0.0125$) as K/R, with $+$ supported, $-$ refuted and ? inconclusive, on covered targets. The training-selected fixed member is a post-hoc instance of A_1. It is ridge on Replogle and pseudobulk-delta on the registered Adamson split, and is refitted on each fresh split. The control-filled arm agrees wherever scored. The linear gene-embedding baseline is post hoc except on the fresh Replogle splits and for STATE with training-data gene features (Appendix P). The last row is the Norman positive control against the registered family, and Table R.1 gives its refutation once an additive baseline is registered.

Model	Screen	A_1 registered	A_1 training-selected	Linear	A_2 selector	A_3 envelope
GEARS	K562	$0/10-$	$0/10-$	$0/10-$	$10/10$?	$0/10-$
GEARS	RPE1	$0/10-$	$0/10-$	$0/10-$	$6/10$?	$0/10-$
STATE	K562	$0/10-$	$0/10-$	$0/10-$	$0/10-$	$0/10-$
STATE	RPE1	$0/10-$	$0/10-$	$0/10-$	$0/10-$	$0/10-$
CPA	K562	$0/10-$	$0/10-$	$0/10-$	$0/10-$	$0/10-$
CPA	RPE1	$0/10-$	$0/10-$	$0/10-$	$0/10-$	$0/10-$
GEARS	K562, 12 splits	$0/12-$	$0/12-$	$0/12-$	$0/12-$	$0/12-$
GEARS	RPE1, 12 splits	$0/12-$	$0/12-$	$0/12-$	$0/12-$	$0/12-$
STATE	K562, features	$0/10-$	$0/10-$	$0/10-$	$0/10-$	$0/10-$
STATE	RPE1, features	$0/10-$	$0/10-$	$0/10-$	$0/10-$	$0/10-$
GEARS	Adamson, 1 split	$0/30-$	$0/30-$	$0/30-$	$0/30-$	$0/30-$
GEARS	Adamson, 30 splits	$0/30-$	$0/30-$	$1/30-$	$0/30-$	$0/30-$
Control	Norman	$30/30+$	n/a	n/a	n/a	$30/30+$

Table G.1: Joint K562/RPE1 reliability, primary coverage, $\rho = 0.75$ and $\delta = 0.0139$. Each run is paired by seed ID and takes the smaller of its two screen-specific excesses. The two screens are not treated as ten biological replicates.

Model	Authority	K/R	p_-	Holm p_-	Simultaneous
GEARS	A1	0/10	9.54×10^{-7}	4.12×10^{-4}	refuted
GEARS	A3	0/10	9.54×10^{-7}	4.12×10^{-4}	refuted
STATE	A1	0/10	9.54×10^{-7}	4.12×10^{-4}	refuted
STATE	A3	0/10	9.54×10^{-7}	4.12×10^{-4}	refuted
CPA	A1	0/10	9.54×10^{-7}	4.12×10^{-4}	refuted
CPA	A3	0/10	9.54×10^{-7}	4.12×10^{-4}	refuted

the fixed comparator and the envelope, and Table G.2 the sensitivity of the single selector-supported cell to both excess thresholds and both corrections.

The reanalysis scores the same sixty prediction profiles at both aggregate and per-target resolution. The smallest distance from an excess to either threshold is 0.000910, exceeding rounding uncertainty, so rounding does not determine any success indicator. Subset analyses use unrounded target scores with the original covered-target and control-filled sets.

The probability-of-improvement control compares each model run with a deterministic comparator on the same screen. It averages win indicators, assigning half weight to exact ties, at zero excess. The bootstrap uses 10,000 resamples (seed 20260914), resampling training-seed blocks synchronously

Table G.2: GEARS/K562 A_2 sensitivity on the same ten retained runs, primary coverage. Both excess thresholds were registered. The individual-test rule uses $\alpha = 0.0125$, and the Holm-adjusted rule uses the complete post-hoc directional family at 0.05.

δ	ρ	K/R	p_+	Holm p_+	Pointwise	Simultaneous
0.0139	0.75	10/10	0.0563	1	inconclusive	inconclusive
0.0346	0.75	6/10	0.922	1	inconclusive	inconclusive
0.0139	0.5	10/10	9.77×10^{-4}	0.324	supported	inconclusive
0.0346	0.5	6/10	0.377	1	inconclusive	inconclusive

Table H.1: Completed A_2 results on the covered targets, $\rho = 0.75$, $\delta = 0.0139$, individual $\alpha = 0.0125$. Each row uses the same ten frozen training runs. The control-filled arm gives the same primary verdict in every cell. For GEARS, the control-filled target counts are 404 and 441 rather than 403 and 437.

Model	Screen	Targets	$s(c_{A_2})$	$\bar{W}$	K/R	Primary verdict
GEARS	K562	403	0.114884	$+0.042233$	10/10	inconclusive
GEARS	RPE1	437	0.350801	$+0.015038$	6/10	inconclusive
STATE	K562	404	0.199684	-0.154725	0/10	refuted
STATE	RPE1	441	0.384652	-0.186860	0/10	refuted
CPA	K562	404	0.199684	-0.152438	0/10	refuted
CPA	RPE1	441	0.384652	-0.433668	0/10	refuted

across the fixed screens. It does not resample biological screens or targets. The observed estimates are zero for all A_1/A_3 cells and all STATE/CPA A_2 cells, and one for both GEARS A_2 screens. The corresponding percentile intervals collapse to $[0, 0]$ and $[1, 1]$. Joint success takes the minimum before evaluating its sign. The separate mean-over-screen summary averages screen-specific indicators. These estimands remain distinct. Boundary intervals can collapse when all ten runs agree and must not be interpreted as population certainty. This is a check of the deterministic-comparator estimand, not a new statistical method. The rliable package reproduces its point estimates exactly (Section 6.3). The package comparison contains 54 matched run-level estimates and their percentile intervals, plus 12 separate target-level descriptive comparisons. All matched run-level estimates are at zero or one and their intervals are degenerate. Their numerical agreement therefore does not establish that task-wise and paired-seed bootstrap procedures are generally equivalent. The joint endpoint has no counterpart in this package comparison.

Matched ordinary inference. A separate implementation evaluates the same one-sided exact-binomial tails and sorts the complete 432-direction family for Holm correction. It reproduces all 216 outcome-coordinate decisions both before and after correction. At $\rho = 0.75$, $\delta = 0.0139$ and primary coverage, both implementations return 24 refuted and three inconclusive outcomes among 27 coordinates. Coordinates overlap across thresholds, authorities, endpoints and coverage arms. They are not independent biological samples. This agreement rules out an incremental decision advantage over ordinary exact-binomial inference under the matched protocol. It does not assess the novelty of comparator specification or compare different estimands as if they were interchangeable.

For the exact individual binomial rule at $\rho = 0.75$ and $\alpha = 0.0125$, all-success support first becomes attainable at $R = 16$. At $R = 30$, the support region is $K \geq 28$. Its probability under a hypothetical true $q = 0.95$ is 0.8122. This analytic design calculation does not establish empirical power for a biological predictor or validate a dependence assumption.

H Completed comparator analyses and mechanism control

These analyses use the original sixty training runs. The A_2 scores were computed after the per-model expression scale was fixed. All six model–screen cells have A_1, A_2, A_3 scores and paired coverage arms. All 432 directional slots are assessed, and the secondary GEARS/K562 A_2 Holm-adjusted support tail is 0.324, with an inconclusive simultaneous verdict. Table H.1 lists the completed selector results for all six cells.

Post-hoc comparator subsets. For each of three models and two excess thresholds, all 63 nonempty subsets are evaluated on the two screens and their paired-seed joint endpoint, under both coverage arms. The 2,268 descriptive rows are overlapping comparisons of the same runs, not new independent observations. Each subset envelope takes the maximum within a target before averaging. Joint excess takes the smaller screen excess within the same seed. The six cumulative subsets are a display order only. They do not constitute a registered nested family or validate ordered testing.

Training-only mechanism control. Within the same six-member family, three-fold out-of-fold training scores select the single fixed comparator. An outer three-fold comparison with inner three-

Table I.1: Complete descriptive comparator-mechanism controls in raw-count and log expression scales. Outer means select the procedure and are not test estimates. All four settings choose fixed ridge. Strong fixed and nested training-only selection therefore have identical test scores. Each screen uses its control-filled target population. Scales share their underlying screens and do not provide four independent biological replications.

Screen	Scale	Targets	Outer A_2	Outer fixed	Test A_2	Test fixed
K562	raw	404	0.102306	0.182408	0.114703	0.192957
RPE1	raw	441	0.363535	0.412100	0.346973	0.394436
K562	log	404	0.179549	0.201705	0.199684	0.229588
RPE1	log	441	0.377751	0.387036	0.384652	0.394274

fold selection chooses between that procedure and the original A_2. Ties favor the fixed procedure. On GEARS/K562 the outer scores are 0.102306 for A_2 and 0.182408 for fixed. The selected fixed member is ridge. The original test targets are evaluated only after selection is fixed. On all 404 frozen targets, original A_2 scores 0.114703, whereas strong fixed and nested training-only selection both score 0.192957. These values use the control-filled population and must not be compared with the 403-target primary scores as if the populations were identical. Nested training-only selection avoids the observed weak selector in this setting. This is a descriptive mechanism control outside the 432-slot family, not proof that all selectors fail, a matched-budget uncertainty comparison, or validation on an independent screen. The original A_2 result is reported unchanged.

Remaining evidence. Real-expression calibration with distinct study and training-run units, random zero-skill envelope padding and the ordered-testing comparison remain unmeasured. Neither the full subset enumeration nor the mechanism controls substitutes for these experiments. The two-screen graded oracle and structured padding are reported in Appendix S, the rliable comparison in Section 6.3, and the completed two-screen, two-scale extension in Appendix I.

