THE CONSTRUCT IS PART OF THE TARGET: RETHINKING PERTURBATION-MODEL EVALUATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Reagent-level variation is biologically established, but perturbation-model benchmarks typically evaluate predictions at the gene level. We ask whether model evaluation changes depending on which reagent is used to perturb a gene and whether that reagent is available in source data. All 2,053 genes observed in all four core CRISPRi panels reuse at least one complete construct across all four contexts, so cellular context holdout alone does not imply construct holdout. Alternative constructs produce recurring response differences beyond matched cell resampling; external CRISPR KO and CRISPRi libraries also show differences beyond same guide resampling. Holding predictions fixed, changing the realized target changes absolute model scores; released State predictions also show sensitivity to target realization beyond same construct resampling. Under global construct exclusion and matched cell counts, all 16 comparisons show higher mean Pearson when the target construct is present in source data than when it is withheld. Construct transfer also changes the relative MAE performance of the two fitted predictors, while same construct resampling does not reproduce the shift. Their mean Pearson ordering also reverses descriptively in all four contexts. A State architecture trained from scratch under the same exclusion design shows material Pearson degradation in RPE1 and Jurkat. Perturbation-model evaluation therefore requires specifying both target realization and which constructs are present in source data.

1 INTRODUCTION

A perturbation benchmark names a gene, but scores a prediction against an experimentally realized response. Different reagents targeting the same gene can produce different realized responses. Here, construct means the complete experimental reagent: a dual-sgRNA cassette in the core CRISPRi panels and a single guide in the external screens. The central question is whether changing the experimental realization of the same gene changes the conclusion drawn from model evaluation.

Evaluation depends separately on which construct realizes the target and which constructs are present in source data. Cellular context holdout alone does not imply construct holdout: all 2,053 genes observed in all four core panels reuse at least one complete construct across all four contexts. Our strict comparison keeps the gene represented through one construct while withholding another complete construct from all source data.

Construct-associated target differences recur across the core panels and exceed same guide resampling in external CRISPR KO and CRISPRi libraries. With predictions fixed, changing the realized target changes absolute model scores; under global construct exclusion, it also changes the relative evaluation of fitted predictors sharing the same gene features (Figure 2). At matched cell counts, changing the target produces a positive shift in the paired MAE gap. The two predictors also show a descriptive reversal in mean Pearson ordering across all four contexts. A State architecture trained from scratch under global construct exclusion also shows material Pearson degradation in RPE1 and Jurkat.

We make three contributions. First, we establish that construct-associated target differences recur and extend beyond same guide resampling (§3). Second, we show that target realization changes model scores for fixed predictions, including released State predictions (§4 and §5). Third, we show that construct transfer changes relative predictor performance under global construct exclusion, while the

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.

same construct control does not reproduce the shift. A State architecture trained from scratch also shows material target sensitivity under the same design (§5).

Relation to prior work. Guide efficacy and off-target activity are established concerns in CRISPR screens (Allen et al., 2019; Meyers et al., 2017; Rohatgi et al., 2024). Benchmarking and metric calibration address generalization and scoring choices (Wu et al., 2025; Wei et al., 2026; Miller et al., 2025). Systema evaluates prediction beyond systematic response variation (Viñas Torné et al., 2026), while reliability analyses relate perturbation quality to model comparison (Wang et al., 2026). Here, we retain the gene label and separate which construct realizes the scored target from whether that construct is present in source data. CPA, GEARS and scGPT provide the broader modeling context (Lotfollahi et al., 2023; Roohani et al., 2024; Cui et al., 2024); we evaluate State both through released predictions (Adduri et al., 2026) and through training from scratch under global construct exclusion.

2 EXPERIMENTAL TARGETS AND EVALUATION CONTRACTS

Reagent formats and response genes. The analyses use two reagent formats. Core comparisons use complete dual-sgRNA cassettes; the external screens use distinct single-guide reagents. Core analyses and State trained from scratch use the shared 6,642-gene response axis. External screens use their own control-defined 2,000-gene spaces, while released State predictions use the release's 2,000 output coordinates. Each result specifies its response-gene set and cohort.

Panels and unit. The CRISPRi panels are K562 and RPE1 (Replogle et al., 2022), and Jurkat and HepG2 (Nadig et al., 2025). All core profiles use the same 6,642 scored genes. All 2,053 perturbation genes observed in all four panels reuse at least one complete construct across all four contexts; 2,047 retain identical complete construct sets, usually singletons. Thus cellular context holdout alone does not imply construct holdout. The global construct exclusion analysis applies no efficacy filter. The CRISPRa panel (Norman et al., 2019) serves as a positional control; its only distinct protospacer case, IER5L, is reported in Appendix D. The principal construct comparison requires at least two distinct sgID_AB cassettes per perturbation, leaving 118/2,057 K562 and 135/2,393 perturbations in each other CRISPRi panel. Population and preparation scope is given in §7.

Profiles and splits. For each perturbation, we form a pseudobulk profile, normalize to counts per 10,000, and compute $\log_2$ fold change against non-targeting controls with pseudocount ε=0.05 unless stated. A random split divides cells into two halves and, on the multi-construct subset, preserves the construct mixture on both sides. A construct split assigns the two complete constructs to opposite halves. Each half uses its own split control. We correlate the two profiles across the 6,642 genes, average Pearson correlations over perturbations and 20 splits, and then apply the Spearman–Brown (SB) transform. The construct-split result is reported as an SB-transformed cross-construct coefficient; the parallel-form interpretation of SB is not established for different constructs. Independent-control reagent-pair concordance is reported separately in Appendix F.

Prediction targets and selection by identifier and cell count. For the core global construct exclusion analysis, each comparison uses two complete registered dual-sgRNA cassettes targeting the same gene: one remains present in source data and the other is withheld. For each gene and held-out cell type, we enumerate feasible ordered pairs of distinct complete IDs. Both constructs require at least 40 cells in the held-out panel, and the construct present in source data requires at least 40 cells in each source panel. We select the lexicographically first feasible pair and exclude the full withheld ID from all source data before response processing. Eligibility and selection use only identifiers and cell counts, not expression, knockdown or prediction scores.

Feature availability determines the shared fitted predictor cohort. For the fitted predictor comparison, requiring supplied GenePT embeddings leaves 57, 60, 49 and 31 evaluation genes on K562, RPE1, Jurkat and HepG2, and 2,033 common training genes. Feature eligibility applies to the perturbed input gene, not the scored output genes; the response axis remains 6,642 genes. Each evaluation gene remains represented in source data, so this tests construct transfer across cell types rather than unseen gene generalization. Appendix G gives the eligibility bookkeeping.

Three targets for one prediction. Let x_R, x_U be construct specific count sums and $h(x) = 10^4 x / \sum_j x_j$. Using a common held out control profile h_0, define $T(x) = \log_2[(h(x)+\varepsilon)/(h_0+\varepsilon)]$ coordinatewise. The represented, selected pair pooled and withheld targets are $T(x_R)$, $T(x_R + x_U)$ and $T(x_U)$. Pooling occurs before normalization and is not an average of log fold changes. A single

gene conditioned prediction is scored against all three; only the realized target changes. The selected pair pool is descriptive context: it combines a represented and a withheld cassette, whereas only $T(x_U)$ evaluates withheld construct transfer.

Scoring. Each frozen gene prediction is scored against represented and withheld profiles using the same held out control, response gene axis and transformation. The suffix Δ denotes the control referenced log fold change response: Pearson-Δ is correlation across response genes and MAE-Δ is mean absolute log fold change error. Scores are averaged equally over perturbation genes within each fold, without SB transformation. The worked example in §4 instead compares all construct pooling with an average of separately scored observed constructs under its own eligibility rule.

Fixed mixture repeatability and cross construct concordance measure operational agreement under specified target realizations, but the observed split statistics do not identify a latent gene level biological variance share (Appendix A).

3 CONSTRUCT REALIZATION CHANGES THE MEASURED TARGET

3.1 CHANGING CONSTRUCT IDENTITY CHANGES THE REALIZED RESPONSE.

Random splits preserve the construct mixture on both sides, whereas construct splits compare alternative realizations of the same gene (Fig. 1A). Before matching cell counts, mean aggregated Spearman–Brown (SB) coefficients are 0.248–0.405 for random splits and 0.025–0.140 for construct splits on the same 118 K562 and 135 targets in each other panel. After matching cell counts, construct split correlations span zero much more often than random split correlations (Fig. 1B,C). With at least 40 cells per construct, the random minus construct SB differences are 0.197, 0.198, 0.149 and 0.178 in K562, RPE1, Jurkat and HepG2, respectively ($n = 96, 102, 95, 45$); every perturbation bootstrap interval is above zero (Fig. 1D; Appendix B, Table A1).

Repeatability depends on what is repeated. Random splits can reproducibly recover the same construct mixture even when distinct constructs of the same gene yield less concordant profiles. Random split and construct split coefficients therefore answer different evaluation questions rather than serving as interchangeable estimates of one target's repeatability. A stable gene pooled benchmark target need not be interchangeable with a separately realized construct target.

The contrast persists among strong responses defined independently. Ranking K562 genes by response magnitude in GenomeWide leaves a positive random minus construct SB difference in the high response Essential stratum ($n = 39$; 0.229 [0.122, 0.343], adjusted interval; Table A3). Because the ranking is defined in a separate experiment, this result is not restricted to genes with weak responses in the evaluated split analysis. Fig. 1F shows the corresponding descriptive raw-Pearson robustness. Appendix B reports all strata and the available depth analysis.

Response coordinates change the magnitude of the contrast. At matched cell counts, random to construct SB ratios are 2.28–8.43 for full profiles and 1.07–1.20 for top 50 and top 200 coordinates selected from the pooled evaluation profile. Both splits use the same selected coordinates, although selection from the evaluation profile can increase concordance. The ratios therefore describe these scoring spaces rather than a single coordinate independent effect size.

3.2 THE CONSTRUCT DIFFERENCE RECURS BEYOND RESAMPLING ONE CELL POOL.

For each gene, the difference between its two most covered complete cassettes cancels the common control reference. Across the K562 Essential and GenomeWide experiments, these difference vectors agree at mean Pearson 0.1562, versus -0.0036 for mismatched perturbations. Within panels, their SB transformed split half recurrence reaches 74–88% of the corresponding mean profile recurrence. Reused construct pairs also show cross cell line agreement of 0.063–0.147 against mismatched pair nulls near zero. Fig. 1E shows the three RPE1-based comparisons. The six cross cell line comparisons reuse construct pairs and are therefore dependent.

Residual construct difference agreement remains above the mismatched pair null in all six cross cell line comparisons after mean response alignment using a scalar coefficient fitted in another cell line. Appendix G, Table A14, reports this transported coefficient analysis; Appendix C reports target

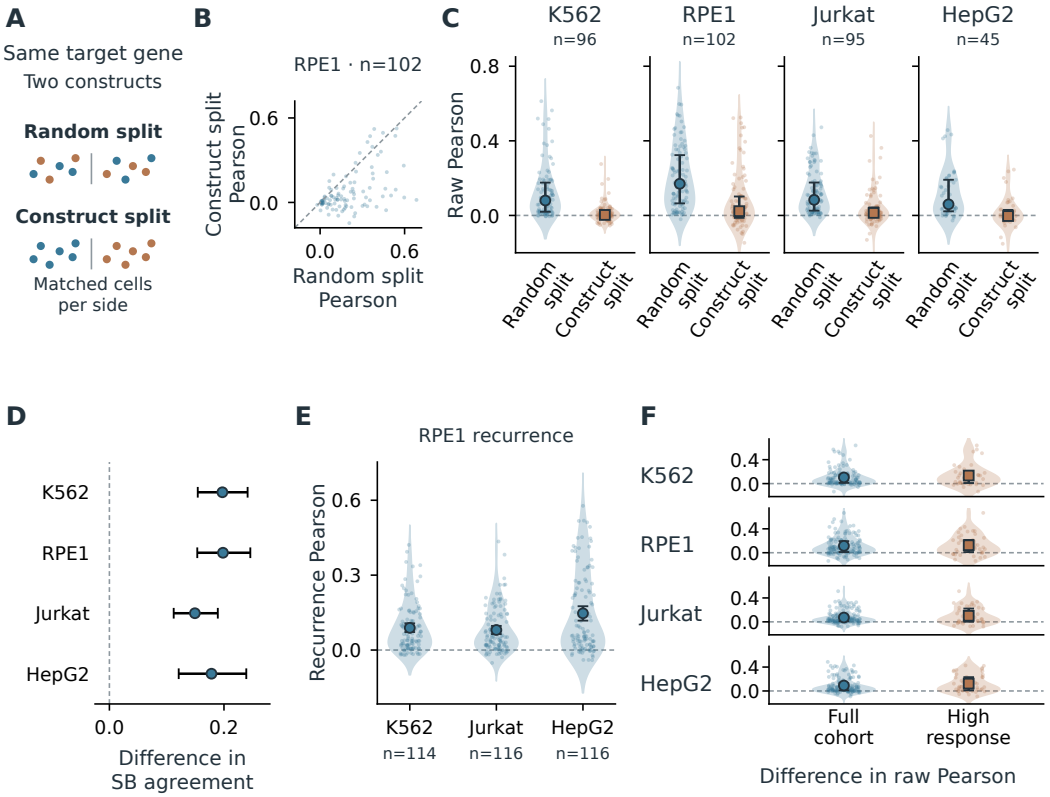

Figure 1: **Alternative constructs targeting the same gene produce less concordant responses than random cell resampling; this difference recurs across contexts. A:** At matched cell counts, random splits preserve the construct mixture; construct splits separate the two complete constructs. **B:** Per-gene RPE1 raw Pearson agreement, averaged over 20 splits. **C:** Raw Pearson across four contexts; gene points, median and IQR; (n) counts paired genes passing the 40-cell floor. **D:** Matched-cell Spearman–Brown (SB) agreement differences on C's cohorts; 95% gene-bootstrap intervals. **E:** RPE1–context recurrence; (n) counts shared genes with the same construct pair. **F:** Random-versus-construct agreement differences for full profiles and independently defined high-response cohorts. K562 response ranking uses an independent experiment; other contexts use transported ranking. Faint points are genes.

knockdown gap diagnostics. The residual agreement shows that recurrence is not explained by the fitted scalar response alignment alone.

3.3 External guide libraries extend the finding across reagent systems.

External CRISPR KO and CRISPRi screens also show greater mean profile disagreement between guides targeting the same gene than between equal cell count resamples of the same guide (Table 1). These model free comparisons use each dataset's control defined response coordinates and do not depend on prediction models or the core panels' SB statistic (Frangieh et al., 2021; Drepanos et al., 2026; Jiang et al., 2025).

Each guide pair provides two disjoint cell halves per guide at the same cell count. We compare cross guide squared disagreement with the average disagreement between resamples of the same guide, average partitions within each gene, and then weight genes equally. Applying the same calculation to control guides defines a separate control floor. Exceeding same guide resampling and exceeding that floor are distinct comparisons.

Both CRISPR KO datasets show excess disagreement beyond same guide resampling: Frangieh's independent Cas9 library shows the primary effect across 205 genes, and Drepanos provides a second

Table 1: **External CRISPR screens show target variation beyond same guide resampling, with dataset specific controls.** Beyond-resampling compares same-gene guides with equal-cell-count resamples of the same guide. Positive control-relative values indicate greater targeting-guide disagreement. Intergenic and non-targeting controls answer different questions; differing response spaces and control definitions preclude direct effect-magnitude comparisons. Appendix H gives statistical procedures, control counts and sensitivity analyses.

Dataset / context / modality	Genes (n)	Excess over same guide resampling [interval]	Control comparison	Difference vs control [interval]	Control result
Frangieh / IFNG / Cas9 KO	205	0.002001 [0.001009, 0.003043]	Intergenic cutting	-0.000376 [-0.007117, 0.005215]	Interval includes 0
Drepanos / K562 D14 / KO	23	0.001321 [0.000506, 0.002652]	Intergenic cutting	0.000816 [-0.000212, 0.002408]	Positive estimate; interval includes 0
Drepanos / K562 D14 / CRISPRi	22	0.005898 [0.002332, 0.011188]	Intergenic binding / guide identity	0.005357 [0.001528, 0.011453]	Positive
Jiang / HT29/TNFA / CRISPRi	44	0.003429 [0.002299, 0.004657]	Non-targeting construct identity	0.002340 [0.000759, 0.004040]	Positive

KO dataset. Neither dataset's intergenic guide comparison resolves an excess over its respective control floor (Table 1).

The CRISPRi datasets additionally exceed their specified control floors. In Drepanos, excess cross guide disagreement exceeds the genome engaging intergenic binding and guide identity floor in the 22 gene ZIM3 KRAB arm. In Jiang, the Dolcetto dCas9 KRAB MeCP2 screen exceeds a non targeting construct identity floor; the registered leave top three sensitivity remains positive after removing the three largest targeting gene contributions (Table A22). These floors answer different questions because Drepanos intergenic guides recruit the effector to genomic loci whereas Jiang non targeting guides do not. Together, the external datasets extend beyond resampling target variation across CRISPR KO and CRISPRi reagent systems (Table 1; Appendix H).

4 FIXED PREDICTIONS, DIFFERENT EVALUATION TARGETS

4.1 CHANGING THE REALIZED TARGET CHANGES FIXED PREDICTION SCORES.

Scoring observed constructs separately rather than against an all construct count pool yields 31–41% lower mean Pearson and 27–52% higher mean MAE for Perturb Mean, Linear-PCA and Linear-GenePT, without changing predictions. In K562, for example, Linear-GenePT mean Pearson changes from 0.120474 for the pooled target to 0.071669 for separately scored constructs on the same 96 genes (Fig. 2A). Model ordering is unchanged; this comparison establishes an absolute score consequence of target realization.

The separately scored target averages eligible constructs equally within gene and then averages genes equally; the pooled target combines all counts before normalization, including constructs below the individual floor. The comparison therefore changes depth, construct mixture and aggregation together. The 96/86/79/33 gene cohorts and full scoring specification are in Appendix G.

4.2 A RELEASED EXTERNAL PREDICTOR IS SENSITIVE TO TARGET REALIZATION.

Scoring released State predictions against alternative complete constructs of the same gene yields different scores at matched cell counts (Adduri et al., 2026). Cross construct Pearson score disagreement exceeds disagreement from resampling the same construct in all three predeclared inferential folds; Table 2A reports the estimates and intervals together with descriptive K562. Scoring the same released prediction against both constructs holds the prediction fixed while target realization changes.

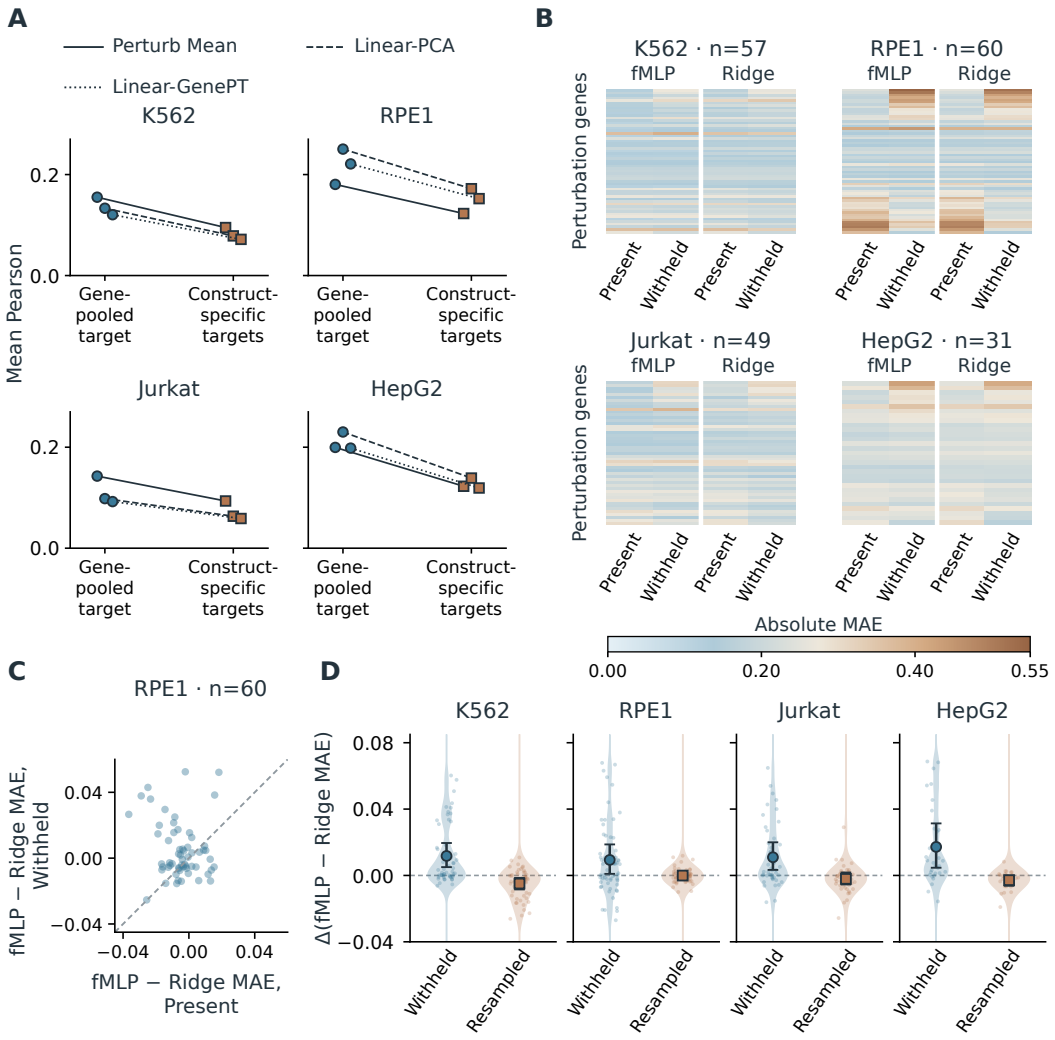

Figure 2: **Target realization changes model evaluation, and construct transfer can change relative performance under global construct exclusion. A:** Mean Pearson for fixed predictions against all-construct pooled versus separately scored eligible targets. **B:** Gene-level absolute MAE for fMLP and ridge regression (Ridge) on matched Present (R') and Withheld (U_d) targets across four contexts under global construct exclusion; genes remain represented and predictions frozen. **C:** RPE1 gene-level gaps $g(R')$ versus $g(U_d)$, where $g(T) = \mathrm{MAE}(\mathrm{fMLP}, T) - \mathrm{MAE}(\mathrm{Ridge}, T)$. Displacement from identity gives the change summarized by D. **D:** Withheld: matched construct-transfer $D = g(U_d) - g(R')$; Resampled: same-construct control $L = g(R') - g(R_0)$. Points show equal-gene means with adjusted paired-gene intervals; faint distributions show gene values. All four D intervals are positive; all L point estimates are non-positive.

The measured change is in scores, not predictions. For each gene, two disjoint halves of each construct provide four targets at the same cell count, and one fixed prediction is scored against all four. We compare score disagreement across constructs with disagreement between resamples of the same construct, using squared Pearson score differences. The difference is averaged over partitions and then equally over genes. Appendix I gives the exact estimator and transformations.

Few genes contribute disproportionately, so fold means exceed medians (Table 2A). The largest gene contributes 27.7%, 49.7% and 39.6% of the net fold sum in RPE1, Jurkat and HepG2, respectively. In a non preregistered influence diagnostic, RPE1 and HepG2 retain positive intervals after removing their three largest contributors, whereas Jurkat loses resolution after removing its two largest (Table

Table 2: **State evidence under two distinct evaluation designs. A:** Fixed released predictions scored against alternative construct targets; positive values indicate squared Pearson score disagreement beyond same construct resampling. K562 is descriptive; RPE1, Jurkat and HepG2 are predeclared inferential folds. **B:** State trained from scratch for 100,000 optimizer steps under global construct exclusion, scored on full-depth targets. Positive values indicate lower Pearson for targets withheld from all source data; RPE1 and Jurkat exceed the predeclared 0.02 materiality margin. Appendix I: full Phase A estimators and sensitivities; Appendix J: Phase B procedures and State Pearson results.

A: Released State — fixed prediction target sensitivity

Fold (n)	Mean excess squared Pearson disagreement [interval]	Median per gene
K562 (14)	0.000965 [-0.000201, 0.002814]	-0.000253
RPE1 (36)	0.012007 [0.002994, 0.024259]	0.001575
Jurkat (31)	0.006277 [0.000362, 0.016792]	0.000200
HepG2 (31)	0.016423 [0.003985, 0.035947]	0.003245

B: State trained from scratch — target sensitivity under global construct exclusion

Fold (n)	Pearson decrease when target construct is withheld [interval]	Result
K562 (57)	0.063878 [-0.006339, 0.139961]	Positive estimate; interval includes 0
RPE1 (60)	0.129821 [0.035025, 0.225598]	Material decrease
Jurkat (49)	0.106916 [0.047568, 0.178545]	Material decrease
HepG2 (31)	0.105918 [-0.022549, 0.251521]	Positive estimate; interval includes 0

A27). The registered inference remains based on the full preregistered cohorts; these removals are descriptive influence diagnostics only.

The released State analysis uses the release's efficacy filtered gene population and 2,000 output coordinates and does not impose global construct exclusion. Appendix I reports all folds, comparator outcomes and sensitivity analyses.

From absolute scores to relative evaluation. We next ask whether construct transfer changes relative model performance at matched cell counts under global construct exclusion (§5).

5 GLOBAL CONSTRUCT TRANSFER CHANGES RELATIVE EVALUATION

Matched cell count evaluation distinguishes targets by whether their construct is present in source data. With predictions fixed, targets from constructs present in source data have higher mean Pearson in all 16 comparisons across the four predictors and lower mean MAE in 15 of 16 than targets from withheld constructs. Thirteen of 32 adjusted intervals resolve an advantage for constructs present in source data, and none resolve the opposite. Appendix E reports the complete endpoints, including the uncertain RPE1 ridge MAE result. Folds reuse constructs. Figure 2B displays the gene-level matched-target MAEs for fMLP and ridge.

Predictors under global construct exclusion. For each evaluation gene, only the selected construct that remains present in source data supplies source responses; all other constructs for that gene, including the withheld complete ID, are removed from every source panel before pooling, normalization, response derived features or fitting (Appendix E). The exclusion is applied before preprocessing caches are built. Linear-GenePT maps fixed 1,536-dimensional GenePT embeddings (Chen & Zou, 2025) to 6,642 response genes by ridge regression. fMLP-GenePT-LFC adapts the fMLP neural core (Miller et al., 2025) to the same embeddings, source responses, exclusions and output coordinates. Its three-seed mean prediction is frozen before scoring, as are all other predictions. The model predicts pseudobulk LFCs rather than individual cells and is fitted here rather than pretrained. Perturb Mean averages each gene's three source-panel responses; Linear-PCA is secondary. Appendix E specifies all fits.

5.1 PREDICTORS SHARING GENE FEATURES INCUR DIFFERENT TRANSFER PENALTIES.

On the full targets, fMLP incurs a larger MAE construct transfer penalty than ridge in all four contexts. All four adjusted intervals for the penalty difference are positive (Table A12); K562 and HepG2 also exceed the predeclared $+0.005$ materiality margin, while RPE1 and Jurkat exclude zero without clearing that margin.

Let N denote fMLP and G ridge, with R_m and U_m denoting their mean MAEs on targets from the represented and withheld constructs. The MAE gap change is

$$I_{NG} = (U_N - U_G) - (R_N - R_G).$$

Inference uses the 40 contrast paired gene family and conditions on the fixed ensemble.

5.2 THE PAIRED GAP CHANGE SURVIVES MATCHING CELL COUNTS.

Define

$$g(T) = \mathrm{MAE}(N, T) - \mathrm{MAE}(G, T).$$

For held out counts n_R and n_U, set $n_P = \min(n_U, n_R - 20)$ and form R' and U_d with exactly n_P cells each. The remaining represented cells form a disjoint R_0 from the same construct. We then define

$$D = g(U_d) - g(R'), \qquad L = g(R') - g(R_0).$$

D is the matched construct transfer contrast, while L is the same construct control. All four adjusted D intervals are above zero, whereas all L point estimates are non-positive (Figure 2D). Only the K562 D interval exceeds the predeclared $+0.005$ materiality margin.

Different transfer penalties change the model comparison. Fig. 2C plots $g(U_d)$ against $g(R')$ for individual RPE1 genes; each point's vertical displacement from the identity line is the paired gap change whose equal-gene mean is D. If changing the realized target added the same MAE penalty to both predictors, $g(T)$ would remain unchanged. In K562, for example, changing from R' to U_d increases MAE by 0.013280 for fMLP but only 0.001528 for ridge, giving $D = 0.011752$ $[0.005021, 0.019553]$. Construct transfer therefore changes relative model performance, not only the absolute score of each predictor. The interval is computed from paired within gene changes, preserving covariance between the predictors' scores.

Mean Pearson ordering also reverses descriptively in all four contexts (Table A5). In RPE1, fMLP and ridge are close on R' (0.178610 versus 0.174186), but on U_d ridge exceeds fMLP (0.111695 versus 0.071114). Thus the predictor with the higher mean Pearson changes with the realized target, despite unchanged predictions. This ordering is descriptive; the paired MAE gap analysis provides the inferential test of relative performance. Appendix E reports the complete endpoints and uncertainty intervals.