I MECHANISM CONTROLS ACROSS SCREENS AND EXPRESSION SCALES

This appendix supports the claim in Section 6.3 that a training-only fixed comparator is stronger than the registered selector in every screen and expression scale. The three additional settings retain the K562/raw-count control's six-member comparator family, inner and outer three-fold training selection, seed 7, 15-neighbor selector and tie rule. They require no additional deep-model training. Selection uses training targets only. Original test targets are scored after the selection is fixed. Table I.1 includes all four settings, without choosing a scale from its test result.

The fixed comparator exceeds original A_2 on each test population. The RPE1/log difference is only 0.009623, so the size of the improvement is not uniform. These observations provide evidence of a weakness in this particular selector. They do not establish that cross-fitting in general fails or that an oracle advantage is unavailable to every deployable comparator. No new simultaneous hypothesis family or selector-selection interval is inferred from these four deterministic comparator settings. These controls do not alter the registered A_2 outcomes, thresholds or negative results.

A stronger selector with the same information. A selector S, registered before its test scoring, retains A_2's two features, standardization and 15-neighbor neighborhood. For each family member, it averages out-of-fold training scores over the neighbors and chooses the member with the highest average. The evaluation uses the same nested folds and reproduces every stored A_2 choice, resolving tied neighbor distances in a fixed order. Table I.2 shows that S recovers most of A_2's deficit to the fixed member but never exceeds it in nested training evaluation. Nested selection therefore keeps the fixed member in all four settings. On the test targets S exceeds the fixed member only on RPE1 in log space, by 6.6×10^{-5}, and captures essentially none of the envelope's advantage. Every model is refuted against S and against the nested choice, at $K = 0/10$ in every cell.

For the exact model-specific target populations, the same GEARS runs change from positive mean excess against A_2 to negative mean excess against ridge on both screens. The intervals quantify variation over model training runs only. The uncertainty in selecting ridge from training targets remains unmeasured.

Table I.2: The stronger selector S against the registered A_2 and the training-selected fixed member (ridge), as macro delta correlation. Outer scores from nested training-only evaluation determine the choice between S and the fixed member, reported as Nested. Test scores use the control-filled test targets. Recovered is the fraction of A_2's test deficit to the fixed member that S recovers. Scales share their screens.

		Outer, training only				Test targets				
Screen	Scale	A_2	S	fixed	Nested	A_2	S	fixed	A_3	Recovered
K562	raw	0.1023	0.1666	0.1824	fixed	0.1147	0.1762	0.1930	0.4037	79%
RPE1	raw	0.3635	0.4106	0.4121	fixed	0.3470	0.3925	0.3944	0.5068	96%
K562	log	0.1795	0.2017	0.2017	fixed	0.1997	0.2284	0.2296	0.2639	96%
RPE1	log	0.3778	0.3870	0.3870	fixed	0.3847	0.3943	0.3943	0.4313	101%

Table I.3: Descriptive paired bootstrap over the same ten retained training runs. s is the comparator score, and $\bar{W}$ is model excess over that comparator. The interval resamples training runs synchronously within each model–screen cell. It does not resample targets or the selector's training choices, is not simultaneous, and does not change any registered verdict.

Model	Screen	$s(A_2)$	$\bar{W}_{A_2}$	95% CI	$s(\text{ridge})$	$\bar{W}_{\text{ridge}}$	95% CI
GEARS	K562	0.1149	$+0.0422$	$[+0.0327, +0.0523]$	0.1934	-0.0363	$[-0.0459, -0.0263]$
GEARS	RPE1	0.3508	$+0.0150$	$[+0.0123, +0.0177]$	0.3987	-0.0329	$[-0.0356, -0.0302]$
STATE	K562	0.1997	-0.1547	$[-0.1697, -0.1393]$	0.2296	-0.1846	$[-0.1996, -0.1692]$
STATE	RPE1	0.3847	-0.1869	$[-0.2017, -0.1723]$	0.3943	-0.1965	$[-0.2114, -0.1819]$
CPA	K562	0.1997	-0.1524	$[-0.1907, -0.1172]$	0.2296	-0.1823	$[-0.2206, -0.1471]$
CPA	RPE1	0.3847	-0.4337	$[-0.4837, -0.3838]$	0.3943	-0.4433	$[-0.4934, -0.3934]$

Information access crossed with selection rule. A_1 and A_3 differ in both information access and selection rule. A fourth cell separates these factors. A_1^* fixes one member for every target, like A_1, but chooses it with held-out truth, as the member with the best mean score over the scored targets. On every screen and scale, the best member in hindsight is ridge, also chosen by training-only selection (Table I.4). Thus fixed-member selection with held-out truth and training-only selection agree, and every model remains refuted under A_1^* at $K = 0/10$. Relative to the registered pseudobulk member, the fixed choice recovers 1.0% to 4.8% of the gap to A_3. Choosing per target changes the comparator in both directions. With training data only it lowers the score, as A_2 does on both screens, and with held-out truth it raises it to A_3. At least 95% of the gap between the registered fixed comparator and the envelope therefore comes from per-target selection with held-out truth. The crossing is post hoc and outside the registered family.

J ENVELOPE SELECTION DIAGNOSTICS

A_3 takes a maximum over a family on the same realization the predictor is scored on, so $\mathbb{E}[\max_b \hat{s}_b] \geq \max_b \mathbb{E}[\hat{s}_b]$ and the empirical envelope is inflated by selection. The reality check of White (2000) addresses the analogous bias for the best model of a specification search. Here selection inflation favors refutation, because an inflated envelope lowers every measured added value and makes refutation under A_3 easier to obtain. A verdict difference involving A_3 could therefore reflect this inflation. The GEARS/K562 secondary, individual-test difference occurs between A_2 and fixed A_1 without using A_3, and primary support is not established. Two diagnostics assess envelope inflation.

Cross-fitted envelope. Selection inflation is removed by choosing the argmax member on one half of each target's test cells and scoring that member on the disjoint half. The disjoint evaluation estimates the selected member's performance without scoring on the cells that selected it. It need not identify the population-best member. The difference between the raw and cross-fitted envelopes is the *max-selection gap*. The max-selection gap must be compared with the authority gap $s(c_{A_3}) - s(c_{A_1})$. Comparing the two gaps shows how much of the authority gap finite-sample selection could explain, although it does not separate information access from the selection rule.

Table I.4: Information access crossed with selection rule, from the retained per-target scores of the covered targets. A_1 registered is pseudobulk-delta. The training-selected and A_1^* columns fix one member for all targets, chosen from training data or with held-out truth. Share is the fraction of the gap from registered A_1 to A_3 that A_1^* recovers. A_3 chooses per target with held-out truth, and A_2, the training-only per-target cell, is in Table I.3. K counts runs that clear $\delta = 0.0139$ against A_1^*, with verdicts at $\rho = 0.75$ that agree in both coverage arms.

Model	Screen	A_1 registered	training-selected	A_1^*	A_3	share	K, A_1^*
GEARS	K562	0.1851	0.1934	0.1934	0.4042	3.8%	0/10 refuted
GEARS	RPE1	0.3931	0.3987	0.3987	0.5093	4.8%	0/10 refuted
STATE	K562	0.2291	0.2296	0.2296	0.2639	1.4%	0/10 refuted
STATE	RPE1	0.3939	0.3943	0.3943	0.4313	1.0%	0/10 refuted
CPA	K562	0.2291	0.2296	0.2296	0.2639	1.4%	0/10 refuted
CPA	RPE1	0.3939	0.3943	0.3943	0.4313	1.0%	0/10 refuted

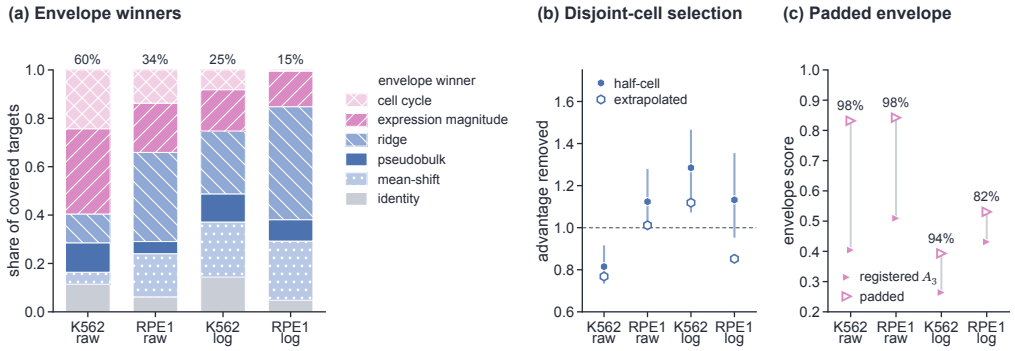


Figure J.1: Envelope diagnostics on the covered targets of each screen and scale, complementing Figure 3. The envelope's strength comes mostly from shortcut members and selection on the scored cells. Enlarging the family raises it further. **(a)** Member that wins each target in the envelope, with the two shortcut members' combined share above each bar. **(b)** Fraction of the envelope's in-sample advantage over A_1 removed when the member is chosen on disjoint cells, with the interval over split repetitions and the extrapolation to an unlimited selection half (open). Values at or above 1 leave no advantage. **(c)** Registered envelope and the envelope padded with every training perturbation, with the share of targets a training perturbation wins.

On the two frozen screens the ratio of the max-selection gap to the authority gap is the selection share of Appendix S, measured at half-cell scoring (Table S.3, Figure J.1(b)). Selection inflation shrinks as scoring noise falls, and the extrapolation to unlimited cells is lower than the half-cell value in all four settings.

Adversarial padding. By Proposition 2, a larger registered family can only revoke a supported verdict. This invites the interpretation that enough zero-skill comparators defeat any predictor, but extreme-value behavior limits their effect. For K independent zero-skill members whose scores carry Gaussian sampling noise of standard deviation σ, the expected maximum grows like $\sigma\sqrt{2\ln K}$, and for sub-Gaussian noise with variance proxy σ^2 it is at most that. The rate is logarithmic in K and small whenever σ is small relative to real differences between comparators. In the earlier Norman study, $K = 1000$ zero-skill members with $\sigma = 0.014$ raised the envelope by only 0.003 (separate earlier-study report in the supplementary archive, not pooled with the present screens).

Random padding is not evaluated on the present screens. Instead, every training perturbation's pseudobulk delta is registered as an additional member, a form of structured padding. Every model–screen cell remains refuted (Appendix S, Table S.4).

Table K.1: Run-level counts for GEARS on K562 under five disjoint control partitions (columns). Each entry gives the covered and control-filled counts out of ten runs, $\delta = 0.0139$. Refutation needs $K \leq 3$ and support is unattainable at $R = 10$. Every other model–screen cell has $K = 0$ against all three comparators in every partition and arm.

Comparator	partition 1	partition 2	partition 3	partition 4	partition 5
A_1 registered	3/3	0/0	0/0	5/6	0/0
A_1 training-selected	0/0	0/0	0/0	3/3	0/0
A_3 envelope	0/0	0/0	0/0	0/0	0/0

Scope of the diagnostics. Both bear on whether A_3 is inflated. The refutations under the fixed comparator do not depend on them, because a refutation under A_3 is implied by a refutation under any single member of the family, by the pointwise inequality of Proposition 2. The diagnostics matter for the opposite case. Where a predictor is *supported* under A_3, a larger or noisier envelope is conservative, and the support is then the finding that has to survive them.

K Disjoint control-reference sensitivity

The primary metric correlates a truth delta with a predicted delta, and both subtract the same control mean. Nicol et al. (2026) argue that such a shared control reference inflates the metric. This analysis removes the shared reference. Each screen's control cells are split into two disjoint, batch-stratified halves that together contain all control cells. The truth delta subtracts the mean of one half, and every predicted delta, from the model or from a comparator, subtracts the mean of the other. The split is repeated for five partition seeds. The sixty retained prediction profiles, the frozen test targets and the fitted comparator recipes are reused, and no model is retrained.

With the unsplit reference, the rebuilt pipeline reproduces the original scores in all six model–screen cells. An independent scoring implementation reproduces every comparator, envelope and model macro score across coverage arms and prediction seeds, with maximum absolute error 4.95×10^{-7}.

Table K.2 reports, for each run, the median over targets of the model's excess over the registered envelope and over the mean-shift comparator, averaged over the five partitions. Every run in every model–screen cell and both coverage arms has a negative median excess against both. GEARS on K562 is the most partition-sensitive cell. Its mean partition standard deviation is 0.0864 in the covered arm, versus at most 0.0172 in the other cells, yet the sign is the same in all ten runs. The analysis is outside the registered protocol and is reported as a sensitivity analysis. Median excess is a descriptive effect size, and perturbations are not treated as independent biological units, so no additional test is claimed.

The registered run-level rule can also be applied directly because each run's and each comparator's macro scores are available under every partition. With the single-screen excess $V(f_r, c, e) = s(f_r, e) - s(c, e)$, $\delta = 0.0139$, $\rho = 0.75$ and $\alpha = 0.0125$, 178 of the 180 decisions over three comparators, five partitions, two arms and six model–screen cells remain refutations. Every STATE, CPA and GEARS/RPE1 decision has $K = 0$. GEARS on K562 is the exception (Table K.1). Against registered A_1 it reaches $K = 5$ and 6 of 10 in the fourth partition, which is inconclusive in both arms, and $K = 3$ in the first, a refutation. Training-selected ridge refutes in every partition, with at most $K = 3$, and the envelope keeps $K = 0$ throughout. The fixed-comparator refutation for GEARS on K562 therefore depends on how the control reference is estimated, while the refutations against training-selected ridge and the envelope do not. Target resampling is a separate question, on which the GEARS/K562 intervals against A_1 and ridge both include zero (Appendix L).

L Target resampling and seed pairing

The run-level count conditions on each screen's frozen target split, and the joint count also conditions on the seed-index pairing across screens. Two descriptive checks address these choices on the covered targets without changing any registered verdict.

Table K.2: Disjoint control-reference sensitivity. Excess is the median over targets of model minus comparator, averaged over five batch-stratified partitions. For the envelope the table gives the mean over ten runs, the range across runs and the mean partition standard deviation. The last two columns count runs with negative excess against the envelope and against the mean-shift comparator. GEARS is reported in both coverage arms, with 403 and 437 covered targets and 404 and 441 after control filling. STATE and CPA cover all 404 and 441 test targets, so their arms coincide.

			Excess over envelope			Excess over	Negative runs	
Model	Screen	Arm	Mean	Range	Part. SD	mean-shift	Env.	M.-s.
GEARS	K562	covered	−0.1635	[−0.1983, −0.1295]	0.0864	−0.0292	10/10	10/10
GEARS	K562	filled	−0.1632	[−0.1961, −0.1295]	0.0881	−0.0289	10/10	10/10
GEARS	RPE1	covered	−0.1052	[−0.1191, −0.0951]	0.0081	−0.0259	10/10	10/10
GEARS	RPE1	filled	−0.1049	[−0.1190, −0.0938]	0.0075	−0.0254	10/10	10/10
STATE	K562	both	−0.1651	[−0.2016, −0.1324]	0.0056	−0.1563	10/10	10/10
STATE	RPE1	both	−0.2036	[−0.2607, −0.1658]	0.0054	−0.1967	10/10	10/10
CPA	K562	both	−0.1550	[−0.2772, −0.1004]	0.0098	−0.1458	10/10	10/10
CPA	RPE1	both	−0.5013	[−0.6610, −0.3657]	0.0172	−0.4963	10/10	10/10

Table L.1: Percentile 95% intervals for the mean excess $\bar{W}$ over the covered targets, resampling the ten training runs and the targets independently, $20{,}000$ replicates. Point estimates equal the mean excesses of Tables F.1 and I.3. GEARS is scored in raw counts, STATE and CPA in log expression. Descriptive only.

Model	Screen	A_1	A_2	A_3	ridge
GEARS	K562	$[-0.099, +0.043]$	$[-0.036, +0.120]$	$[-0.307, -0.188]$	$[-0.103, +0.030]$
GEARS	RPE1	$[-0.039, -0.016]$	$[-0.002, +0.033]$	$[-0.159, -0.128]$	$[-0.044, -0.021]$
STATE	K562	$[-0.212, -0.158]$	$[-0.182, -0.129]$	$[-0.244, -0.196]$	$[-0.212, -0.159]$
STATE	RPE1	$[-0.217, -0.177]$	$[-0.209, -0.166]$	$[-0.254, -0.214]$	$[-0.217, -0.177]$
CPA	K562	$[-0.231, -0.139]$	$[-0.201, -0.110]$	$[-0.264, -0.176]$	$[-0.231, -0.140]$
CPA	RPE1	$[-0.504, -0.385]$	$[-0.495, -0.376]$	$[-0.540, -0.424]$	$[-0.505, -0.385]$

Target resampling. For a model–screen cell and a comparator c, the per-target excess averaged over runs, $\bar{e}_t = R^{-1}\sum_r s(f_r, t) - s(c, t)$, has mean $\bar{W}$ over targets. Table L.1 gives percentile intervals for $\bar{W}$ from 20,000 replicates that resample the ten runs and the covered targets independently with replacement. Each comparator's per-target choice is held fixed, since A_1, A_2 and ridge are selected on training data and A_3 target by target. Every STATE and CPA interval lies below zero, as do the GEARS intervals against A_3 on both screens and against A_1 and ridge on RPE1. The GEARS/K562 intervals against A_1 and ridge, $[-0.099, +0.043]$ and $[-0.103, +0.030]$, include zero and δ. Another target draw from the same screen could therefore place the mean excess over the fixed comparator above the margin. That refutation therefore describes this split, and it is the only A_1 cell of which this is true. The GEARS intervals against A_2 include zero on both screens.

Seed pairing. Each screen's runs are trained separately, so a joint count pairs them by seed index. For per-screen counts K_a and K_b out of R, a one-to-one pairing can give any joint count from $\max(0, K_a + K_b - R)$ to $\min(K_a, K_b)$. Across three models, three authorities, two coverage arms and two margins, every case has $K = 0$ or $K = R$ on at least one screen. The range therefore collapses to one value, and all 72 joint verdicts at both reliability thresholds are independent of the pairing.

M Complementary rank-metric sensitivity

This appendix checks whether the GEARS mean advantage over A_2 in Section 6.2 depends on the Pearson score. This post-hoc analysis reuses all sixty retained prediction profiles on the same two screens. For each target, Spearman correlation compares the ranks of the truth and predicted expression deltas across all genes, using average ranks for ties. A constant vector receives score zero, as in the primary scoring convention. Target scores are averaged equally within each run. The tables

Table M.1: Pearson mean excess (training-run standard deviation) on the retained populations. Each cell uses ten runs. N counts targets, not independent biological studies. Covered rows score the model's covered targets. Filled rows score all test targets, using the control profile for uncovered targets. The Ridge column is the training-selected fixed member, which nested training-only selection also chooses. GEARS uses raw counts. STATE and CPA use log expression. These are descriptive effects, not new reliability decisions.