5.3 OTHER PREDICTOR COMPARISONS AND RECIPROCAL SENSITIVITY.

Ridge also incurs a smaller MAE construct transfer penalty than Perturb Mean: all four adjusted intervals for their penalty difference lie below -0.005 under the 24 contrast family (Table A4). On the strict 22 gene reciprocal cohort, exchanging the roles of the two constructs preserves positive fMLP minus ridge MAE gap changes in all four contexts. This reciprocal analysis is reported descriptively outside the primary 40 contrast family. Appendix E reports the complete intervals, orientation averaging and denominator gated Pearson comparisons.

A modern deep model remains sensitive to target realization under global construct exclusion. A State architecture trained from scratch for 100,000 optimizer steps via the source-only bridge (Appendix J) shows lower Pearson scores on full-depth targets realized by constructs withheld from all source data, with the difference exceeding the preregistered materiality margin in RPE1 and Jurkat (Table 2B; Table A29). This extends the global construct exclusion evidence to a modern deep model and shows that whether a construct is present in source data can change model evaluation.

6 IMPLICATIONS FOR PERTURBATION-MODEL EVALUATION

Choose the evaluation target before comparing predictors. A gene label and cellular context holdout do not fully specify the response against which a model should be judged. If the intended target is a gene's pooled experimental population, combine counts using a declared construct mixture before normalization and scoring. If the question is agreement with alternative realized responses, score eligible constructs separately and declare how constructs and genes are weighted. The fixed prediction comparison in §4 shows that these choices change absolute scores even when mean model ordering is retained. A pooled score evaluates the declared mixture; agreement with each contributing construct requires separate scoring.

Specify which constructs are present in source data. A context holdout and a complete construct holdout ask different generalization questions. For construct transfer, keep the gene represented through one construct while withholding the complete construct from all source data. Apply this exclusion before pooling, normalization, response derived features and preprocessing caches. The withheld unit must be the registered complete construct ID, not a partly shared spacer. Released State predictions in §4 do not impose this contract and therefore test fixed prediction target sensitivity. The State architecture trained from scratch in §5 is evaluated under global construct exclusion.

The fMLP–ridge comparison in §5 shows that predictors incur different construct transfer penalties, changing relative model performance. Report absolute scores on both realized targets and the paired change in the MAE gap with paired uncertainty; mean Pearson ordering remains descriptive. Evaluate the same frozen predictions using the same eligible genes, response coordinates, transformation and held out reference, and match target cell counts to separate target realization from sampling depth. Specify the cohort, sampling rule and uncertainty family for each comparison.

Make the benchmark declaration reproducible. Record the target construct IDs, which construct IDs are present in source data, the pooling or score averaging rule, target cell counts, scored response coordinates and how they were selected, and the control reference construction. Random splits measure repeatability of a fixed construct mixture, whereas construct splits measure agreement across alternative realizations. Full profiles and top response coordinates selected from the pooled evaluation profile can yield different contrast magnitudes (Appendix B); shared control references can also change pair agreement (Appendix F). External control floors answer different questions and should be reported with their specific definition. These declarations distinguish evaluations that share the same gene label and metric but ask different prediction questions.

7 SCOPE AND CONCLUSION

Empirical scope. Construct reuse spans the shared gene population in the core panels, but direct alternative construct comparisons require multiple constructs and therefore do not estimate effects for genes represented by a single construct. The core recurrence comparisons share reagent lineage. External CRISPR KO and CRISPRi screens extend the target variation evidence across other libraries and a different repressor fusion, on their registered cohorts and response spaces; they do not extend the fitted model construct transfer experiment to those datasets.

Interpretation. The observed construct differences do not identify a unique mechanism, such as efficacy or preparation. Under global construct exclusion, the evaluation effects likewise do not isolate a single mechanism for why withholding a construct from source data changes model performance. External control floors constrain specific alternatives rather than all guide related mechanisms; in particular, Jiang's non targeting floor does not control genomic effector recruitment. Target gene LFC is an efficacy proxy rather than evidence of efficacy balance (Appendix E).

Conclusion. A gene label does not fully specify the realized target used to evaluate a perturbation model. Construct-associated target differences recur across the core panels and exceed same guide resampling in external CRISPR KO and CRISPRi screens. Changing the realized target changes absolute scores for fixed predictions, including released State predictions; under global construct exclusion, construct transfer also changes the relative performance of fitted predictors. A State architecture trained from scratch extends target realization sensitivity under the same exclusion design to a modern deep model. Perturbation model evaluation should specify target realization and which constructs are present in source data, together with the predictor and scoring rule.

AI USE STATEMENT

Generative AI tools were used to provide feedback on research methodology and experimental design, assist with interpretation of results, and review mathematical reasoning. They were also used for manuscript editing and language polishing, software and workflow assistance, and scientific figure design and editing. All AI-assisted analyses, claims, code, and manuscript changes were reviewed and verified by the authors, who take responsibility for the final content of this work.

REPRODUCIBILITY STATEMENT

The manuscript and appendices specify the evaluation cohorts, construct-selection and exclusion rules, target construction, estimands, uncertainty procedures, and model-training protocols used in the reported analyses. Appendix E documents global construct exclusion and matched-target evaluation; Appendix H specifies the external CRISPR-screen analyses, controls, and inference; and Appendices I and J describe released-State prediction evaluation and State training from scratch under global construct exclusion, respectively. The reproducibility and provenance section records the relevant data sources, model artifacts, randomization procedures, and available analysis outputs. Code and execution scripts supporting the reported analyses will be released upon publication.

REFERENCES

Abhinav K. Adduri, Dhruv Gautam, Beatrice Bevilacqua, Mohsen Naghipourfar, Alishba Imran, Rohan Shah, Noam Teyssier, Rishi Verma, Christopher Carpenter, Basak Eraslan, Francis Chalissery, Rajesh Ilango, Vishvak Subramanyam, Chiara Ricci-Tam, Sanjay Nagaraj, Aidan Winters, Mingze Dong, Stefanie Fellinger, Adam Krejci, Tilmann Burckstummer, Sravya Tirukkovular, Jeremy Sullivan, Brian S. Plosky, Nicholas D. Youngblut, Jure Leskovec, Luke A. Gilbert, Silvana Konermann, Patrick D. Hsu, Alexander Dobin, Dave P. Burke, Hani Goodarzi, and Yusuf H. Roohani. 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.

Felicity Allen, Fiona Behan, Anton Khodak, Francesco Iorio, Kosuke Yusa, Mathew Garnett, and Leopold Parts. JACKS: joint analysis of CRISPR/Cas9 knockout screens. *Genome Research*, 29 (3):464–471, 2019. doi: 10.1101/gr.238923.118.

W. Brown. Some experimental results in the correlation of mental abilities. *British Journal of Psychology*, 3(3):296–322, 1910.

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

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.

Laura M. Drepanos, Berta Escude Velasco, Abigail J. Chase, et al. Direct comparison of CRISPR knockout and interference with Perturb-seq. *bioRxiv*, 2026. doi: 10.64898/2026.07.04.736492. Preprint.

Chris J. Frangieh, Johannes C. Melms, Pratiksha I. Thakore, et al. Multimodal pooled Perturb-CITE-seq screens in patient models define mechanisms of cancer immune evasion. *Nature Genetics*, 53: 332–341, 2021. doi: 10.1038/s41588-021-00779-1.

Kiyoshi F. Fukutani, Thomas H. Hampton, Mattheus Luís Alves Santiago, Eddy José Francisco de Oliveira, and Todd A. MacKenzie. When normalization induces correlation: shared and unstable references can create hidden dependencies. *npj Systems Biology and Applications*, 12:90, 2026. doi: 10.1038/s41540-026-00722-2. open access; PMCID PMC13287743.

Longda Jiang, Carol Dalgarno, Efthymia Papalexi, et al. Systematic reconstruction of molecular pathway signatures using scalable single-cell perturbation screens. *Nature Cell Biology*, 27: 505–517, 2025. doi: 10.1038/s41556-025-01622-z.

Mohammad Lotfollahi, Anna Klimovskaia Susmelj, Carlo De Donno, Leon Hetzel, Yuge Ji, Ignacio L Ibarra, Sanjay R Srivatsan, Mohsen Naghipourfar, Riza M Daza, Beth Martin, Jay Shendure, Jose L McFaline-Figueroa, Pierre Boyeau, F Alexander Wolf, Nafissa Yakubova, Stephan Günnemann, Cole Trapnell, David Lopez-Paz, and Fabian J Theis. Predicting cellular responses to complex perturbations in high-throughput screens. *Molecular Systems Biology*, 19(6):e11517, 2023. doi: 10.15252/msb.202211517.

Robin M Meyers, Jordan G Bryan, James M McFarland, Barbara A Weir, Ann E Sizemore, Han Xu, Neekesh V Dharia, Phillip G Montgomery, Glenn S Cowley, Sasha Pantel, Amy Goodale, Yenarae Lee, Levi D Ali, Guozhi Jiang, Rakela Lubonja, William F Harrington, Matthew Strickland, Ting Wu, Derek C Hawes, Victor A Zhivich, Meghan R Wyatt, Zohra Kalani, Jaime J Chang, Michael Okamoto, Kimberly Stegmaier, Todd R Golub, Jesse S Boehm, Francisca Vazquez, David E Root, William C Hahn, and Aviad Tsherniak. Computational correction of copy number effect improves specificity of CRISPR–Cas9 essentiality screens in cancer cells. *Nature Genetics*, 49:1779–1784, 2017. doi: 10.1038/ng.3984.

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*, 10 2025. doi: 10.1101/2025.10.20.683304.

Ajay Nadig, Joseph M. Replogle, Angela N. Pogson, Mukundh Murthy, Steven A. McCarroll, Jonathan S. Weissman, Elise B. Robinson, and Luke J. O'Connor. Transcriptome-wide analysis of differential expression in perturbation atlases. *Nature Genetics*, 57(5):1228–1237, 2025. doi: 10.1038/s41588-025-02169-3.

Phillip B. Nicol, Shriya Shivakumar, and Rafael A. Irizarry. Spurious correlation inflates performance in single-cell perturbation prediction. *bioRxiv*, 2026. doi: 10.64898/2026.05.07.723486. Preprint, v2, posted 11 August 2026.

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, Torsten Gross, Joseph Min, Samuele Garda, Bo Yuan, Linus J. Schumacher, Jake P. Taylor-King, Debora S. Marks, Augustin Luna, Nils Blüthgen, and Chris Sander. scPerturb: Harmonized single-cell perturbation data. *Nature Methods*, 21(3):531–540, 2024. doi: 10.1038/s41592-023-02144-y.

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

Neha Rohatgi, Jean-Philippe Fortin, Ted Lau, Yi Ying, Yue Zhang, Bettina L. Lee, Michael R. Costa, and Rohit Reja. Seed sequences mediate off-target activity in the CRISPR-interference system. *Cell Genomics*, 4:100693, 2024. doi: 10.1016/j.xgen.2024.100693.

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.

Kendall R. Sanson, Ruth E. Hanna, Mudra Hegde, et al. Optimized libraries for CRISPR-Cas9 genetic screens with multiple modalities. *Nature Communications*, 9:5416, 2018. doi: 10.1038/s41467-018-07901-8.

C. Spearman. Correlation calculated from faulty data. *British Journal of Psychology*, 3(3):271–295, 1910. doi: 10.1111/j.2044-8295.1910.tb00206.x.

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

Xi Wang, Jack Kuipers, Florian Hugi, Randall Jeffrey Platt, and Niko Beerenwinkel. Reliable single-cell perturbations explain and improve model performance. *bioRxiv*, 2026. doi: 10.64898/2026.08.11.744177. URL https://www.biorxiv.org/content/10.64898/2026.08.11.744177v1. Preprint, v1, posted 12 August 2026.

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

Yan Wu, Esther Wershof, Sebastian M. Schmon, Marcel Nassar, Błażej Osiński, Ridvan Eksi, Zichao Yan, Rory Stark, Kun Zhang, and Thore Graepel. PerturBench: Benchmarking machine learning models for cellular perturbation analysis. In *Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track*, 2025. URL https://openreview.net/forum?id=PPPDuyiZaG. arXiv:2408.10609.

A OBSERVABLE SPLIT ESTIMANDS

The construct split does not identify a latent gene-level share. Consider a centered profile over scored genes, $y_c^{(a)} = \beta + \delta_c + \epsilon_c^{(a)} - r_c^{(a)}$, where c is construct identity and a indexes repeated measurement. Here β is gene-associated, δ_c is a stable construct deviation, and ϵ, r are sampling and control-reference errors. Define $B = \mathrm{Var}(\beta)$, $D = \mathrm{Var}(\delta_c)$ and $K = \mathrm{Cov}(\delta_1, \delta_2)$. For this compact argument take equal construct variances, independent centered errors across measurements and an orthogonal parameterization $\mathrm{Cov}(\beta, \delta_c) = 0$. Orthogonality is a choice of parameterization, not an empirical biological finding.

What a split preserves determines its covariance. Repeating one construct preserves both β and its deviation. Switching constructs preserves β and whatever deviation covariance K the pair shares. Taking a difference cancels β and a common control. With $d^{(a)} = y_1^{(a)} - y_2^{(a)}$, the stable covariance terms are

$$\begin{aligned} \text{same construct:} \quad & \mathrm{Cov}(y_c^{(a)}, y_c^{(b)}) = B + D, \\ \text{cross construct:} \quad & \mathrm{Cov}(y_1^{(a)}, y_2^{(b)}) = B + K, \\ \text{repeated difference:} \quad & \mathrm{Cov}(d^{(a)}, d^{(b)}) = 2(D - K). \end{aligned}$$

A random cell split instead approximates a fixed mixture. If construct weights are w and $1 - w$, its deviation is $\delta_w = w\delta_1 + (1 - w)\delta_2$, with $D_w = [w^2 + (1 - w)^2]D + 2w(1 - w)K$. The repeated-mixture covariance is $B + D_w$, rather than $B + D$; at equal weights $D_w = (D + K)/2$. Finite-cell imbalance enters the sampling variability. Thus the positive random-split distribution and the smaller cross-construct distribution measure different stable objects even at equal depth.

The latent attribution is not identified. For any covariance-admissible shift s, replacing (B, D, K) by $(B + s, D - s, K - s)$ leaves all four quantities $B + D$, $B + K$, $D - K$ and $B + D_w$ unchanged. Observable recurrence therefore cannot determine how much common covariance belongs to B rather than K, nor generally the sign of K. In particular, the construct-split covariance need not upper-bound B. Converting covariances to correlations and applying SB does not resolve that ambiguity: denominators differ across constructs and mixtures, and averaging precedes a nonlinear transform. The admissibility bounds follow below.