Model, screen	Coverage	N	A_1	A_2	A_3	Ridge
GEARS K562	Covered	403	-0.0280 (0.0166)	$+0.0422$ (0.0166)	-0.2471 (0.0166)	-0.0363 (0.0166)
GEARS K562	Filled	404	-0.0280 (0.0166)	$+0.0420$ (0.0166)	-0.2470 (0.0166)	-0.0362 (0.0166)
GEARS RPE1	Covered	437	-0.0273 (0.0046)	$+0.0150$ (0.0046)	-0.1435 (0.0046)	-0.0329 (0.0046)
GEARS RPE1	Filled	441	-0.0263 (0.0046)	$+0.0155$ (0.0046)	-0.1443 (0.0046)	-0.0319 (0.0046)
STATE K562	Covered	404	-0.1841 (0.0258)	-0.1547 (0.0258)	-0.2190 (0.0258)	-0.1846 (0.0258)
STATE K562	Filled	404	-0.1841 (0.0258)	-0.1547 (0.0258)	-0.2190 (0.0258)	-0.1846 (0.0258)
STATE RPE1	Covered	441	-0.1961 (0.0253)	-0.1869 (0.0253)	-0.2335 (0.0253)	-0.1965 (0.0253)
STATE RPE1	Filled	441	-0.1961 (0.0253)	-0.1869 (0.0253)	-0.2335 (0.0253)	-0.1965 (0.0253)
CPA K562	Covered	404	-0.1818 (0.0626)	-0.1524 (0.0626)	-0.2167 (0.0626)	-0.1823 (0.0626)
CPA K562	Filled	404	-0.1818 (0.0626)	-0.1524 (0.0626)	-0.2167 (0.0626)	-0.1823 (0.0626)
CPA RPE1	Covered	441	-0.4429 (0.0848)	-0.4337 (0.0848)	-0.4804 (0.0848)	-0.4433 (0.0848)
CPA RPE1	Filled	441	-0.4429 (0.0848)	-0.4337 (0.0848)	-0.4804 (0.0848)	-0.4433 (0.0848)

report the mean and sample standard deviation across the ten model-training runs. The standard deviation describes training randomness conditional on a fixed screen, not biological generalization or uncertainty in comparator selection.

The target populations, coverage handling and training-selected comparators are unchanged. The primary arm uses targets covered by every retained run. The filled arm retains all frozen test targets and uses the original control-fill rule for missing predictions. No targets are silently skipped. GEARS is evaluated in raw-count space and STATE and CPA in log space. Recomputed Pearson scores match the frozen per-target scores exactly for every comparator and model run. STATE's stored control profile is not an evaluation target.

A_1 retains the fixed pseudobulk-delta comparator. A_2, the training-selected fixed comparator and the nested training-selected procedure retain their original Pearson-based training choices. None is retuned for Spearman. The training-selected fixed and nested procedures both select ridge in these settings and yield identical scores. A_3 takes the per-target maximum over the same six-member family under the metric being reported. It remains a hindsight diagnostic, not a deployable predictor.

Tables M.1 and M.2 report both coverage arms for every model–screen pair. GEARS's primary-arm mean excess against A_2 changes from $+0.0422$ and $+0.0150$ under Pearson to -0.0648 and -0.0583 under Spearman on K562 and RPE1, respectively. The filled arm has the same sign reversal. STATE and CPA have negative mean excess against A_2 under both metrics, and every model–screen pair has negative mean excess against the strong fixed comparator under both metrics and coverage arms. Mean excess against A_1 and A_3 is also negative throughout.

Zero is only a descriptive sign reference here. Rank agreement does not measure effect-magnitude accuracy, and the registered meaningful-effect margin is not recalibrated or transferred to the new metric. No new reliability verdict, hypothesis family or significance claim is issued. The primary Pearson analysis is unchanged. The GEARS/A_2 sign reversal limits claims of metric-general added value. It does not determine which metric is preferable or show that rank-tuned comparator selection would behave identically.

N Independent-screen test on the Adamson screen

This appendix reports the independent-screen test of Section 6.4. It uses the unfolded-protein-response Perturb-seq screen of Adamson et al. (2016), a separate experiment with its own guide library, target set and processing in the same K562 cell line. The test therefore concerns transfer to a new screen, not a new cell type.

Table M.2: Spearman mean excess (training-run standard deviation), with the same ten runs, targets and coverage rules as Table M.1. Training-selected comparators remain fixed to their Pearson-based choices. All means are negative, including the two GEARS/A_2 comparisons whose Pearson means are positive. Standard deviations do not quantify uncertainty over biological screens or comparator training.

Model, screen	Coverage	N	A_1	A_2	A_3	Ridge
GEARS K562	Covered	403	-0.1162 (0.0126)	-0.0648 (0.0126)	-0.1824 (0.0126)	-0.1148 (0.0126)
GEARS K562	Filled	404	-0.1159 (0.0125)	-0.0646 (0.0125)	-0.1820 (0.0125)	-0.1144 (0.0125)
GEARS RPE1	Covered	437	-0.1130 (0.0176)	-0.0583 (0.0176)	-0.1628 (0.0176)	-0.1145 (0.0176)
GEARS RPE1	Filled	441	-0.1134 (0.0174)	-0.0592 (0.0174)	-0.1636 (0.0174)	-0.1149 (0.0174)
STATE K562	Covered	404	-0.1512 (0.0199)	-0.1277 (0.0199)	-0.1910 (0.0199)	-0.1508 (0.0199)
STATE K562	Filled	404	-0.1512 (0.0199)	-0.1277 (0.0199)	-0.1910 (0.0199)	-0.1508 (0.0199)
STATE RPE1	Covered	441	-0.1692 (0.0183)	-0.1637 (0.0183)	-0.2133 (0.0183)	-0.1698 (0.0183)
STATE RPE1	Filled	441	-0.1692 (0.0183)	-0.1637 (0.0183)	-0.2133 (0.0183)	-0.1698 (0.0183)
CPA K562	Covered	404	-0.1486 (0.0474)	-0.1250 (0.0474)	-0.1884 (0.0474)	-0.1482 (0.0474)
CPA K562	Filled	404	-0.1486 (0.0474)	-0.1250 (0.0474)	-0.1884 (0.0474)	-0.1482 (0.0474)
CPA RPE1	Covered	441	-0.3212 (0.0613)	-0.3157 (0.0613)	-0.3653 (0.0613)	-0.3218 (0.0613)
CPA RPE1	Filled	441	-0.3212 (0.0613)	-0.3157 (0.0613)	-0.3653 (0.0613)	-0.3218 (0.0613)

Data. The version of this screen distributed with GEARS (Roohani et al., 2024) has 68,603 cells and 5,060 genes, and its control labels include perturbed cells (Wong et al., 2025). Control and perturbation labels are rebuilt from the original guide and barcode calls, following the published correction. Thirty cells in fifteen groups whose source identity cannot be resolved uniquely are excluded rather than assigned. The unfolded-protein-response experiment then contains 49,786 cells, including 4,595 controls, and 82 targets with at least twenty cells. The expression values are used as distributed, and their normalization provenance is not certified.

Split and protocol. The 82 targets are split at the target level with a fixed seed into 66 training and 16 test targets, giving 36,091 perturbed training cells and 9,100 test cells. GEARS is trained thirty times with the recipe of the main campaign, twenty epochs and hidden width 64, differing only in the seed. GEARS covers 15 of the 16 test targets. TIMM23, which has no connection in its gene-ontology graph, receives the control profile in the control-filled arm. The six-member family is fitted on the training targets. The primary comparator is the training-selected instance of A_1, the member with the best out-of-fold training score. On this screen it is pseudobulk-delta, the registered member. The registered pseudobulk comparator, the cross-fit selector, the nested training-only selection and the per-target envelope are reported as sensitivities. The margin $\delta = 0.0139$ is carried over unchanged from the Replogle noise floor, with $\rho = 0.75$ and individual $\alpha = 0.0125$. At $R = 30$ support requires $K \geq 28$ and refutation $K \leq 16$.

Scope. The split, comparator rule, thresholds and the two-arm rule were fixed before any Adamson test score was computed. One run had already produced predictions, unscored, when the scoring rule was fixed. Test scores were computed once, after all thirty runs had finished. The thirty runs are repetitions within one biological source, not thirty studies.

Result. Table N.1 reports every comparator in both arms. GEARS clears the margin in none of the thirty runs against any comparator in either arm. On the 15 covered targets its mean excess over the primary comparator is -0.171, the runs lie between -0.319 and -0.092, and the refutation tail is 8.67×10^{-19}. The comparators are close to one another on this screen, with covered-target scores from 0.497 for the selector to 0.508 for the envelope, so the refutation does not depend on the authority. Nested training-only selection chooses the fixed procedure, whose outer training score of 0.4424 exceeds the selector's 0.4385.

Thirty fresh splits. The registered test repeats thirty seeds on one split. A second test was registered before any run on a new split existed. It draws thirty further splits with thirty new seeds by the same procedure (targets sorted, sixteen test targets drawn by NumPy's default generator). Each split has one GEARS run with its own training seed, and every comparator is refitted on that split's training targets. The splits cover 13 to 16 of their 16 test targets. The training-selected member

Table N.1: Independent-screen test on the Adamson screen, $R = 30$ GEARS runs, $\delta = 0.0139$, $\rho = 0.75$, individual $\alpha = 0.0125$. $s(c)$ is the comparator's macro score, $\bar{W}$ the mean run excess over it, and p_+ and p_- the exact upper and lower binomial tails. The covered arm scores the 15 targets GEARS can address, and the control-filled arm scores all 16. The first row of each arm is the primary comparator, and the other rows are sensitivities.

Arm	Comparator	$s(c)$	$\bar{W}$	K/R	p_+	p_-	Verdict
covered (15)	A_1 trained (primary)	0.5022	-0.1713	0/30	1	8.67×10^{-19}	refuted
	A_1 registered	0.5022	-0.1713	0/30	1	8.67×10^{-19}	refuted
	nested selection	0.5022	-0.1713	0/30	1	8.67×10^{-19}	refuted
	A_2 selector	0.4973	-0.1664	0/30	1	8.67×10^{-19}	refuted
	A_3 envelope	0.5080	-0.1772	0/30	1	8.67×10^{-19}	refuted
control-filled (16)	A_1 trained (primary)	0.5153	-0.2051	0/30	1	8.67×10^{-19}	refuted
	A_1 registered	0.5153	-0.2051	0/30	1	8.67×10^{-19}	refuted
	nested selection	0.5153	-0.2051	0/30	1	8.67×10^{-19}	refuted
	A_2 selector	0.5115	-0.2013	0/30	1	8.67×10^{-19}	refuted
	A_3 envelope	0.5216	-0.2114	0/30	1	8.67×10^{-19}	refuted

Table N.2: Adamson multi-split test, thirty fresh target splits with one GEARS run each and every comparator refitted on each split's training targets, $\delta = 0.0139$, $\rho = 0.75$, individual $\alpha = 0.0125$. K is given for the covered and control-filled arms, and $\bar{W}$, min and max are over the thirty runs in the covered arm. Every count is a refutation at $p = 0.25^{30}$.