The evaluation quantity follows the target specification. Once a reagent distribution and exposure rule are specified, fixed-mixture repeatability, observed-construct concordance and transfer to a globally excluded construct define distinct operational questions. The non-identification result separates these measurable comparisons from a latent biological decomposition. A benchmark can evaluate a particular realized response without claiming that its split coefficient measures a gene-level share. Section 5 uses this distinction to test relative evaluation: each prediction is held fixed while the evaluator changes the construct supplying the target.

Covariance admissibility. The following gives the admissible range for the decomposition above and its relation to correlations.

Identification. The transformation $B' = B + s$, $D' = D - s$, $K' = K - s$ leaves $B + D$, $B + K$, $D - K$ and $B + D_w$ unchanged. Any shift satisfying $B + s \geq 0$ and $D - s \geq |K - s|$ is covariance-admissible. In particular, when the observed single-construct and cross-construct stable covariances satisfy $0 < C < S$, with $S = B + D$ and $C = B + K$, feasible B spans $0 \leq B \leq (S + C)/2$, so $K = C - B$ can have either sign. These observables therefore do not in general identify the sign of K or the latent gene-associated share. The construct-split coefficient is neither that share nor its identified numerical upper bound.

Correlations and the reported statistic. Writing the independent sampling and reference variances as E and R, under the equal-variance assumptions, raw same-construct and cross-construct correlations are $(B + D)/(B + D + E + R)$ and $(B + K)/(B + D + E + R)$. The fixed-mixture correlation instead uses $B + D_w$ and its own sampling/reference variances. A common reference adds its variance to cross-measurement covariance; disjoint references remove that shared term under the independence assumption. Empirically, §3 averages Pearson correlations over perturbations and splits and then applies $\mathrm{SB}(r) = 2r/(1 + r)$. Heterogeneous denominators, finite-sample mixtures and the nonlinear transform prevent recovery of B, D, K from those summaries. Parallel forms are not established for different constructs. The identified operational comparison is fixed-mixture

repeatability versus cross-construct concordance under the specified sampling protocol: split design determines the evaluation quantity.

B DEPTH AND COVERAGE

Cell-count restrictions in the construct comparison. With at least 40 cells per construct, the matched-cell-count random-minus-construct SB differences remain positive on K562, RPE1, Jurkat and HepG2, with 96, 102, 95 and 45 eligible genes (Table A1). The 20- and 30-cell restrictions also retain positive differences. These restricted cohorts differ from the 118/135-gene comparisons without the per-construct cell-count restriction and must not be substituted into their statistics.

Table A1: Matched-cell-count random and construct split on the same pairs, each construct with at least 40 total cells; split controls, 6,642 shared genes, ε=0.05, Pearson per perturbation, mean over perturbations and 20 splits, then Spearman–Brown. Intervals resample perturbations; the difference is the primary comparison.

panel	n	random	construct	difference	95% CI
K562	96	0.223	0.026	0.197	[0.154, 0.241]
RPE1	102	0.337	0.139	0.198	[0.153, 0.246]
Jurkat	95	0.217	0.067	0.149	[0.112, 0.189]
HepG2	45	0.210	0.032	0.178	[0.121, 0.239]

Available depth. Both split arms are capped to a common depth, and fits require at least ten contributing perturbations. §3 limits conclusions to available depth; testing greater coverage would ask whether the capped difference narrows toward zero as depth grows. The deepest fitted points are 140 cells per side on K562, 175 on RPE1, 140 on Jurkat and 70 on HepG2. No perturbation on any panel reaches 350 cells per side. These measurements support the reported range, not an extrapolation beyond available coverage.

Distribution and aggregate units. Figure 1B/C report split-averaged raw Pearson values on the coverage-filtered 96/102/95/45-gene cohorts; Figure 1F separately uses the full/high cohorts without the per-construct cell-count restriction. Table A1, which requires at least 40 cells per construct, instead averages correlations before SB transformation on its restricted cohort. Neither establishes a cell-count-invariant ratio of SB-transformed aggregates.

Response space changes the magnitude, while the measured difference stays positive. On the same 118/135-gene cohorts without the per-construct cell-count restriction, the random-minus-construct SB differences with the random arm capped to the construct-side cell count range from 0.127–0.182 over the full 6,642-coordinate profile, 0.093–0.136 on top-50 coordinates and 0.046–0.106 on top-200 coordinates. The corresponding random/construct ratios are 2.28–8.43, 1.14–1.20 and 1.07–1.16. Controls, normalization, pseudocount 0.05 and mean-Pearson-before-SB aggregation are unchanged. Top-DE coordinates are ranked by absolute LFC on each perturbation's pooled evaluation profile; selection can raise both arms. These measured contrasts therefore characterize the specified scoring spaces. The full-profile ratio is not coordinate-invariant, and ratios of SB-transformed aggregates do not inherit the raw-correlation depth-cancellation identity of a simplified variance model.

Response strength from an independent K562 experiment. We rank the fixed multi-construct cohorts by response magnitude in the separate K562 GenomeWide experiment. For positive-cell source construct row k, let $c_{kj} = 10^4 x_{kj} / \sum_{l=1}^{8248} x_{kl}$ and average c_{kj} equally across source rows for each perturbation g. The reference averages all 585 positive-cell non-targeting rows equally. Define $L_{gj}^{GW} = \log_2[(\bar{c}_{gj} + 0.05)/(\bar{c}_{NT,j} + 0.05)]$ and $q_g = 6546^{-1} \sum_{j \in J} |L_{gj}^{GW}|$, where J is the fixed stable-ID intersection. Normalization precedes restriction to this intersection; source rows are not cell-count weighted. Evaluation retains all 6,642 response coordinates and its original count-pooling convention.

Every frozen gene has a finite source score. Source coverage is 6,546/6,642 coordinates; no evaluation coordinate or perturbation is removed. Rank terciles contain 40/39/39 K562 genes and 45/45/45 in each other panel, all exceeding the minimum 20 genes for inference with no boundary ties. K562

Essential is primary; RPE1, Jurkat and HepG2 use the same GenomeWide ranking transported from K562 and cannot rescue a failed primary result. GenomeWide and Essential are disjoint experiments but share library and laboratory lineage. The stratifier is defined from an independent K562 experiment; in the other panels it is transported rather than independently re-estimated locally. It measures response magnitude, not target-experiment repeatability, efficacy or pure biological strength; it includes shared responses and source noise.

For each stratum and arm, average the original raw correlations equally over genes and 20 splits before applying $SB(r) = 2r/(1 + r)$. Define $\Delta = SB(\bar{r}_{random}) - SB(\bar{r}_{construct})$. The saved random arms are capped to the smaller original construct side; the two whole-construct sides need not have equal counts. Separate split controls, all 6,642 coordinates, count pooling before CP10K and pseudocount 0.05 remain fixed. No profiles or split assignments are reconstructed. The descriptive ratio is printed only when both aggregate raw correlations are positive, as they are in every reported stratum.

Stratification rules and memberships were fixed after the historical full-cohort analysis and before stratified outcomes were evaluated. Resample genes within each frozen high tercile, retaining paired arms and all splits, then recompute raw means and SB. A four-contrast family uses 20,000 draws, seed 1707 restarted per panel, and linear percentile tails 0.00625/0.99375 for approximate simultaneous 95% intervals. Inference conditions on measured source scores and fixed memberships; source-score uncertainty is not resampled. Low, middle and full-cohort estimates are descriptive. No materiality threshold is imposed; overlapping panels and splits are not independent replications.

Table A2 retains every stratum. The primary K562 high group has a positive adjusted interval, as do the transported secondary groups (Table A3). High-group point differences exceed low-group differences in every panel, but the pattern is not uniformly monotonic. This addresses confinement to weak responses under the independent K562 ranking and its transported application to the other contexts. It does not establish statistical moderation or rule out target-repeatability explanations. The single prespecified continuous diagnostic is descriptive Spearman association between q_g and each gene's mean raw random-minus-construct correlation; no significance test or searched functional form is used.

Table A2: Response-strength strata defined by the independent K562 GenomeWide ranking and its transported application to the other contexts. The random arm is capped to the construct-side cell count and compared with whole-construct splits, with separate split controls, 6,642 evaluation coordinates and pseudocount 0.05. Raw correlations are averaged equally over genes and 20 splits, then SB-transformed. All rows are retained; ratios are descriptive. K562 is primary; the other panels are secondary transported-ranking analyses.

Panel	Stratum	n	Raw rand.	Raw con.	SB rand.	SB con.	Δ	Ratio
K562	full	118	0.115352	0.012424	0.206843	0.024543	0.182301	8.427838
K562	low	40	0.041821	0.005855	0.080285	0.011641	0.068643	6.896482
K562	middle	39	0.152167	0.012415	0.264141	0.024525	0.239616	10.770242
K562	high	39	0.153951	0.019171	0.266825	0.037620	0.229205	7.092644
RPE1	full	135	0.190993	0.075531	0.320729	0.140453	0.180276	2.283526
RPE1	low	45	0.115366	0.025850	0.206866	0.050397	0.156470	4.104776
RPE1	middle	45	0.211058	0.079982	0.348552	0.148116	0.200435	2.353227
RPE1	high	45	0.246555	0.120762	0.395578	0.215500	0.180079	1.835633
Jurkat	full	135	0.101754	0.029810	0.184713	0.057894	0.126819	3.190556
Jurkat	low	45	0.039724	0.006171	0.076412	0.012266	0.064145	6.229341
Jurkat	middle	45	0.103551	0.022677	0.187669	0.044348	0.143320	4.231704
Jurkat	high	45	0.161988	0.060581	0.278812	0.114242	0.164571	2.440549
HepG2	full	135	0.114349	0.023280	0.205231	0.045501	0.159729	4.510428
HepG2	low	45	0.050901	0.004590	0.096872	0.009138	0.087734	10.601480
HepG2	middle	45	0.129139	0.025493	0.228739	0.049718	0.179021	4.600724
HepG2	high	45	0.163008	0.039759	0.280321	0.076477	0.203845	3.665457

Table A3: High-response SB inference and the continuous descriptive diagnostic. High-group Δ is the difference of SB-transformed raw means in Table A2; intervals use the four-contrast family. Spearman association uses all frozen genes and each gene's mean raw random-minus-construct correlation; it is descriptive, without a significance test. Figure 1F instead shows descriptive raw-Pearson robustness; its whiskers are not these adjusted SB intervals.

Panel	High n	Δ	Adjusted interval	All n	Spearman
K562	39	0.229205	[0.121615, 0.342587]	118	0.297339
RPE1	45	0.180079	[0.092479, 0.278513]	135	0.070920
Jurkat	45	0.164571	[0.084772, 0.249785]	135	0.140976
HepG2	45	0.203845	[0.121231, 0.287188]	135	0.226236

C CONSTRUCT-PAIR DIAGNOSTICS

Measured target-gene LFC gaps in Figures A1 and A2 are absolute differences in the target gene's log fold change against the control pool. Quartiles are formed among targets whose target gene is scored. They are descriptive bands, not efficacy-matched interventions. Measured target-gene LFC is a proxy for efficacy; these comparisons do not rule out nonlinear or unmeasured efficacy effects.

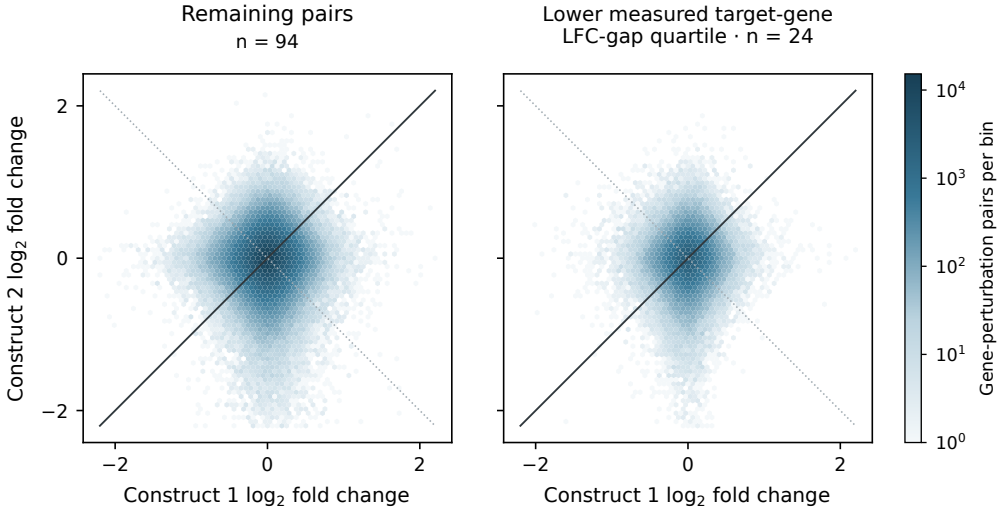

Figure A1: **Observed construct pairs show different profiles.** Each point is one gene within one perturbation on K562: construct 1's $\log_2$ fold change against construct 2's, shaded by density. Spread *along* the diagonal is effect the two reagents share; spread *across* it is where they disagree. The two panels partition the 118 perturbations carrying two constructs: the 24 in the bottom quartile of the measured target-gene LFC gap (*right*), and the 94 outside that selection (*left*). The quartile uses pairs with a measured target coordinate; remaining pairs include those without one. This quartile is descriptive and does not establish efficacy matching. K562, the 6,642 shared genes, ε = 0.05.

D THE SINGLE REAGENT PAIR IN THE CRISPRA PANEL

Our metadata analysis finds that 47 of 48 genes occupying both vector positions use the same spacer. IER5L is the sole distinct-protospacer case in the CRISPRa positional-control panel (§2). Over 20 splits at $\varepsilon = 0.05$, its mean construct-split and random-depth-matched raw correlations are 0.5533 and 0.5614, respectively, at a median cap of 183 cells per side. We report this single-gene case descriptively: its greater available depth and single-protospacer CRISPRa setting differ from the dual-sgRNA CRISPRi comparison.