Comparator	K covered	K filled	$\bar{W}$	min W	max W	Verdict
A_1 training-selected	0/30	0/30	-0.135	-0.313	-0.035	refuted
A_1 registered	0/30	0/30	-0.136	-0.313	-0.035	refuted
A_2 selector	0/30	0/30	-0.135	-0.314	-0.033	refuted
nested selection	0/30	0/30	-0.135	-0.313	-0.033	refuted
A_3 envelope	0/30	0/30	-0.148	-0.326	-0.061	refuted

is mean-shift in 24 splits, ridge in 5 and pseudobulk-delta in 1, and nested training-only selection keeps the fixed member in 15 splits and the selector in 15. No run clears the margin against any comparator of Table N.2 in either arm, and none does at $\delta = 0.0346$ or $\rho = 0.5$ either. The excess over the training-selected comparator averages -0.135 with standard deviation 0.062 across splits, against -0.171 and 0.058 across the thirty seeds of the registered split. Across these thirty draws, changing the target split shifts the mean by much less than the model's deficit and adds little spread.

The linear baseline. The linear gene-embedding baseline of the Replogle tests (Appendix O), fitted on each split's training targets, covers every Adamson test target. On the registered split it scores 0.471 on the covered targets, below every comparator of the family, and GEARS clears the margin against it in none of the thirty seeds, with mean excess -0.140. On the fresh splits one run of thirty clears it, by 0.031, and the mean excess is -0.095. Both counts are refutations, at $p = 8.67 \times 10^{-19}$ and 7.89×10^{-17}. The comparison is post hoc.

O REPLOGLE MULTI-SPLIT TEST ON THE REFERENCE PREPROCESSING

This appendix reports the Replogle test of Section 6.4, registered before any of its splits was built. It changes three things at once relative to the confirmatory campaign. GEARS receives its reference preprocessing instead of raw counts, every run has its own target split, and GEARS and the comparators are fitted on the same training cells. It also adds the linear gene-embedding baseline of Ahlmann-Eltze et al. (2025) as a comparator.

Splits and preprocessing. Each screen has twelve splits drawn by the registered procedure, which sorts the targets with at least twenty cells, shuffles them with the split's seed and holds out the first fifth. With the campaign's own seed, this procedure returns the confirmatory held-out targets on both screens, 404 of 2,021 on K562 and 441 of 2,204 on RPE1. Every cell is scaled to 10^4 counts over the

Table O.1: Replogle multi-split test, twelve fresh target splits per screen with one GEARS run each on its reference preprocessing and every comparator refitted on that split's training cells, $\delta = 0.0139$, $\rho = 0.75$, individual $\alpha = 0.0125$. K is given for the covered and control-filled arms, and $\bar{W}$, min and max are over the twelve runs in the covered arm. Every count is a refutation at $p = 0.25^{12}$.

Comparator	K covered	K filled	$\bar{W}$	min W	max W	Verdict
K562, $R = 12$						
A_1 training-selected	0/12	0/12	-0.082	-0.095	-0.069	refuted
A_1 registered	0/12	0/12	-0.082	-0.095	-0.069	refuted
linear gene embedding	0/12	0/12	-0.100	-0.115	-0.085	refuted
A_2 selector	0/12	0/12	-0.079	-0.092	-0.061	refuted
nested selection	0/12	0/12	-0.082	-0.095	-0.069	refuted
A_3 envelope	0/12	0/12	-0.096	-0.109	-0.085	refuted
RPE1, $R = 12$						
A_1 training-selected	0/12	0/12	-0.097	-0.140	-0.058	refuted
A_1 registered	0/12	0/12	-0.097	-0.140	-0.058	refuted
linear gene embedding	0/12	0/12	-0.101	-0.146	-0.060	refuted
A_2 selector	0/12	0/12	-0.096	-0.140	-0.058	refuted
nested selection	0/12	0/12	-0.097	-0.140	-0.058	refuted
A_3 envelope	0/12	0/12	-0.117	-0.160	-0.075	refuted

gene panel and log-transformed, and the panel is not reduced to highly variable genes. Controls and training targets are subsampled once to 80,000 cells using the GEARS wrapper's own rule and the run's training seed. The rule retains every control, 10,691 on K562 and 11,485 on RPE1. GEARS and every comparator are fitted on exactly these cells, and the evaluation truth comes from 20,000 held-out cells drawn by the registered rule.

Runs and comparators. GEARS is trained once per split with the recipe of the main campaign, twenty epochs and hidden width 64, and a training seed of its own. On each split the six-member family, the training-selected member, the cross-fit selector, nested training-only selection and the per-target envelope are refitted on that split's training cells. The linear baseline follows the authors' code with its defaults. Gene embeddings are the first ten principal components of the training pseudobulk, including the control. A target's embedding is its gene's row, and the bilinear map is fitted by ridge regression with penalty 0.1. A target whose gene lies outside the panel receives the mean training change. The primary comparator is the training-selected member, and at $R = 12$ refutation needs $K \leq 4$, while support is unattainable. All 24 runs are included. Seven of them ran with PyTorch 2.7.1 instead of 2.5.1, which their GPUs required, and their excesses fall inside the range of the other runs on both screens.

Result. The run-level test refutes GEARS against every comparator on both screens in both coverage arms, $K = 0$ of 12 throughout, at $p = 0.25^{12} = 5.96 \times 10^{-8}$ (Table O.1). The training-selected member is ridge on 23 of the 24 splits and mean-shift on one. Against this member, the mean excess is -0.082 on K562 and -0.097 on RPE1, with the best run at -0.069 and -0.058. The verdicts are the same at $\delta = 0.0346$ and at $\rho = 0.5$. The selector A_2 no longer leaves GEARS with a positive mean excess, -0.079 and -0.096, and GEARS scores 0.100 and 0.101 below the linear baseline. With its reference preprocessing, a fresh target split for every run and comparators fitted on the same cells, GEARS is therefore refuted on both Replogle screens.

P STATE WITH TRAINING-DATA GENE FEATURES

This appendix reports a test registered before any of its runs, which places STATE inside its supported regime. In the campaign STATE encodes each perturbation by a one-hot index, so a held-out target reaches weights that received no gradient. Here every target receives a feature vector computed from training cells alone, its gene's row of the linear baseline's embedding (Appendix O). The embedding comprises the first ten principal components of the training pseudobulk with the control included, each scaled to unit standard deviation over the gene panel. The control and any target outside the panel receive the zero vector. The recipe, software versions, log-scale datasets and training

Table P.1: STATE with training-data gene features, $R = 10$ runs per screen, $\delta = 0.0139$, $\rho = 0.75$, individual $\alpha = 0.0125$. K is given for the covered and control-filled arms, and $\bar{W}$, min and max are the run excesses in the covered arm. The last line of each screen compares each run with the registered one-hot run of the same seed on all held-out targets.

Comparator	K covered	K filled	$\bar{W}$	min W	max W	Verdict
K562, $R = 10$						
A_1 training-selected	0/10	0/10	-0.210	-0.258	-0.154	refuted
A_1 registered	0/10	0/10	-0.209	-0.258	-0.154	refuted
linear gene embedding	0/10	0/10	-0.223	-0.271	-0.167	refuted
A_2 selector	0/10	0/10	-0.180	-0.229	-0.125	refuted
nested selection	0/10	0/10	-0.210	-0.258	-0.154	refuted
A_3 envelope	0/10	0/10	-0.244	-0.293	-0.189	refuted
Features minus one-hot, same seed: mean -0.025, range -0.056 to $+0.031$, $n = 10$						
RPE1, $R = 10$						
A_1 training-selected	0/10	0/10	-0.193	-0.245	-0.155	refuted
A_1 registered	0/10	0/10	-0.193	-0.245	-0.154	refuted
linear gene embedding	0/10	0/10	-0.195	-0.247	-0.156	refuted
A_2 selector	0/10	0/10	-0.183	-0.236	-0.145	refuted
nested selection	0/10	0/10	-0.193	-0.245	-0.155	refuted
A_3 envelope	0/10	0/10	-0.230	-0.282	-0.192	refuted
Features minus one-hot, same seed: mean $+0.003$, range -0.031 to $+0.037$, $n = 10$						

seeds 1 to 10 are those of the registered STATE runs, and the feature file is the only change. Each run is scored against the campaign's comparators and the linear baseline in both coverage arms. At $R = 10$ refutation needs $K \leq 3$ and support is unattainable. No run is excluded.

Result. The run-level test refutes STATE against every comparator on both screens in both coverage arms, $K = 0$ of 10 throughout (Table P.1). The mean excess over the registered fixed comparator is -0.209 on K562 and -0.193 on RPE1, with the best run at -0.154 on both, and the envelope lies further away. Relative to the registered one-hot run of the same seed, the gene features change the score by -0.025 on K562 and $+0.003$ on RPE1 on average. Placing held-out targets inside STATE's supported regime therefore leaves its refutation unchanged.

Q THE LINEAR BASELINE AS THE PREDICTOR UNDER TEST

This appendix reports a test registered before any of its runs, which applies the rule to a real predictor whose mean excess over the registered fixed comparator lies near the margin. The predictor is the linear gene-embedding baseline of Appendix O. Each run refits the baseline on a bootstrap resample of the registered training cells, drawn with replacement at the original sample size, using the registered control reference. Each of the four registered screen and scale settings has thirty runs. The comparators are the campaign's, fitted once on the registered training set. The primary arm scores the held-out targets whose gene is in the panel, 371 on K562 and 386 on RPE1. The second arm scores all held-out targets. At $R = 30$ support needs $K \geq 28$ and refutation $K \leq 16$.

Result. Table Q.1 shows all three verdicts. Against the training-only selector A_2 the baseline is supported in three of the four settings, with every run clearing the margin and mean excess from $+0.047$ to $+0.082$. Against the registered fixed comparator the same runs are refuted in raw counts on both screens, including RPE1, where the mean excess of $+0.0136$ lies just below δ and 12 of 30 runs clear it. On K562 in log space the verdict is inconclusive, with mean excess $+0.0147$ just above δ. The margin is cleared by 22 of 30 runs in the primary arm, an inconclusive count, and by 13 of 30 in the second, a refutation, so the arms disagree. Against training-selected ridge the verdicts are the same, and against the envelope every setting is refuted. Bootstrap runs vary little, with standard deviations of 0.0009 to 0.0029, so a count leaves its extremes only when the mean excess lies within a few standard deviations of δ. The test illustrates the rule near the margin and a supported verdict against a weak selector but not against the fixed comparator. It does not show a verdict that differs between the two strong fixed comparators, which agree in every setting.

Table Q.1: The linear gene-embedding baseline as the predictor under test, $R = 30$ bootstrap runs per registered dataset, $\delta = 0.0139$, $\rho = 0.75$, individual $\alpha = 0.0125$. K is given for the primary arm, the held-out targets whose gene is in the panel, with the final verdict ($+$ supported, $-$ refuted, ? inconclusive), which requires agreement with the arm over all held-out targets.

Screen	Scale	A_1 registered	training-selected	A_2	nested	A_3
K562	raw	0/30 $-$	0/30 $-$	30/30 $+$	0/30 $-$	0/30 $-$
RPE1	raw	12/30 $-$	0/30 $-$	30/30 $+$	0/30 $-$	0/30 $-$
K562	log	22/30 ?	20/30 ?	30/30 $+$	20/30 ?	0/30 $-$
RPE1	log	0/30 $-$	0/30 $-$	4/30 $-$	0/30 $-$	0/30 $-$

Table Q.2: The sixty registered runs against the linear gene-embedding baseline, fitted on each registered training set, $\delta = 0.0139$, $\rho = 0.75$, individual $\alpha = 0.0125$. $s(c)$ is the baseline's score on the model's covered targets, and $\bar{W}$ the mean run excess over it. GEARS is scored in raw counts and the other models in log space, as registered. This comparison is post hoc.

Model	Screen	$s(c)$	$\bar{W}$	K covered	K filled	Verdict
GEARS	K562	0.1898	-0.0327	0/10	0/10	refuted
GEARS	RPE1	0.4070	-0.0411	0/10	0/10	refuted
STATE	K562	0.2425	-0.1976	0/10	0/10	refuted
STATE	RPE1	0.3957	-0.1979	0/10	0/10	refuted
CPA	K562	0.2425	-0.1953	0/10	0/10	refuted
CPA	RPE1	0.3957	-0.4447	0/10	0/10	refuted

The linear baseline on the registered splits. The same baseline is fitted on each of the four registered training sets, one per screen and scale, and the sixty registered runs are counted against it with the registered rule, post hoc and outside the registered family. Its gene panel contains the genes of 371 of the 404 K562 and 386 of the 441 RPE1 held-out targets. The remaining targets receive the mean training change. Every model is refuted in both coverage arms (Table Q.2). The baseline scores slightly below training-selected ridge on K562 in raw counts, 0.1898 against 0.1934 on the targets GEARS covers, and above it in the other three settings, so it is the strongest fixed comparator in three of the four.

R Run-level positive control on the Norman screen

This appendix reports the positive control of Section 6.5, registered with its decision rule before any run-level score existed. It shows that the rule supports a predictor with a real advantage over the registered family and withdraws that support once a comparator explaining the advantage is registered. It is a method control, not a confirmatory claim, because the Norman screen was excluded from the confirmatory claim before the campaign.

Data and predictor. The earlier study's frozen split of the Norman K562 genetic-interaction screen (Norman et al., 2019) has 71 held-out perturbations and 5,045 genes in log-transformed expression, and no new split is drawn. A perturbation's feature is the sum of its genes' 512-dimensional standard-normal embeddings. Ridge regression with penalty 10 on standardized features maps this sum to the mean shift and is fitted to the training perturbations. Each of the thirty runs draws a new embedding for every gene, so runs differ only in the random feature map. Because the feature of a pair is the sum of its genes' vectors and the head is linear, the head can learn each training gene's marginal effect and predict a held-out pair as their sum.

Comparators and rule. The four comparators are the registered fixed comparator, the envelope over the registered six-member family, an additive-identity baseline alone, and the envelope over the registered family plus that baseline. The additive-identity baseline is a ridge regression, with penalty 1, of the training perturbations' mean shifts on a one-hot gene-membership vector, so it predicts a held-out pair as the sum of its genes' marginal effects. With one environment $W_r = V_r$, and the rule uses $\delta = 0.0139$, $\rho = 0.75$ and individual $\alpha = 0.0125$ with $R = 30$.

Table R.1: Run-level positive control on the Norman screen, $R = 30$ runs of a random-feature ridge head, $\delta = 0.0139$, $\rho = 0.75$, individual $\alpha = 0.0125$. $s(c)$ is the comparator's mean per-target score, $\bar{V}$ the head's mean excess over it, and p_+ and p_- the exact upper and lower binomial tails.

Comparator	$s(c)$	$\bar{V}$	K/R	p_+	p_-	Verdict
A_1 registered	0.5150	$+0.2725$	30/30	1.79×10^{-4}	1	supported
A_3 registered family	0.5361	$+0.2514$	30/30	1.79×10^{-4}	1	supported
additive identity	0.8000	-0.0125	1/30	1	7.89×10^{-17}	refuted
A_3 family + additive	0.8486	-0.0611	0/30	1	8.67×10^{-19}	refuted

Agreement with the earlier study. With the earlier study's deterministic embedding, this implementation reproduces its published values, a median per-target excess of 0.2757 over the earlier five-member envelope with 63 of 71 targets positive, and a median delta correlation of 0.8937 for the additive-identity baseline.

Result. The head's mean score over the thirty runs is 0.787, against 0.515 for the fixed comparator, 0.536 for the registered envelope and 0.800 for the additive baseline (Table R.1). The comparators are fixed, so every excess varies across runs only through the head, whose score has standard deviation 0.0138. The head clears the margin against the registered family in every run, so the thirty-run test supports it. Once the additive baseline, which encodes the genes' marginal effects, is registered, the test refutes the head. The earlier study observed the same pattern with one embedding table. This control therefore shows the rule's behavior on a known positive, rather than discovering one.

S CALIBRATION AND POWER DIAGNOSTICS

The following constructions belong to the earlier battery, whose completed results appear in the earlier-study report in the supplementary archive. Two are evaluated below under the run-level rule on real expression. The battery's operating characteristics were measured on controlled predictors with known ground truth, namely a *shared-effect* predictor (shared, environment-invariant effects, the setting of Peters et al., 2016 and Arjovsky et al., 2019), an environment-specific *shortcut* predictor (accurate in one environment, noise elsewhere), a pure-*noise* predictor, a negative method control by analogy with Lipsitch et al. (2010), and a *stable-confounder* predictor (an environment-invariant axis shared by the effects, with no perturbation-specific resolution). The earlier-study report gives the battery's false-positive and true-positive rates across seeds at $\alpha \in \{0.05, 0.1, 0.2\}$. That sweep characterizes the battery's operating curve and does not change the decision rule. Every verdict in this extension uses the registered level $\alpha = 0.0125$ per individual directional test, fixed in Section 5. The separate simultaneous sensitivity uses Holm at 0.05. Section 6.5 identifies which rates were measured for the earlier battery and which remain unmeasured for the new rule.

Constructed predictors under the new rule. Two of the four constructions above were evaluated on the frozen screens with the identical scoring path and prediction format, without a separate evaluation route. The noise arm predicts the control mean plus a random direction at the median norm of the training deltas, and so carries no skill by construction. The stable-confounder arm uses the dominant shared axis of the training per-perturbation deltas. Each target's loading is obtained from its own control cells and extrapolated by a map fitted on training targets alone. The predictor is therefore environment-invariant and correlated with the response but has no perturbation-specific resolution. Both constructions are specified to use training responses and evaluation control profiles only. The construction code verifies that the evaluation response array is unmodified, which does not by itself show that the array was unread. The shared-effect and shortcut constructions are not reported, because expression magnitude and cell cycle are members of the registered family, and scoring one against an envelope that contains it returns a non-positive excess by definition.

Noise runs draw new random directions, and confounder runs bootstrap the training perturbations. Across both arms, both screens, both value spaces and all three authorities, $K = 0$ of 10 in all 24 cells at $p = 9.54 \times 10^{-7}$ each, and the run-to-run standard deviation ranges from 0.0004 to 0.0425. A count of zero in every setting bounds the battery rather than estimating a rate. Table S.1 therefore reports the best of the ten runs relative to the margin, in units of δ.

Table S.1: Shortfall of the best of ten runs from the registered threshold, in units of δ, for two constructed predictors scored through the identical path. Every cell is refuted at $K = 0$ of 10. The single cell within one δ is the stable confounder against the training-only selector on K562 in raw counts. Four of its ten runs score above that selector outright (largest $+0.00545$), so the registered margin, rather than the comparator score alone, prevents a success.

Constructed predictor	Screen and space	A_1	A_2	A_3
noise	K562, raw counts	14.2	9.2	30.0
noise	K562, log	17.4	15.3	19.9
noise	RPE1, raw counts	28.9	25.9	37.4
noise	RPE1, log	29.3	28.6	32.0
stable confounder	K562, raw counts	5.6	0.6	21.4
stable confounder	K562, log	3.7	1.5	6.2
stable confounder	RPE1, raw counts	8.2	5.2	16.7
stable confounder	RPE1, log	18.9	18.2	21.6

Real-data power diagnostic (post hoc). A graded predictor that mixes mean-shift with an oracle read from half of each target's test cells probes real-data power after scoring. Its description and results are in the earlier-study report of the supplementary archive, and the same construction is applied to the new protocol below.