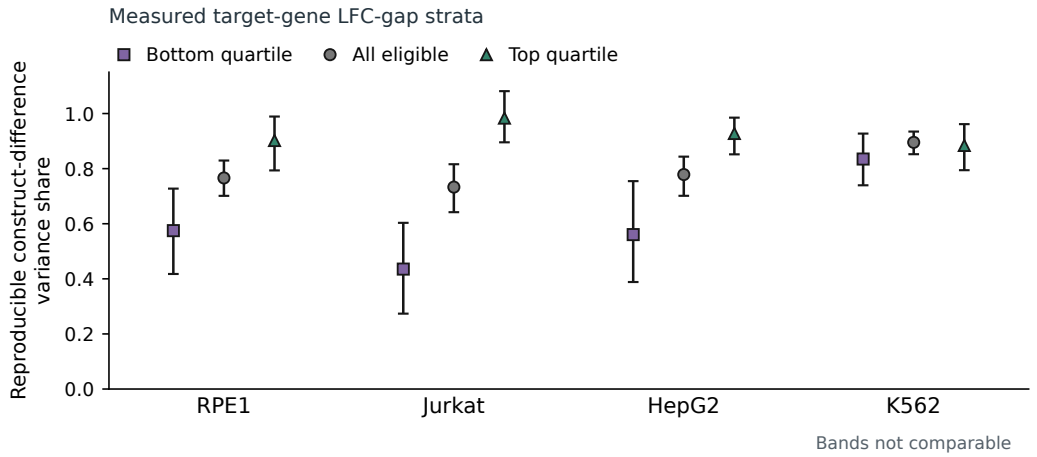

Figure A2: **Reproducible pair differences across measured target-gene LFC-gap quartiles.** Reproducible construct-difference variance share (estimated reproducible construct-difference variance divided by total construct-difference variance) by quartile of the measured target-gene LFC gap, with bootstrap intervals over perturbations. This operational quantity is distinct from the latent gene-level share discussed in Appendix A. On RPE1, Jurkat and HepG2 the bottom quartile still reaches 0.57, 0.44 and 0.56 against 0.90–0.98 in the top quartile. The cross-context statement applies to RPE1, Jurkat, and HepG2; K562 is shown separately. Its bottom quartile has a median gap of 0.60 against RPE1's 0.21. Quartiles do not equate efficacy across constructs or panels. All panels, the 6,642 shared genes, $\varepsilon = 0.05$.

E GLOBAL CONSTRUCT EXCLUSION SPECIFICATION

Navigation. Matched-depth absolute endpoints (Tables A5 and A6) and historical full-target absolute scores (Tables A10 and A11) below include all four methods; the neural tables retain every primary contrast. The main text specifies the common GenePT representation, source-only training rows, exclusion rule and target transformations. Linear-PCA uses training-response features with a rank cap of 50 and is secondary. The construct withheld from all source data is the complete registered construct ID, not every spacer sharing part of its sequence.

The primary MAE margin is 0.005 $\log_2$ fold-change units and the Pearson margin is 0.02 (full-target MAE interaction: Table A4). Primary inference requires at least 30 paired finite genes; all four cohorts retain all eligible genes. Bootstrap seed 1707 and 20,000 draws are fixed. Pearson additionally requires the represented Perturb Mean mean to exceed 0.05 both in the observed sample and in every bootstrap sample. HepG2 fails in 47 stored bootstrap draws; no substitute interaction is reported. Failed-gate slots remain in the 24-contrast family.

Table A4: Linear-GenePT minus Perturb Mean MAE transfer penalties, I_{GB}, for the §5 full-target global construct-exclusion comparison. Equal-gene paired estimates and approximate simultaneous intervals from the 24-contrast family; all upper endpoints are below -0.005. This is the full-target comparison, not the matched D/L comparison in Figure 2D; the uncertainty families are distinct.

Cell type	n	I_{GB}	Adjusted interval
K562	57	-0.018214	[-0.029979, -0.008436]
RPE1	60	-0.018706	[-0.032316, -0.007415]
Jurkat	49	-0.019969	[-0.031847, -0.010271]
HepG2	31	-0.026450	[-0.046360, -0.009722]

The strict sensitivity cohort contains the same 22 reciprocal genes in each cell type. Forward and reverse scores are averaged within gene before resampling, so the effective sample size remains 22. Its intervals are descriptive 95% intervals. All four strict Linear-GenePT–Perturb Mean MAE

interaction intervals lie below the primary negative margin; this consistency is not a separate primary success test. Historical full-target means and Linear-PCA comparisons are descriptive; the original primary family does not provide intervals for them. The matched-depth absolute comparison has its own 32-contrast family.

The fixed-prediction aggregation experiment (§4) uses all-construct pooled targets; the global construct-exclusion experiment (§5) uses a count-pooled selected pair only as context. Neither the cohort nor the pooled target is interchangeable across the two experiments. The primary inference concerns interactions between targets from the construct present in source data and from the withheld construct, with equal perturbation weighting and unchanged predictions.

Feature asset. Gene embeddings use the ada-002 asset in Zenodo record 10833191 (GenePT embedding release v2; `https://doi.org/10.5281/zenodo.10833191`), not the alternative text-embedding-3-large asset or a GenePT-w/s cell embedding. We use the supplied normalized vectors without additional fitted scaling. Ridge on the mean of k source responses uses penalty λ/k, equivalent to penalty λ on the stacked squared-error sum; here $k = 3$ and $\lambda = 1$.

Nonlinear model interface. fMLP-GenePT-LFC is a declared pseudobulk LFC adaptation of the pinned fMLP neural core Miller et al. (2025). Its widths are $1536 \to 512 \to 512 \to 128 \to 512 \to 512 \to 6642$, with affine biases, four LeakyReLU stages of slope 0.2, a linear latent layer and linear output. It uses the same fixed GenePT embedding and source LFC rows as the ridge fit. The adaptation replaces the individual-cell prediction interface with a gene-embedding-to-pseudobulk-LFC map; it is neither a pretrained response model nor the original cell-level wrapper. Training minimizes equally weighted squared error across source rows and response genes. Three fixed seeds are trained for 100 epochs, and their prediction vectors are averaged before scoring.

Fixed training specification. Each of the 12 exclusion jobs is fitted with seeds 42, 43 and 44, for 100 epochs at batch size 256. Adam uses learning rate 0.001, betas (0.9, 0.999), epsilon 10^{-8} and zero weight decay; StepLR multiplies the learning rate by 0.2 every 30 epochs. Training uses shuffled rows, the final incomplete batch, standard linear-layer initialization and the final epoch, without early stopping or hyperparameter selection. All jobs' seed predictions and ensembles are frozen before evaluation targets or reference scores are accessed. The nonlinear extension uses the same eligibility, response axis, exclusion sets and normalization as the linear comparison.

Scope of prospective specification. The same-dataset neural extension was specified before its training and outcomes, with the Linear-GenePT/Perturb Mean results already known. Its primary uncertainty conditions on the fixed ensemble and does not estimate variability across future training runs. Architecture, optimization, regularization and ensembling jointly define the fitted methods; matched features do not isolate nonlinearity as a causal mechanism. Equal additive score penalties give zero interaction; model-dependent responses to target noise can give nonzero interaction.

Original Pearson comparison. For Linear-GenePT against Perturb Mean, the RPE1 Perturb Mean-calibrated interaction is -0.084773 [-0.163684, -0.020540], on 60 genes. This is the original 24-contrast adjusted interval, wholly below the prespecified -0.02 margin; it is not reconstructed from the later neural comparison.

Neural inference family. Within each fold and metric, use the common finite gene set for N (fMLP), G (Linear-GenePT) and B (Perturb Mean), against targets from the construct present in source data R and the construct withheld from all source data U. All eligible genes remain finite. MAE penalties are $P_m = U_m - R_m$, with interactions $I_{NB} = P_N - P_B$ and $I_{NG} = P_N - P_G$. Pearson penalties are $R_m - U_m$; its interactions instead use $k = B_U/B_R$, $J_{NB} = kN_R - N_U$ and $J_{NG} = (kN_R - N_U) - (kG_R - G_U)$. The ratio gate requires $B_R > 0.05$ in the sample and every bootstrap draw. Primary HepG2 fails that gate; its two Pearson interaction slots remain unavailable, not zero.

Three method penalties plus two interactions, across two metrics and four folds, form 40 primary slots. Paired-gene resampling shares indices across methods and contrasts within a fold/metric, with 20,000 draws and seed 1707. Approximate simultaneous percentile intervals use tails $0.05/(2 \times 40) = 0.000625$. Signed materiality requires the entire interval above +0.005 or below -0.005 for MAE, and above +0.02 or below -0.02 for Pearson. An interval excluding zero need not meet that criterion. Failed ratio slots remain in the family. The earlier 24-contrast family is retained separately.

Complete primary outcomes. Tables A12 and A13 report every neural-family slot, including unavailable entries. Exactly the K562 and HepG2 neural-minus-linear MAE intervals clear their materiality margin among the new primary interactions. RPE1 and Jurkat do not. No fMLP–Perturb Mean primary interaction clears its margin; this is not an equivalence conclusion. Paired interactions can be more precise than either method's own penalty because they exploit within-gene covariance.

Strict reciprocal outcomes. Average two orientation-specific scores within gene before resampling, retaining 22 genes per fold. The neural-minus-linear MAE estimates and descriptive 95% intervals are K562 +0.012341 [0.004197, 0.022410], RPE1 +0.014783 [0.005718, 0.025518], Jurkat +0.014496 [0.005960, 0.024624], and HepG2 +0.017776 [0.007559, 0.029357]. Only the latter three intervals lie wholly above +0.005. K562 and Jurkat strict Pearson interactions are unavailable because respectively 2 and 6 bootstrap draws fail the denominator gate. Reciprocal intervals are descriptive sensitivity, not further primary tests or independent replications.

Same construct control. The control retains the primary genes, cassette orientations and frozen fMLP/GenePT-ridge predictions. The same construct control rules were fixed before its outcomes, with the original interactions already known. For held-out counts n_R, n_U, set $n_P = \min(n_U, n_R - 20)$. Partition represented cells into disjoint R' of size n_P and R_0 containing the remainder; subsample U_d to the same n_P. Counts are pooled before full-axis normalization over 6,642 response genes; the pseudocount is 0.05 and the held-out non-targeting reference is unchanged. No model is refitted.

Let $g(T) = \mathrm{MAE}(N, T) - \mathrm{MAE}(G, T)$. The same construct control is $L = g(R') - g(R_0)$, the resampled cross-construct interaction is $C = g(U_d) - g(R_0)$, and the matched construct transfer contrast is $D = C - L = g(U_d) - g(R')$. The historical full-target interaction minus the same construct control is $H = I_{NG} - L$; its changed baseline prevents interpreting it as a pure construct excess. Average 20 seeded partitions within gene, then weight genes equally within fold. Paired-gene bootstrap intervals use 20,000 draws, seed 1707 and tails 0.000625/0.999375, retaining the conservative 40-slot adjustment. These control intervals do not enlarge or replace the original neural family. The materiality margin is +0.005; Monte Carlo SE above 0.0005 suppresses a materiality verdict. All folds pass the precision gate. Inference conditions on frozen predictions, observed cells and the partition average.

The same construct control L point estimates are non-positive, whereas all matched construct transfer D intervals exclude zero (Table A7; Figure 2D); only K562 clears the materiality margin. This weakens the tested target-resampling explanation without separating exposure from cassette identity, efficacy or preparation. Exact historical U depth is attainable for 28/57, 31/60, 23/49 and 5/31 genes; other genes use the registered common cap. The exact-U-depth sensitivity is inferential only in RPE1, with $D = 0.009615$ [0.000738, 0.020887], unresolved for materiality. The other three subsets are descriptive only.

Signed cell-count and target-gene LFC diagnostics. On the unchanged primary pairs, report $n_U - n_R$ and the target coordinate of $T(x_U) - T(x_R)$, using full-axis count normalization and the common held-out control. Positive target-gene LFC difference means U is less downregulated, not automatically less efficacious. Missing target coordinates remain missing. Means weight genes equally; intervals are descriptive paired-gene 95% percentile intervals from 20,000 draws with seed 1707, without pooled-fold inference. U has fewer cells on average in every fold, with resolved signed cell-count differences in K562 and RPE1. No target-LFC mean interval excludes zero (Table A8). These diagnostics establish coverage asymmetry, not efficacy equivalence. The separate strict reciprocal diagnostics retain both orientations: opposite signed differences are a consistency check, not independent evidence of balance; absolute differences are reported once per pair (Table A9).

Matched-cell-count absolute performance. For each of the four frozen predictors, use the same 20 partitions of targets from the construct present in source data R' and withheld construct U_d defined above on the unchanged 57/60/49/31 primary genes. Both targets contain n_P cells. Counts are pooled before CP10K normalization over all 6,642 coordinates, with pseudocount 0.05 and the same held-out control. Define $A_P = \mathrm{Pearson}(\hat{y}, R') - \mathrm{Pearson}(\hat{y}, U_d)$ and $A_M = \mathrm{MAE}(\hat{y}, U_d) - \mathrm{MAE}(\hat{y}, R')$, where the target symbols denote transformed profiles. Positive values favor targets from the construct present in source data. Average partitions within gene, then weight genes equally within fold. These absolute quantities differ from $D = g(U_d) - g(R')$, the relative N–G gap change; algebraically $A_M(N) - A_M(G) = D$.

The matched-cell-count endpoints were preregistered before target access; the earlier full-target and relative-control analyses had already been completed. All targets and predictions were finite on all frozen genes; no row was dropped. One 32-contrast family uses 20,000 paired-gene bootstrap draws, seed 1707 restarted per fold, shared indices across methods/metrics, and linear percentile tails 0.00078125/0.99921875 for approximate simultaneous 95% intervals. Inference conditions on frozen predictions, observed cells, the common control and the 20-partition average. An interval entirely above zero resolves a represented advantage; entirely below zero resolves an excluded advantage; otherwise it is uncertain. No new materiality threshold is applied.

Tables A5 and A6 show represented-favoring mean Pearson in 16/16 comparisons and mean MAE in 15/16. Eight Pearson and five MAE intervals resolve that advantage; none resolve the opposite. The RPE1 Linear-GenePT MAE advantage is -0.003546 $[-0.047190, +0.041085]$, uncertain. These are finite-set counts, not independent fold replications. Matching at the common cap does not estimate performance at arbitrary depth or separate exposure from cassette identity, efficacy or preparation.

Table A5: Matched-cell-count Pearson absolute performance. Equal-gene means after averaging 20 partitions per gene. B: Perturb Mean; G: Linear-GenePT; N: fMLP-GenePT-LFC; P: Linear-PCA. Positive advantage favors targets from the construct present in source data. R: resolved advantage for targets from the construct present in source data; ?: uncertain; U: resolved advantage for targets from the withheld construct. Intervals use the separate 32-contrast family. Cohorts: K562 57, RPE1 60, Jurkat 49, HepG2 31.

Fold	Method	$R\prime$ mean	U_d mean	Advantage	Adjusted interval	Result
K562	B	0.106761	0.008349	+0.098412	[+0.040371, +0.168039]	R
K562	G	0.064742	0.048845	+0.015897	[-0.058733, +0.091923]	?
K562	N	0.079886	0.036974	+0.042913	[-0.030066, +0.119195]	?
K562	P	0.077808	0.029220	+0.048587	[-0.023755, +0.126977]	?
RPE1	B	0.142301	0.022204	+0.120097	[+0.053900, +0.196002]	R
RPE1	G	0.174186	0.111695	+0.062491	[-0.064563, +0.187932]	?
RPE1	N	0.178610	0.071114	+0.107496	[+0.006053, +0.212961]	R
RPE1	P	0.210358	0.086949	+0.123409	[+0.002940, +0.250172]	R
Jurkat	B	0.127280	0.012924	+0.114357	[+0.058169, +0.180931]	R
Jurkat	G	0.078160	0.038898	+0.039262	[-0.033178, +0.115658]	?
Jurkat	N	0.097064	0.026010	+0.071054	[+0.003819, +0.146769]	R
Jurkat	P	0.092178	0.020469	+0.071710	[+0.001498, +0.149681]	R
HepG2	B	0.127180	0.013248	+0.113931	[+0.008674, +0.243206]	R
HepG2	G	0.109780	0.096452	+0.013327	[-0.160535, +0.185107]	?
HepG2	N	0.135943	0.084366	+0.051577	[-0.116880, +0.232618]	?
HepG2	P	0.137017	0.094453	+0.042564	[-0.142928, +0.243621]	?