S.1 Synthetic latent-signal diagnostic

This diagnostic tests positive detection under a specified synthetic expression generator, not biological calibration or a learned predictor. Each of 256 independently generated studies has 32 targets and 64 genes. A centered unit vector b represents the shared effect, and each target has a centered unit signal t_j orthogonal to b. The truth is $b + t_j$, the comparator is b, and the oracle prediction in run r is

$$\widehat{y}_{jr} = b + a t_j + \sigma_r \epsilon_{jr}, \qquad \log \sigma_r \sim \mathcal{N}(-0.7, 0.8^2).$$

Noise directions ϵ_{jr} are centered unit vectors orthogonal to both b and t_j. The predictor is given the latent target signal. It is not fitted to data. Each run uses the same correlation calculation as the primary analysis, averaging its excess over the shared-effect comparator equally across targets. A run succeeds only above $\delta = 0.0139$, and the existing rule uses $\rho = 0.75$ and $\alpha = 0.0125$.

The five signal amplitudes and noise distribution were fixed before this diagnostic was run. Random directions and noise magnitudes are shared across amplitudes within each study. The ten-run experiment uses the first ten runs of the thirty-run stream. Table S.2 reports every amplitude and outcome. At the strongest prespecified amplitude, $a = 1$, only 80 of 256 studies are supported at $R = 30$, that is 31.25%, with an exact binomial 95% Monte Carlo interval of $[25.62\%, 37.32\%]$ across independent synthetic studies. None is supported at $R = 10$, where support is unattainable under this rule even if every run succeeds. Because all directions are orthogonal unit vectors, a run's excess has the closed form $2^{-1/2}\{(1 + a)/(1 + a^2 + \sigma_r^2)^{1/2} - 1\}$. A run therefore succeeds with probability 0.896 at $a = 1$, and support would be expected in 38% of studies. The simulated runs succeed at a rate of 0.890, for which 34% is expected, and the observed 31% lies within one Monte Carlo standard deviation of it. These frequencies are conditional on the generator, not estimates of biological power or error rates. In particular, the zero-signal arm is not calibrated to the boundary null $q = \rho$, and the positive arm does not establish uniformly high detection power.

S.2 Real-expression controls on the frozen splits

The separate graded-oracle probe uses the two confirmatory environments, the frozen splits and the frozen scoring path. Each target's evaluation cells are partitioned at random into an A half available to the oracle and a disjoint B half used for all scoring. A target enters a split only if it has at least four evaluation cells in total, with at least two cells per half and per nested oracle subset. We draw 100 splits per setting with fixed seeds. The oracle's nested budgets are 0.25, 0.5 and 1.0 of the A pool, so headroom can be assessed as a function of the available cells. Every comparator delta is recomputed on the same rows it is scored on, because the control reference is not exactly constant within a

Table S.2: Synthetic latent-oracle diagnostic. Each amplitude has 256 independent synthetic studies, with paired ten- and thirty-run streams. Counts are supported, refuted and inconclusive studies under the same individual decision rule. Targets and runs are not counted as independent studies, and the predictor is an oracle rather than a trained biological model.

Amplitude	$R = 10$			$R = 30$		
	Supported	Refuted	Inconclusive	Supported	Refuted	Inconclusive
0	0	256	0	0	256	0
0.1	0	110	146	0	250	6
0.25	0	10	246	0	47	209
0.5	0	1	255	10	0	246
1	0	0	256	80	0	176

target and the raw-count predictions are stored in single precision. This scoring path reproduces the retained per-target comparator and model scores for all ten seeds exactly, the macro statistic of the registered verdict to 4.5×10^{-7} and the frozen path to 1.0×10^{-8}. Randomizing the held-out truth leaves every probe delta unchanged. The half-split match was planned at a tolerance of 10^{-12} and checked at 10^{-7}, because the two paths accumulate sums in a different order.

Headroom. Table S.3 reports the pure oracle at the largest A half and its extrapolation to unlimited cells per target, the intercept of a regression of the single-screen excess V on the reciprocal of the oracle's cell budget across the three nested fractions (r^2 between 0.929 and 0.970). The oracle sees about 24.5 cells per target on K562 and 22.4 on RPE1, roughly half of a median target. The fitted trajectory suggests additional headroom at larger cell budgets but does not measure the full-budget outcome. Against the fixed comparator the extrapolated advantage exceeds δ in all four settings. Against the same comparator on RPE1 in log space, the largest measured oracle budget still gives -0.003557 excess, and the positive extrapolation is not an observed positive result. Against the in-sample envelope it exceeds δ only in raw-count space. On the same halves the models score below the fixed comparator in every setting, from -0.028 to -0.377. A control that deranges the mapping between a target's oracle estimate and its scored half returns -0.14 to -0.22 against the fixed comparator, as it must.

Selection share and structured padding. Figure J.1 collects the diagnostics below. The envelope's advantage over the fixed comparator is measured again by choosing its member on one half of a target's cells and scoring that choice on the other. The selection share is the fraction of the in-sample advantage that this cross-fitting removes, $(A_3 - A_3^{\mathrm{cf}})/(A_3 - A_1)$, so a share at or above 1 leaves the cross-fit envelope no better than the fixed comparator. The selection share is 0.81 on K562 in raw counts and at or above 1 in the other three settings. The reverse direction agrees to within 0.04. Extrapolation to an unlimited selection half, with the scoring half fixed, gives values between 0.77 and 1.12. Enlarging the family from six members to every training perturbation, using each one's pseudobulk delta as a candidate, raises the per-target envelope by 0.099 to 0.428 (Table S.4). A training delta is the binding member on 82.3 to 98.4 percent of targets, and every model–screen cell is still refuted at $K = 0$ of 10.

Variance of the run-level statistic. The macro statistic over ten training seeds and the resampled halves has a between-split standard deviation of 0.0006 to 0.0117 and a between-seed standard deviation of 0.0052 to 0.0681. The seed term carries 63 and 16 percent of the total variance for GEARS on K562 and RPE1 and 98 to 100 percent for STATE and CPA. A constructed predictor's success frequency over $R = 10$ runs is therefore degenerate on these data, not informative about the rule's operating characteristics. The four-construction calibration is consequently reported as headroom, not empirical error rates.

Figure J.1 in Appendix J visualizes these diagnostics. The tables below report the corresponding numerical values and the scope of each extrapolation.

Table S.3: Headroom probe on the frozen splits. V is the oracle's macro advantage over the named comparator authority within one screen, measured at the largest oracle budget and extrapolated to unlimited cells per target. A_3^{cf} is the cross-fit envelope. The selection share is the fraction of the envelope's in-sample advantage over the fixed comparator lost under cross-fitting, $(A_3 - A_3^{\mathrm{cf}})/(A_3 - A_1)$.

Screen	Cells	V vs A_1	ext.	V vs A_3	ext.	ext. A_3^{cf}	share
K562 raw	24.5	$+0.260$	$+0.289$	$+0.036$	$+0.065$	$+0.239$	0.81
RPE1 raw	22.4	$+0.192$	$+0.230$	$+0.058$	$+0.096$	$+0.232$	1.12
K562 log	24.5	$+0.018$	$+0.036$	-0.020	-0.002	$+0.041$	1.29
RPE1 log	22.4	-0.004	$+0.026$	-0.046	-0.017	$+0.021$	1.13

Table S.4: Structured padding of the comparator family with every training perturbation's pseudobulk delta. The padded envelope is higher in every row, a training delta is the binding member on most targets, and no verdict changes.

Screen and space	Model	A_3	padded	training delta wins	K	verdict
K562 raw	GEARS	0.404	0.832	97.5%	0/10	refuted
RPE1 raw	GEARS	0.509	0.842	98.4%	0/10	refuted
K562 log	STATE	0.264	0.393	93.8%	0/10	refuted
K562 log	CPA	0.264	0.393	93.8%	0/10	refuted
RPE1 log	STATE	0.431	0.531	82.3%	0/10	refuted
RPE1 log	CPA	0.431	0.531	82.3%	0/10	refuted

T Relation to multiverse and specification-curve analysis

Reporting a result across defensible analytic choices that could have been made, rather than under just one choice, is not new. Steegen et al. (2016) call it multiverse analysis, and Simonsohn et al. (2020) give the specification curve, in which a result is plotted against every defensible specification and the distribution of outcomes is the finding. Our analysis applies that method to one degree of freedom.

The restriction to one degree of freedom is deliberate. A full multiverse over metric, aggregation unit, split and comparator would produce a distribution that no ten runs per cell could resolve, and it would not carry an error guarantee. Fixing everything except the comparator defines the axis. An ordered error rate additionally needs a specified nested sequence and a valid testing procedure. The registration fixes three unordered authorities, not a six-rung sequence. The completed six-member cumulative display and all 63 nonempty subsets are therefore post-hoc descriptive analyses, without an ordered-testing guarantee. The result therefore concerns only one axis, not the full space of analytic choices.

A specification curve also summarizes the spread of results across defensible specifications. Comparator authorities are not interchangeable because they differ in what a practitioner could actually deploy. A_1 is available at prediction time and A_3 is not, so the two ends of our axis carry different operational meanings rather than being two equally valid analyst choices. That is why we report a verdict per authority rather than a distribution over authorities.

Order of registration. The error control of Proposition 1 is exact only for scoring choices fixed before the outcomes they could influence, and Corollary 1 extends it to the expression scale fixed after 17 of the sixty runs. Table T.1 lists the registration records in order, with the outcomes already observed and the choices each fixed. Each inference can thus be matched to the choices that preceded it. Analyses with no registration record, namely the training-selected ridge comparator, the Holm family, target resampling, the run-level rule under disjoint control references and the rank metric, are post hoc and are reported as such.

Table T.1: Registration records in the order they were written. The middle column lists the outcomes available when each record was written. Every record except row 6 was timestamped externally when written.