F SHARED-REFERENCE DIAGNOSTIC

Shared control references increase apparent concordance. A construct profile is a log-fold change against a control-reference estimate. When two profiles reuse the same reference estimate, its sampling error enters both with the same sign. Reagent-pair concordance then includes a covariance term from the reference, even though that term says nothing about how similarly the two constructs respond. The mechanism is known from prior analyses of control-referenced correlations Nicol et al. (2026); Fukutani et al. (2026); here we intervene on the control construction for pairs of reagents targeting the same gene.

The intervention isolates the artifact at matched control-cell count. On the five panels, halving a shared control pool raises median apparent pair agreement; giving each construct its own independent half lowers it (Figure A3). The paired median reduction is 0.0126–0.0242, with perturbation-bootstrap intervals above zero on every panel. On the selected HepG2 pairs the reduction is approximately as large as the half-pool shared-reference concordance, whose denominator is near zero; the quoted share is a ratio of medians, not a per-pair biological fraction. The CRISPRa panel is a same-spacer positional control, not a further different-reagent replication.

A covariance identity explains the direction. With two construct profiles $y_c = \beta + \delta_c + \varepsilon_c - r$ and equal-depth exchangeable references, replacing a shared r with disjoint r_1, r_2 changes their

Table A6: Matched-cell-count MAE absolute performance. Equal-gene means after averaging 20 partitions per gene. B: Perturb Mean; G: Linear-GenePT; N: fMLP-GenePT-LFC; P: Linear-PCA. Positive advantage favors targets from the construct present in source data. R: resolved advantage for targets from the construct present in source data; ?: uncertain; U: resolved advantage for targets from the withheld construct. Intervals use the separate 32-contrast family. Cohorts: K562 57, RPE1 60, Jurkat 49, HepG2 31.

Fold	Method	$R\prime$ mean	U_d mean	Advantage	Adjusted interval	Result
K562	B	0.204658	0.223519	+0.018861	[+0.008730, +0.031802]	R
K562	G	0.198068	0.199596	+0.001528	[-0.008058, +0.012641]	?
K562	N	0.182535	0.195815	+0.013280	[+0.001878, +0.027021]	R
K562	P	0.186589	0.199035	+0.012446	[+0.001974, +0.025701]	R
RPE1	B	0.259143	0.271828	+0.012685	[-0.027810, +0.057207]	?
RPE1	G	0.250057	0.246511	-0.003546	[-0.047190, +0.041085]	?
RPE1	N	0.245341	0.251156	+0.005814	[-0.038101, +0.053250]	?
RPE1	P	0.241813	0.251052	+0.009239	[-0.033482, +0.055676]	?
Jurkat	B	0.221805	0.240749	+0.018944	[+0.006124, +0.035144]	R
Jurkat	G	0.214725	0.216013	+0.001289	[-0.014411, +0.017805]	?
Jurkat	N	0.204297	0.216543	+0.012246	[-0.004079, +0.031603]	?
Jurkat	P	0.207400	0.218693	+0.011293	[-0.004804, +0.030143]	?
HepG2	B	0.243698	0.269492	+0.025794	[+0.001854, +0.058500]	R
HepG2	G	0.243840	0.245322	+0.001482	[-0.027423, +0.030696]	?
HepG2	N	0.231560	0.250244	+0.018684	[-0.007552, +0.051392]	?
HepG2	P	0.234877	0.250431	+0.015554	[-0.007287, +0.047916]	?

Table A7: Same construct control and matched construct transfer comparison. Equal-gene means after within-gene partition averaging; adjusted paired-gene intervals. D compares endpoints at equal n_P; H compares with the historical full-target interaction. Only K562 clears +0.005 for D.

Fold	n	Contrast	Estimate	Adjusted interval
K562	57	L	-0.004885	[-0.008429, -0.001471]
K562	57	C	+0.006867	[-0.000501, +0.015026]
K562	57	D	+0.011752	[+0.005021, +0.019553]
K562	57	H	+0.018996	[+0.010126, +0.028768]
RPE1	60	L	-0.000072	[-0.001703, +0.001681]
RPE1	60	C	+0.009288	[+0.000845, +0.018490]
RPE1	60	D	+0.009360	[+0.000963, +0.018701]
RPE1	60	H	+0.012688	[+0.002035, +0.024855]
Jurkat	49	L	-0.002057	[-0.005434, +0.001761]
Jurkat	49	C	+0.008900	[+0.001150, +0.018262]
Jurkat	49	D	+0.010957	[+0.003334, +0.019977]
Jurkat	49	H	+0.015764	[+0.006590, +0.025967]
HepG2	31	L	-0.002836	[-0.006226, -0.000136]
HepG2	31	C	+0.014366	[+0.002505, +0.027637]
HepG2	31	D	+0.017202	[+0.004641, +0.031376]
HepG2	31	H	+0.022337	[+0.007223, +0.039298]

Table A8: Signed cell-count and target-gene LFC diagnostics on the exact primary pairs. Differences are $U - R$; intervals are descriptive 95%, not the neural or same-construct-control simultaneous family. Target-LFC summaries use available target coordinates only; A/M is available/missing.

Fold	n	Cell difference [95% CI]	A/M	Target-LFC difference [95% CI]
K562	57	-33.12 [-62.67, -5.63]	44/13	+0.220 [-0.322, +0.755]
RPE1	60	-53.13 [-122.15, -3.28]	48/12	+0.146 [-0.441, +0.721]
Jurkat	49	-52.76 [-131.37, +1.82]	38/11	+0.118 [-0.492, +0.716]
HepG2	31	-5.74 [-26.90, +10.13]	22/9	-0.553 [-1.351, +0.284]

Table A9: Strict reciprocal diagnostics, descriptive means. Each fold has 22 unique pairs; target LFC is available for 16 and missing for six. Signed differences retain each orientation; absolute differences are reported once per pair. Reversed signs do not establish balance.

Fold	Quantity	Forward $U - R$	Reverse $U - R$	Absolute difference
K562	Cells	-4.409	+4.409	74.409
K562	Target LFC	-0.807	+0.807	1.801
RPE1	Cells	-26.409	+26.409	110.227
RPE1	Target LFC	-0.737	+0.737	1.960
Jurkat	Cells	-6.045	+6.045	97.591
Jurkat	Target LFC	-0.916	+0.916	1.913
HepG2	Cells	-7.682	+7.682	32.682
HepG2	Target LFC	-0.537	+0.537	1.945

Table A10: Primary Pearson-Δ scores. Means for targets from the construct present in source data, the count-pooled selected pair and the withheld construct are descriptive, with equal gene weighting. All use the same frozen predictions and scoring rules. Linear-PCA is secondary.

Cell type	Method	n	Represented	Pooled	Withheld
K562	fMLP-GenePT-LFC	57	0.097961	0.114128	0.042875
K562	Linear-GenePT	57	0.082942	0.116173	0.057095
K562	Perturb Mean	57	0.130550	0.113774	0.011728
K562	Linear-PCA	57	0.099287	0.113141	0.035457
RPE1	fMLP-GenePT-LFC	60	0.194502	0.195507	0.076564
RPE1	Linear-GenePT	60	0.195569	0.228159	0.118560
RPE1	Perturb Mean	60	0.160558	0.137636	0.027738
RPE1	Linear-PCA	60	0.231195	0.233245	0.092422
Jurkat	fMLP-GenePT-LFC	49	0.108961	0.100571	0.028682
Jurkat	Linear-GenePT	49	0.089738	0.102982	0.044604
Jurkat	Perturb Mean	49	0.149939	0.126051	0.017457
Jurkat	Linear-PCA	49	0.104326	0.097106	0.024164
HepG2	fMLP-GenePT-LFC	31	0.152271	0.203090	0.101441
HepG2	Linear-GenePT	31	0.127286	0.197468	0.117049
HepG2	Perturb Mean	31	0.142062	0.124630	0.015999
HepG2	Linear-PCA	31	0.156280	0.212470	0.111893

Table A11: Primary MAE-Δ scores. Means for targets from the construct present in source data, the count-pooled selected pair and the withheld construct are descriptive, with equal gene weighting. All use the same frozen predictions and scoring rules. Linear-PCA is secondary.

Cell type	Method	n	Represented	Pooled	Withheld
K562	fMLP-GenePT-LFC	57	0.151942	0.123740	0.175098
K562	Linear-GenePT	57	0.169754	0.137081	0.178799
K562	Perturb Mean	57	0.177421	0.157368	0.204680
K562	Linear-PCA	57	0.156392	0.128939	0.178482
RPE1	fMLP-GenePT-LFC	60	0.200929	0.140840	0.223328
RPE1	Linear-GenePT	60	0.207366	0.140267	0.217150
RPE1	Perturb Mean	60	0.218392	0.167351	0.246882
RPE1	Linear-PCA	60	0.197145	0.137772	0.223223
Jurkat	fMLP-GenePT-LFC	49	0.171441	0.135562	0.193494
Jurkat	Linear-GenePT	49	0.183947	0.142703	0.192292
Jurkat	Perturb Mean	49	0.192018	0.165232	0.220331
Jurkat	Linear-PCA	49	0.175468	0.139830	0.196349
HepG2	fMLP-GenePT-LFC	31	0.184668	0.139117	0.209569
HepG2	Linear-GenePT	31	0.199317	0.144571	0.204717
HepG2	Perturb Mean	31	0.200025	0.164425	0.231876
HepG2	Linear-PCA	31	0.188769	0.142741	0.210428

Table A12: All primary neural-comparison MAE contrasts. Penalties are positive for degradation; interaction definitions are given above. Equal-gene, per-fold estimates and 40-family intervals; unavailable ratio-gated slots retain their place in the family.

Cell type	Contrast	n	Estimate	Adjusted interval
K562	P_N	57	+0.023156	[+0.003178, +0.043464]
K562	P_G	57	+0.009045	[-0.010433, +0.027609]
K562	P_B	57	+0.027259	[+0.008661, +0.045294]
K562	$N - B$	57	-0.004103	[-0.011292, +0.001405]
K562	$N - G$	57	+0.014111	[+0.006289, +0.023391]
RPE1	P_N	60	+0.022400	[-0.041559, +0.087582]
RPE1	P_G	60	+0.009784	[-0.054607, +0.072197]
RPE1	P_B	60	+0.028490	[-0.028664, +0.087409]
RPE1	$N - B$	60	-0.006090	[-0.015418, +0.001437]
RPE1	$N - G$	60	+0.012616	[+0.002251, +0.024225]
Jurkat	P_N	49	+0.022052	[-0.005006, +0.051449]
Jurkat	P_G	49	+0.008345	[-0.017152, +0.035544]
Jurkat	P_B	49	+0.028313	[+0.004951, +0.053708]
Jurkat	$N - B$	49	-0.006261	[-0.015202, +0.000377]
Jurkat	$N - G$	49	+0.013708	[+0.004907, +0.023821]
HepG2	P_N	31	+0.024901	[-0.009145, +0.062994]
HepG2	P_G	31	+0.005401	[-0.031192, +0.041201]
HepG2	P_B	31	+0.031851	[+0.000506, +0.070588]
HepG2	$N - B$	31	-0.006950	[-0.016716, -0.000208]
HepG2	$N - G$	31	+0.019500	[+0.005380, +0.034857]

Table A13: All primary neural-comparison Pearson contrasts. Penalties are positive for degradation; interaction definitions are given above. Equal-gene, per-fold estimates and 40-family intervals; unavailable ratio-gated slots retain their place in the family.

Cell type	Contrast	n	Estimate	Adjusted interval
K562	P_N	57	+0.055086	[-0.029750, +0.146973]
K562	P_G	57	+0.025848	[-0.058411, +0.113927]
K562	P_B	57	+0.118823	[+0.050394, +0.201409]
K562	$N - B$	57	-0.034075	[-0.074833, -0.003410]
K562	$N - G$	57	+0.015569	[-0.000811, +0.034563]
RPE1	P_N	60	+0.117938	[+0.012013, +0.231449]
RPE1	P_G	60	+0.077009	[-0.056169, +0.208105]
RPE1	P_B	60	+0.132820	[+0.064131, +0.214461]
RPE1	$N - B$	60	-0.042962	[-0.096970, +0.001720]
RPE1	$N - G$	60	+0.041812	[-0.003955, +0.096284]
Jurkat	P_N	49	+0.080278	[+0.004303, +0.164382]
Jurkat	P_G	49	+0.045134	[-0.033277, +0.129921]
Jurkat	P_B	49	+0.132482	[+0.068222, +0.206427]
Jurkat	$N - B$	49	-0.015997	[-0.052388, +0.015039]
Jurkat	$N - G$	49	+0.018160	[-0.001539, +0.038580]
HepG2	P_N	31	+0.050830	[-0.138159, +0.253396]
HepG2	P_G	31	+0.010237	[-0.190650, +0.205783]
HepG2	P_B	31	+0.126063	[+0.011154, +0.267320]
HepG2	$N - B$	31	—	unavailable
HepG2	$N - G$	31	—	unavailable

covariance by $R = \mathrm{Var}(r)$, while leaving their marginal variances V_1, V_2 unchanged. The idealized Pearson inflation is therefore $R/\sqrt{V_1 V_2}$, regardless of the common or different construct components.

The direct disjoint-reference comparison necessarily uses half the available control pool. Any quoted full-pool reduction is therefore an *estimate* extrapolated from measured half-pool behavior, not a measured disjoint full-pool statistic. Reporting reagent-pair concordance requires naming whether the compared profiles reused one reference estimate; when that choice changes, the biological comparison has not changed but the measured correlation has.

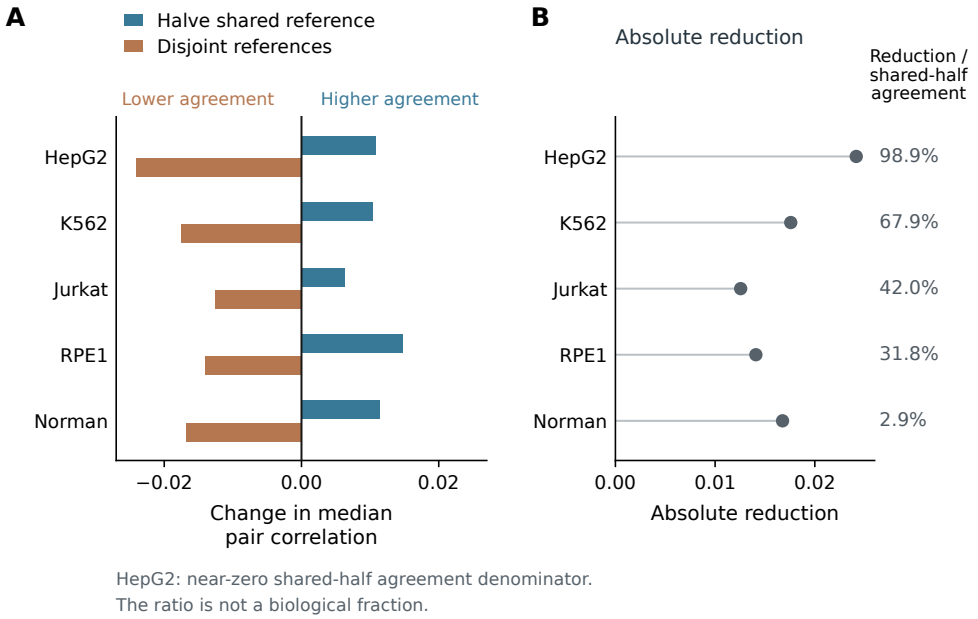

Figure A3: **Shared control references change apparent agreement.** *Left:* both directions the mechanism predicts, on five panels — halving the shared control pool makes the reference noisier and *raises* apparent agreement, while giving each construct its own half lowers it. *Right:* the reduction is of the same order on every panel (0.0126–0.0242) while the ratio of median reduction to median shared-half agreement ranges from 2.9% to 98.9% across these selected pairs. This ratio has a near-zero denominator on HepG2 and is not a per-pair biological fraction. Five panels, each panel's own gene set, $\varepsilon = 0.05$, constructs of ≥ 40 cells.

G METHOD DETAIL

Disjoint split controls (§2). The construct scheme partitions the control constructs by their own draw; the random scheme draws a side per cell over all cells, including controls. Both form two disjoint control estimates rather than sharing one, so the shared-reference term in Appendix F does not enter the split coefficients §3 reports.

Full-panel random-split reference. Table A15 compares mean and median aggregation on each panel's own response-gene set and perturbation cohort; the multi-construct coefficients use the distinct cohort and axis specified in §3.

Convergent cross-experiment evidence. Across 2,053 matched K562 perturbations, random-split reliability at ε=0.05 tracks Essential-versus-GenomeWide agreement using the raw DESeq2-profile Pearson field (rawcorr_pearson) in Nadig et al. Supplementary Table 9 Nadig et al. (2025), selected as the criterion for our analysis: Spearman ρ=0.880; partial rank correlation controlling log cell count, 0.863. K562-Essential contributes to both quantities and the experiments share the reagent library, so this is convergent evidence rather than independent validation. This K562 overlap is not the four-panel intersection of the same size.

Aggregation and cohort. The Spearman–Brown transformation Spearman (1910); Brown (1910) supplies historical attribution, not parallel-form validity for different constructs. Raw per-perturbation Pearson correlations are averaged arithmetically over perturbations and splits before applying $2r/(1+r)$. The unrestricted aggregate coefficients quoted in §3 use the same eligible targets in both arms without matching cells per side. Figure 1B–D use the coverage-filtered matched-cell-count split comparison, while Figure 1F uses the separate unrestricted full/high cohorts. The primary $\geq$40-cell comparison in §3 matches both target membership and sampled cell counts; Table A1 uses a different estimator/cohort from Table A15.

Observed construct-ID eligibility. Eligibility is determined from distinct observed construct IDs within a target, before the cell-count restriction for the cell-count-restricted comparisons. The observed assignment records do not establish why these targets received extra constructs. The numerical comparison is scoped to this eligible set (§2).

Three-construct route. The three-construct comparison splits one shared construct into disjoint cell halves and correlates its difference from each of two other constructs, averaging over which construct is shared. A common control pool cancels from both difference vectors. This tests recurrence of construct-associated deviation, not individual-spacer identity. Its shared set comprises 19 genes across three cell lines; it is retained here as supporting recurrence evidence and does not isolate individual-spacer effects.

Measured target-gene LFC gap alone also leaves a difference. On RPE1 and Jurkat, construct pairs in the bottom quartile of measured target-gene LFC gaps have median gaps 0.21 and 0.22 $\log_2$ units; their reproducible construct-difference variance shares are 57% and 44% with intervals above null. The high-gap quartiles show larger reproducible shares; this association does not isolate an efficacy mechanism. Measured target-gene LFC is only one proxy for efficacy. Full bands and the larger-gap K562 bottom quartile are in Figure A2.

Recurrence reference. The mean-profile positive control retains a shared reference term that cancels from the construct difference. The residual analysis concerns cross-line comparisons and was not performed on the second K562 experiment.

Transported scalar alignment of construct differences. For each perturbation, center its construct difference d and mean-effect profile m across response genes, fit $\alpha = (d^\top m)/(m^\top m)$, and analyze $r = d - \alpha m$. The primary test fits α in one cell line and transports it to another. Residual agreement exceeds the mismatched-pair null in all six comparisons, with perturbation-bootstrap intervals above zero (Table A14). Across panels, the median fitted squared-norm share is 8.5–18.5%; the mean is larger, reflecting an aligned tail. Separately fitted residuals provide a sensitivity check, exceeding their null by 0.040–0.066. This is residual agreement beyond a fitted scalar component, not identification of a dose or efficacy mechanism. The analysis was not performed on the second K562 experiment.

Table A14: Mean across-perturbation Pearson correlation of the original construct-pair difference and the primary residuals after transported scalar alignment. Above null subtracts mismatched-pair correlation; intervals bootstrap perturbations. The six comparisons reuse pairs and are not independent.

cell lines	n	original	transported	above null	95% CI
K562–RPE1	115	0.088	0.067	0.069	[0.053, 0.085]
K562–Jurkat	115	0.093	0.066	0.071	[0.057, 0.088]
K562–HepG2	115	0.085	0.061	0.062	[0.048, 0.078]
RPE1–Jurkat	127	0.082	0.043	0.045	[0.036, 0.055]
RPE1–HepG2	129	0.143	0.045	0.043	[0.031, 0.057]
Jurkat–HepG2	128	0.063	0.039	0.036	[0.025, 0.047]

Cross-context comparison cohorts. The 0.147 maximum in §3 is RPE1–HepG2 on 116 genes; Table A14 reports 0.143 on the 129-gene subset used for its scaling comparison.

Input-gene feature availability determines the final cohort. The metadata rule gives 59, 63, 51 and 33 eligible genes on K562, RPE1, Jurkat and HepG2. Requiring a GenePT embedding for the perturbed input gene removes two, three, two and two genes, respectively, leaving 57, 60, 49 and 31. This deterministic feature rule was fixed before experiment outcomes; there is no embedding

imputation, replacement gene or outcome-based pair reselection. The same rule reduces the common training universe from 2,053 to 2,033 genes. It leaves the scored response axis unchanged: feature eligibility concerns the perturbed gene used as input, not the expression genes used as outputs. This is construct transfer for represented genes across cell types, not unseen-gene generalization.

Count-pooled versus separately scored targets. The fixed-prediction comparison in §4 (Pearson summary: Figure 2A) uses 96, 86, 79 and 33 eligible perturbations on K562, RPE1, Jurkat and HepG2. Each has at least two observed constructs with at least 40 cells. For the separately scored construct target, score every eligible construct separately, average equally over constructs within gene, then equally over genes. For the pooled target, pool all cells for the gene before normalization and scoring, including constructs below the per-construct 40-cell threshold. Both use the same held-out non-targeting control, the 6,642 shared genes, counts per 10,000 and $\log_2$ fold change with $\varepsilon = 0.05$, without SB transformation. On the same 96 K562 genes, Linear-GenePT's pooled and separately scored construct mean Pearson values are 0.120474 and 0.071669, and its mean MAEs are 0.143600 and 0.182295. Cell count, mixture weighting and aggregation jointly define the comparison. Point ordering is retained for both metrics in every fold. The pooled target here is an all-construct count pool, unlike the selected-pair count pool used descriptively in §5.

Table A15: Split-half reliability with the Spearman–Brown correction, random split, split control, $\varepsilon = 0.05$, over 20 splits, on each panel's own gene set and perturbation cohort. n is each panel's own perturbation count. The two columns differ only in whether perturbations are averaged by mean or by median. The choice reorders the panels, which is why the aggregation and cohort must be specified. These full-panel values are not the multi-construct or fitted-predictor cohorts used elsewhere.

Panel	n	mean-aggregated	median-aggregated
K562	2,057	0.3333	0.2011
RPE1	2,393	0.4614	0.4175
Jurkat	2,393	0.2731	0.2000
HepG2	2,393	0.3179	0.1727
CRISPRa (Norman)	236	0.6810	0.7038

H External guide-realization validation

External target variation: shared question, distinct instruments. These analyses ask whether two distinct single-guide reagents targeting the same gene yield realized expression profiles that disagree beyond same-guide cell resampling. They do not evaluate model predictions or impose source-construct exclusion. The common estimator family operates on expression-profile differences, whereas the released-model analysis in Appendix I operates on differences between prediction scores. The core panels' SB coefficients are a third quantity. We report each experiment in its own response space and inferential family, without pooling estimates across datasets or modalities.

Matched-cell-count profile comparison. For a selected guide pair A, B, set $m = \lfloor \min(n_A, n_B)/2 \rfloor$. Each of 20 frozen partitions creates disjoint halves A_1, A_2 and B_1, B_2, each containing m cells. Sum counts before normalization and define $T_j = \log_2(10^4 c_j / N + 0.05)$ on the registered response coordinates, with N summed over the full deposited gene axis. The partition-level excess is the mean squared cross-guide profile difference minus one half of the sum of the two same-guide half-split mean squared differences. Average partitions within each selected pair and then weight targeting genes equally. A common reference cancels from profile differences and is not subtracted in these analyses. Equal cell counts do not imply equal UMI totals.

Response coordinates and inferential units. Each response space contains 2,000 genes selected by mean per-cell CP10K expression in the designated control-guide cells; the selection rule and ascending gene-name/column-index tie-breaking were fixed before outcomes, and the realized coordinate lists were sealed before targeting pseudobulks. Control populations differ across experiments (Table A16). Frangieh uses non-targeting and intergenic controls; Drepanos uses modality-specific intergenic controls at K562 Day 14 and pooled controls at A549 Day 7; Jiang uses non-targeting controls. Differences in coordinate membership and populations prevent interpretation of raw cross-system effect-size differences. Targeting uncertainty resamples genes. Frangieh's intergenic cutting/guide-identity comparison resamples disjoint preassigned control-guide pairs; Drepanos's

intergenic comparisons and Jiang's non-targeting construct-identity comparisons instead resample control guides because their all-pairs control comparisons reuse guides. The numbers of overlapping pairs are not their effective control sample sizes.

Table A16: External screens test the same target-specification question with different control instruments. Rows describe only the analyzed cohorts, not every available dataset stratum. Frangieh uses disjoint intergenic cutting/guide-identity pairs; Drepanos uses ten intergenic guides with modality-specific attribution at K562 Day 14 and a pooled, modality-ambiguous control at A549 Day 7; Jiang uses fourteen non-targeting guides. Response coordinates are selected separately from control cells. These instruments and coordinate spaces are not interchangeable, and the screens evaluate no prediction model.

Dataset / condition	Reagent / effector	Cohort / guide threshold	Response space	Control instrument / unit	Stratum role
Frangieh IFN-gamma	Cas9 KO	205 genes; exact single guide; $\geq$30 cells/guide	2,000 genes from non-targeting + intergenic cells	17 disjoint intergenic pairs	primary
Frangieh Control	Cas9 KO	182 genes; exact single guide; $\geq$30 cells/guide	2,000 genes from non-targeting + intergenic cells	14 disjoint intergenic pairs	secondary
Frangieh Co-culture	Cas9 KO	195 genes; exact single guide; $\geq$30 cells/guide	2,000 genes from non-targeting + intergenic cells	16 disjoint intergenic pairs	descriptive
Drepanos K562 D14 KO	Cas9 KO	23 genes; dominant guide; deposited 100-cell filter	2,000 genes from intergenic cells	10 guides; modality-specific	primary
Drepanos K562 D14 CRISPRi	ZIM3-KRAB CRISPRi	22 genes; dominant guide; deposited 100-cell filter	2,000 genes from intergenic cells	10 guides; modality-specific	primary
Drepanos A549 D7 KO	Cas9 KO	24 genes; dominant guide; deposited 100-cell filter	2,000 genes from intergenic cells	10 guides; pooled, modality-ambiguous	secondary
Drepanos A549 D7 CRISPRi	ZIM3-KRAB CRISPRi	24 genes; dominant guide; deposited 100-cell filter	2,000 genes from intergenic cells	10 guides; pooled, modality-ambiguous	secondary
Jiang HT29/TNFA	Dolcetto / KRAB-MeCP2 CRISPRi	44 genes; deposited guide; $\geq$30 cells/guide	2,000 genes from non-targeting cells	14 non-targeting guides	primary
Jiang HAP1/IFNB	Dolcetto / KRAB-MeCP2 CRISPRi	51 genes; deposited guide; $\geq$30 cells/guide	2,000 genes from non-targeting cells	14 non-targeting guides	secondary
Jiang HT29/IFNB	Dolcetto / KRAB-MeCP2 CRISPRi	51 genes; deposited guide; $\geq$30 cells/guide	2,000 genes from non-targeting cells	14 non-targeting guides	secondary
Jiang MCF7/IFNB	Dolcetto / KRAB-MeCP2 CRISPRi	51 genes; deposited guide; $\geq$30 cells/guide	2,000 genes from non-targeting cells	14 non-targeting guides	descriptive

H.1 Independent knockout-library evidence. Frangieh et al.'s Perturb-CITE-seq screen Frangieh et al. (2021) supplies an independent Cas9-knockout library in patient-derived melanoma cells. We retain exact single-guide assignments using the original sgRNA field, require at least 30 eligible cells per guide and select the lexicographically first pair of guides differing at at least three nucleotide positions. IFN-gamma is primary, Control is inferential secondary and Co-culture is descriptive. The conditions share one transduced pool and are not independent replications. The response spaces are selected separately for each condition.

For Frangieh, the frozen endpoint identifiers are targeting excess (E1), intergenic cutting/guide-identity comparison (E2), and targeting-minus-intergenic comparison (E3). The IFN-gamma primary targeting excess is 0.00200063 [0.00100929, 0.00304332] over 205 genes. The Control secondary targeting excess is 0.00364499 [0.00189992, 0.00547924] over 182 genes. The targeting-minus-intergenic comparisons are -0.000375545 [-0.00711704, 0.00521499] and 0.000272466 [-0.00502703, 0.00527581], respectively. The targeting excess is therefore not distinguished from the intergenic cutting/guide-identity comparison; this does not establish equivalence or show that cutting explains it. Only 17 and 14 disjoint intergenic control pairs are usable, making the control comparison substantially less precise than the targeting mean. Non-targeting-guide comparisons are descriptive.