	Observed when written	Fixed
1	Development-split GEARS runs only	Metric, δ, ρ, α, R, family, fixed A_1 member, splits, models, coverage rule
2	Three one-epoch pipeline runs on the test split and model-free comparator scores	No choice changed
3	17 of 60 confirmatory runs	Per-model expression scale, three control arms
4	Four of six cells	The two readings of the GEARS verdict
5	GEARS verdicts and the first A_2 results	The post-hoc reading of A_2
6	One Adamson run, unscored	Adamson split, comparator rule, thresholds
7	The earlier study's published values only	Norman run-level positive control
8	Stronger selector, training side only	Its test scoring and the nested choice
9	Registered Adamson result	Adamson multi-split test
10	Every registered result, including the Adamson multi-split test	Replogle multi-split test with reference preprocessing and the linear baseline
11	No multi-split run finished, the linear baseline's post-hoc refutations	STATE with training-data gene features
12	The linear baseline's scores on the registered datasets	The linear baseline as the predictor under test

U ADDITIONAL CONTEXT FOR THE COMPARATOR AUTHORITIES

Operational context for the endpoints. The fixed and oracle-max endpoints correspond to comparator choices discussed in separate literatures.

At the upper end, Equation 2 is the envelope of a family of competing models, and tests of a method against that envelope are established outside machine learning. White (2000) asks whether the best model encountered in a specification search truly beats a benchmark, and builds a null distribution that accounts for the search having taken a maximum. Hansen (2005) recovers power with a studentized statistic and a sample-dependent null distribution. These results establish selection inflation for a maximum over searched models, and a per-target maximum over a baseline family can be inflated in the same way, as the disjoint-cell diagnostic measures. None of the perturbation-prediction evaluations cited here applies that correction.

At the lower end, A_1 is stronger than the comparators that the evaluation literature documents in practice. Leech et al. (2024) catalog questionable research practices and name two that act directly on the comparator. *Baseline hacking* is choosing weak comparators to compare against, which their table marks as not plausibly accidental, and *baseline nerfing* is tuning a competitive comparator less than one tunes one's own method, which they note is easy to do unintentionally. Where weak-baseline use has been measured, its rate is high. Reviewing articles that use machine learning on fluid-related partial differential equations and claim to beat a standard numerical method, McGreivy & Hakim (2024) find that 79% of them, 60 of 76, compare against a weak baseline.

A_1 and A_3 therefore bracket comparator choices discussed in different bodies of work. This does not imply that any particular paper computes Equation 2, nor does it constitute an audit of published baseline tables. The claim is that added value depends on comparator authority, and both endpoint choices have precedents in the literature.

V EXTENDED RELATED WORK

Run-level evidence standards for benchmarks. The quantity we test, the probability that a single complete training run delivers added value, comes from earlier work. Agarwal et al. (2021) define the average probability of improvement and, on Atari 100k, find that confidence intervals on sample medians need closer to 50–100 runs than to 20 or 30. Bouthillier et al. (2021) recommend a threshold of $\gamma = 0.75$ on $P(A > B)$ and, under their testing procedure, a sample size of $N = 29$ to detect it, and we adopt that threshold as the primary one. We claim neither ingredient, and Section 6.5 gives the exact budgets that the registered rule requires.

Relation to run-level evaluation. Three features distinguish our statistic from that standard, and only the third is a contribution rather than an adaptation. First, the aggregation is conjunctive rather than marginal. The average probability of improvement rewards a run that wins on some tasks and loses on others. We instead take the worst case over environments, so a run counts only if the model is ahead everywhere at once. Second, we threshold the magnitude of the improvement at a pre-registered δ, addressing a limitation Agarwal et al. (2021) state of their own metric, namely that it does not account for the size of an improvement. Third, the paper's central construction indexes every verdict by comparator authority, the information and rule used for selection. The same frozen runs are judged against a fixed comparator, a training-only selector and a per-target maximum over the registered family, and each authority receives its own verdict rather than one label.

These authorities are an unordered axis and not rungs of a nested ladder. A maximum over one realization is monotone in the registered family, whereas a comparator chosen by a fitted rule is not. Section 4 states the scope and Appendix A gives the counterexample.

Three-state verdicts and ordered error control. Neither ingredient of the decision rule is new and we do not present them as such. Margin-based decisions with an inconclusive outcome are standard. Two one-sided tests declare equivalence within prespecified bounds (Schuirmann, 1987; Lakens, 2017), and a Bayesian region of practical equivalence separates practically null from practically relevant differences (Benavoli et al., 2017). Our three verdicts apply the same idea to a one-sided hypothesis about runs. Controlling directional error along a nested family is the sequential rejection principle of Goeman & Solari (2010), under which a fixed-sequence test on a monotone family that meets its conditions needs no per-layer penalty. Our use of these methods is an application. The power gain of the ordered form over a Holm correction on the same counts remains unmeasured (Section 7).

Algorithm selection and dynamic selection. The three authorities have direct counterparts in algorithm selection, which reports a per-instance selector against the single best solver and the virtual best solver and summarizes it by the fraction of the gap between them that it closes (Rice, 1976; Xu et al., 2008; Bischl et al., 2016; Lindauer et al., 2019; Kerschke et al., 2019). Dynamic classifier selection uses the per-sample oracle in the same way, as an optimistic reference rather than an attainable target (Cruz et al., 2018). In that literature the portfolio is the proposed method, whereas here it supplies the comparators. The question is whether a fixed predictor still has added value once the comparator's position between these references is stated. The stronger selector of Appendix I is a standard per-instance selector with a single-best fallback, and nested training-only selection follows the nested cross-validation that reduces selection bias in estimated error (Varma & Simon, 2006; Cawley & Talbot, 2010).

Selection budgets and seeds. Domain generalization ties each reported result to the data its model selection could use, from training-domain validation to a test-domain oracle, and finds that carefully selected empirical risk minimization matches the methods it was compared with (Gulrajani & Lopez-Paz, 2021). Language modeling and generative modeling found that gains shrink once baselines receive the same tuning budget (Melis et al., 2018; Lucic et al., 2018), and Dodge et al. (2019) ask that the budget be reported with each comparison. Those studies index the selection procedure of every method, whereas we index only the comparator's. The number of seeds needed to compare trained agents is analyzed by Colas et al. (2018), and counting wins with a sign test across data sets is a documented option (Demšar, 2006), where the units are independent data sets rather than training runs. Giacomini & White (2006) make the forecasting method, including its estimation scheme and information set, the object of comparison, which is the econometric counterpart of indexing a comparator by its authority.

Data snooping and the maximum over a model family. Our per-target maximum has precedents in the literature. White (2000) shows that the maximum of a specification search is biased upward by the search itself, and builds a bootstrap null that accounts for having taken the maximum. Hansen (2005) recovers power with a studentized statistic and a sample-dependent null distribution. Our A_3 takes that maximum target by target, which is at least as large as the maximum of the members' average scores, and we do not claim this bootstrap methodology.

The distinction is the quantity of interest. Those procedures treat the inflation as a nuisance to be removed, so that a valid test of whether the best of a family beats a benchmark can be recovered. We instead vary the information available to the comparator and ask whether the verdict about a fixed predictor survives that change. The evaluation-practice literature discussed in Appendix U shows the comparator is pushed in the opposite direction at least as reliably, which is why its position cannot be assumed and has to be reported.

Evaluation of perturbation prediction. Whether deep perturbation predictors add value over simple baselines is an open disagreement in this field rather than a settled question. Ahlmann-Eltze et al. (2025) report that they do not yet beat linear baselines, and Miller et al. (2025) report gains for deep models over mean, control and linear baselines under calibrated metrics. The second paper attributes the disagreement to the metric, finding that mean squared error and control-referenced delta correlation are often poorly calibrated while weighted and rank-based alternatives are not. Nicol et al. (2026) criticize the same metric on different grounds, arguing that a shared control reference inflates it. The metric is therefore already understood in this field to be a choice that changes the answer.

Baseline choice is already studied. Systema (Viñas Torné et al., 2026) emphasizes perturbation-specific effects and compares perturbed-mean and matching-mean baselines, and Vollenweider & Bühlmann (2026) group baselines by the data they may use, unlearned, learned on training data or oracle, and use them as performance floors beside metric diagnostics and empirical bounds. Their grouping parallels the information-access axis of our authorities, which also differ in selection rule and therefore do not form a nested sequence. In its reported evaluations, the grouping is not tied to a decision about whether a trained model beats a floor, and repeated training runs are not the inference unit. Our rule adds that decision and unit. scPertEval (Schäfer et al., 2026) decomposes representation, metric, transformation and reporting choices. The in-the-wild benchmark shows that rankings also depend on the task and metric (Mao et al., 2026). We ask a narrower question. For a fixed predictor, score and margin, does a run-level added-value claim survive a declared change in the comparator's information and selection rule? Our tests index each verdict by that rule. Section 6.1 reports the observed effect and its limits.

W Relation to the earlier five-context study

This work extends an earlier battery that tested excess over registered observational baselines with one training seed per configuration, using targets or Reactome clusters as units. That study supported none of 15 GEARS, STATE and CPA model-context pairs, nor a registered prospective transfer prediction on Jurkat. Its methods, proofs, results and controls are retained in full in a separate earlier-study report in the supplementary archive, including the foundation-model head with its random-embedding and additive-identity controls and the pathway-cluster analyses. The two studies use different inference units, so their counts and p-values are not pooled, and the earlier calibration at $\alpha = 0.05$ does not validate the run-level rule at $\alpha = 0.0125$ used here. The earlier cross-fitted envelope selects on half of a target's held-out cells, whereas A_2 selects on training targets only, so the two are reported separately.

Contribution of the run-level study. Three observations come from the earlier study and are not claimed as findings of this one. Deep models did not beat an envelope of simple baselines, whose composition determined the difficulty of the comparison. An additive-identity baseline revoked the apparent support of the foundation-model head, which a random embedding had matched. New here are the comparator authority as a verdict index with a deployable member, a selector and an oracle, the run-level unit with its exact binomial rule and design analysis, the sixty-run campaign with its selection adjustment, both Adamson tests, the Replogle multi-split test on the reference preprocessing, STATE with training-data gene features, the linear baseline as the predictor under test, the thirty-run positive control, the linear gene-embedding and stronger-selector controls, and the target-resampling and disjoint-reference analyses.