The primary targeting excess remains positive in all registered targeting sensitivities (Table A20), including the higher-depth subset, all-pairs aggregation, sequence-overlap exclusions and removal of the three largest gene contributions. Sensitivity of the targeting-minus-intergenic comparison was not registered. The IFN-gamma excess is approximately 1.5% of its same-guide resampling term at the registered depth; its largest gene contributes 11.9% of the net sum. Signs of noisy per-gene estimates do not estimate prevalence. The 744 targeting spacers have no overlap with the original ST20 reagents; some overlap with other design libraries, and shared Perturb-seq methodology remains. This evidence extends the beyond-resampling finding, not the recurrence or model-transfer experiments.

Complete endpoints for this reanalysis of Frangieh et al. (2021) are in Table A17.

Table A17: Complete Frangieh guide-realization endpoints. Estimates average twenty equal-cell-count partitions within each unit and weight targeting genes or disjoint control pairs equally. IFNγ targeting excess is primary; Control targeting excess and the IFNγ/Control targeting-minus-intergenic comparisons form the three-slot secondary family. Other endpoints are descriptive. Primary/descriptive percentile tails are 0.025/0.975; secondary tails are 0.0083333/0.9916667. The unresolved targeting-minus-intergenic comparisons do not establish equivalence to the intergenic cutting/guide-identity comparison. Conditions share one transduced pool.

Condition	Endpoint	Population n	Estimate [interval]	Family	Status
IFN-gamma	Targeting excess (E1)	205 genes	0.002001 [0.001009, 0.003043]	primary	positive
IFN-gamma	Intergenic cutting / guide-identity comparison (E2)	17 pairs	0.002376 [-0.002115, 0.007680]	descriptive	uncertain
IFN-gamma	Targeting - intergenic comparison (E3)	205 genes; 17 pairs	-0.000376 [-0.007117, 0.005215]	secondary	uncertain
IFN-gamma	Non-targeting comparison (E4)	18 pairs	0.001826 [-0.000207, 0.004096]	descriptive	uncertain
Control	Targeting excess (E1)	182 genes	0.003645 [0.001900, 0.005479]	secondary	positive
Control	Intergenic cutting / guide-identity comparison (E2)	14 pairs	0.003373 [-0.000473, 0.007437]	descriptive	uncertain
Control	Targeting - intergenic comparison (E3)	182 genes; 14 pairs	0.000272 [-0.005027, 0.005276]	secondary	uncertain
Control	Non-targeting comparison (E4)	18 pairs	0.001207 [-0.001734, 0.004569]	descriptive	uncertain
Co-culture	Targeting excess (E1)	195 genes	0.002779 [0.001994, 0.003596]	descriptive	positive

Table A17 continued

Condition	Endpoint	Population n	Estimate [interval]	Family	Status
Co-culture	Intergenic cutting / guide-identity comparison (E2)	16 pairs	0.005048 [0.002484, 0.007830]	descriptive	positive
Co-culture	Targeting - intergenic comparison (E3)	195 genes; 16 pairs	-0.002268 [-0.005158, 0.000431]	descriptive	uncertain
Co-culture	Non-targeting comparison (E4)	18 pairs	0.001031 [-0.000203, 0.002303]	descriptive	uncertain

Targeting and control populations are both retained for contrasts. Family and tail assignments are specified in the caption. E3 stores targeting-gene n and disjoint intergenic-control-pair n separately; neither population is replaced by the other.

H.2 Modality-matched intergenic control comparisons. The head-to-head knockout/interference dataset of Drepanos et al. Drepanos et al. (2026) uses the same ten intergenic guides in both libraries. K562 Day 14 is primary because separate encapsulation permits their modality assignment. A549 Day 7 uses pooled, modality-ambiguous controls and supplies targeting-excess secondary endpoints only; no modality-specific targeting-minus-intergenic contrast is available there. The population follows deposited dominant-guide assignments after the authors' 100-cell guide filter. The computed purity statistic was not applied and cannot be reconstructed from the deposit; the population is not described as purity-filtered or exact MOI one.

Targeting guides pass both Hamming distance at least three and longest-common-substring length less than ten nucleotides in either orientation. The frozen endpoint identifiers are targeting excess (XCD), intergenic comparison (XID), and targeting-minus-intergenic contrast (XCF). For XID, the KO instrument is an intergenic cutting/guide-identity comparison, whereas the CRISPRi instrument is an intergenic binding/guide-identity comparison. At K562 Day 14, targeting excess (XCD) is 0.00132128 [0.00050599, 0.00265193] for knockout (23 genes) and 0.00589829 [0.00233155, 0.01118781] for CRISPRi (22 genes). The targeting-minus-intergenic contrast (XCF) is 0.00081598 [-0.00021210, 0.00240814] for KO and 0.00535722 [0.00152799, 0.01145319] for CRISPRi. The KO contrast is unresolved; the CRISPRi contrast exceeds its measured intergenic binding/guide-identity comparison. The latter does not identify guide-sequence causality or a gene-specific efficacy mechanism.

Ten control guides supply 45 overlapping pairs per modality. Uncertainty for the intergenic comparison draws guides with replacement and averages valid distinct-guide pairs within each bootstrap draw; it does not resample 45 independent pairs. The targeting panel is small, and the CRISPRi targeting-minus-intergenic claim has no registered leave-top-gene interval: the removal sensitivities were specified for targeting excess (XCD) alone (Table A21). The largest CRISPRi targeting gene contributes about 34% of the sum and the top three about 57%. Some sequence-removal subsets fall below the inferential gene-count rule and remain descriptive. The CRISPRi ZIM3-KRAB-dCas9 fusion shares repressor lineage with the core panels. The study is a preprint. Its Day-14 modality arms use different response-coordinate memberships; their raw XCD difference is not interpreted as a biological modality effect.

Complete endpoints for this reanalysis of Drepanos et al. (2026) are in Table A18.

Table A18: Complete Drepanos guide-realization endpoints. XCD is targeting excess; XID is the intergenic comparison, with cutting/guide-identity for KO and binding/guide-identity for CRISPRi; XCF is the registered targeting-minus-intergenic contrast. K562 Day-14 XCD endpoints form the two-slot primary family (percentile tails 0.0125/0.9875); both XCF and A549 XCD endpoints form the four-slot secondary family (0.00625/0.99375). XID endpoints are descriptive (0.025/0.975). Intergenic-comparison uncertainty resamples ten guides, not forty-five independent pairs. The A549 intergenic control is pooled and modality-ambiguous; no A549 XCF is registered. Different response-coordinate sets preclude a raw cross-modality effect-size comparison.

Condition	Endpoint	Population n	Estimate [interval]	Family	Status
K562 D14 KO	XCD	23 genes	0.001321 [0.000506, 0.002652]	primary	positive
K562 D14 KO	XID	10 guides	0.000505 [0.000226, 0.000856]	descriptive	positive
K562 D14 KO	XCF	23 genes; 10 guides	0.000816 [-0.000212, 0.002408]	secondary	uncertain
K562 D14 CRISPRi	XCD	22 genes	0.005898 [0.002332, 0.011188]	primary	positive
K562 D14 CRISPRi	XID	10 guides	0.000541 [0.000278, 0.000882]	descriptive	positive
K562 D14 CRISPRi	XCF	22 genes; 10 guides	0.005357 [0.001528, 0.011453]	secondary	positive
A549 D7 KO	XCD	24 genes	0.001590 [0.000577, 0.002881]	secondary	positive
A549 D7 KO	XID	10 guides	0.000395 [0.000235, 0.000576]	descriptive	positive
A549 D7 CRISPRi	XCD	24 genes	0.014703 [0.003895, 0.028724]	secondary	positive
A549 D7 CRISPRi	XID	10 guides	0.000395 [0.000235, 0.000576]	descriptive	positive

Targeting and control populations are both retained for contrasts. Family and tail assignments are specified in the caption. Saved contrast n uses the smaller effective count of targeting genes and control guides. The A549 XID entries repeat the same pooled intergenic comparison, not two independent measurements.

H.3 A different CRISPRi library and repressor fusion. Jiang et al.'s pathway screens Jiang et al. (2025) use Dolcetto-derived guides Sanson et al. (2018) and dCas9-KRAB-MeCP2. The dataset offers 30 cell-line/pathway strata, but this analysis registers four: HT29/TNFA primary, HAP1/IFNB and HT29/IFNB inferential secondary, and MCF7/IFNB descriptive. Only these four support the reported outcomes. The primary was selected from metadata before outcome access. The population follows deposited guide assignments because numeric purity and MOI filters are unavailable. Guides require at least 30 assigned cells and pass the same two-part distinctness rule used for Drepanos; a deterministic first eligible pair is selected per gene.

The HT29/TNFA targeting excess (CGD) is 0.00342891 [0.00229934, 0.00465738] over 44 genes. Its non-targeting construct-identity comparison (NTD) is 0.00108940 [0.00048487, 0.00179586], descriptively; the registered targeting-minus-control contrast (CGF = CGD - NTD) is 0.00233951 [0.00075926, 0.00404028]. Fourteen non-targeting guides supply 91 overlapping pairs; uncertainty resamples the fourteen guides, not the pairs. A non-targeting guide does not recruit the effector to a genomic locus, so this comparison does not bound nonspecific KRAB-MeCP2 recruitment, chromatin-context effects or efficacy mechanisms. The positive contrast is specific to this measured instrument.

Both registered secondary targeting means are positive, but no targeting-minus-control contrast is registered outside the primary stratum. MCF7/IFNB and all non-targeting construct-identity (NTD) estimates remain descriptive. Every stratum uses its own response-coordinate set, so larger raw values in another stratum are not stronger evidence for a biological effect. Separate infections and pathway experiments supply context breadth, not interchangeable replications of one estimand.

In the primary, the registered leave-top-three targeting-minus-control (CGF) sensitivity remains positive at 0.00160916 [0.00022720, 0.00297285]; this robustness was specified before outcomes

(Table A22). The largest gene supplies 9.98% and the top three 26.7% of the targeting sum. The aggregate is not driven by a few genes, but the 97.7% positive noisy gene estimates are not an affected-gene prevalence. The targeting excess (CGD) is approximately 6.5% of its within-guide resampling term. All 654 targeting spacers match Dolcetto; nine coincide with ST20 and 104 with hCRISPRi-v2. None of the selected primary pairs contains an ST20-coincident spacer, so that registered exclusion removes no eligible primary pair (a no-op). Removing hCRISPRi-v2-coincident guides retains both positive primary endpoints. The evidence concerns a different guide-library lineage and repressor fusion, not independence of all methodology or absence of all shared KRAB biology. The fourteen non-targeting spacer sequences are not public and their overlap cannot be checked.

Complete endpoints for this reanalysis of Jiang et al. (2025) are in Table A19.

Table A19: Complete Jiang guide-realization endpoints. CGD is targeting excess, NTD is the non-targeting construct-identity comparison, and CGF is the registered targeting-minus-control contrast. HT29/TNFA CGD is primary (percentile tails 0.025/0.975); its CGF and both inferential secondary-stratum CGD endpoints form a three-slot secondary family (0.0083333/0.9916667). All NTD and MCF7 endpoints are descriptive (0.025/0.975). Uncertainty for the non-targeting comparison resamples fourteen guides, not ninety-one independent pairs. CGF is registered only for HT29/TNFA. Non-targeting controls do not bound nonspecific genomic recruitment of the KRAB-MeCP2 effector.

Condition	Endpoint	Population n	Estimate [interval]	Family	Status
HT29/TNFA	CGD	44 genes	0.003429 [0.002299, 0.004657]	primary	positive
HT29/TNFA	NTD	14 guides	0.001089 [0.000485, 0.001796]	descriptive	positive
HT29/TNFA	CGF	44 genes; 14 guides	0.002340 [0.000759, 0.004040]	secondary	positive
HAP1/IFNB	CGD	51 genes	0.005246 [0.002713, 0.008638]	secondary	positive
HAP1/IFNB	NTD	14 guides	0.001428 [0.000602, 0.002421]	descriptive	positive
HT29/IFNB	CGD	51 genes	0.005537 [0.002861, 0.009021]	secondary	positive
HT29/IFNB	NTD	14 guides	0.001172 [0.000267, 0.002136]	descriptive	positive
MCF7/IFNB	CGD	51 genes	0.032035 [0.023367, 0.042347]	descriptive	positive
MCF7/IFNB	NTD	14 guides	0.025449 [0.016564, 0.035460]	descriptive	positive

Targeting and control populations are both retained for contrasts. Family and tail assignments are specified in the caption. Saved contrast n uses the smaller effective count of targeting genes and control guides.

Registered inference. All three external analyses use 20,000 percentile bootstrap draws with seed 1707 and the frozen implementation's linear-interpolated percentile convention. The Frangieh family contains one primary and three secondary slots; Drepanos contains two primary and four secondary slots; Jiang contains one primary and three secondary slots. Primary/secondary tails follow these family sizes, and descriptive intervals remain unadjusted. The gene-count rule is at least 20 for inference, 10–19 for descriptive reporting and fewer than 10 unavailable; the control-unit rule is at least 10 for inference, 3–9 descriptive and fewer than 3 unavailable. The count rule establishes inferential eligibility, not an upgrade of a descriptive family; contrast status cannot exceed the weaker component. Each workstream was frozen before its own outcome access; these sequential workstreams are not one jointly preregistered experiment.

H.4 Registration, complete outcomes and reproducibility. Endpoint tables report each experiment's primary, secondary and descriptive roles, original family sizes, percentile tails, sample-size rules and unavailable slots. Sensitivity tables distinguish registered analyses from descriptive verification diagnostics, including exclusions that remove no eligible units. Table A23 summarizes the per-gene distributions.

Table A20: Registered sensitivity results by study. Frangieh sensitivities concern targeting excess (E1) only; no targeting-minus-intergenic (E3) sensitivity was registered. Drepanos sensitivities concern targeting excess (XCD) and do not provide influence-robust targeting-minus-intergenic (XCF) intervals. Jiang's primary sensitivities cover targeting excess (CGD) and targeting-minus-control (CGF); other strata have CGD sensitivities only. Rows below the inferential sample-size threshold are descriptive. Exclusions that remove no eligible units are labeled no-op; unavailable saved intervals remain unavailable.

Condition	Registered sensitivity / endpoint	Remaining population	Estimate [saved interval]	Role / saved status / note
IFN-gamma	S1 selected pair $\geq$50 cells; E1	155 genes	0.001883 [0.000947, 0.002871]	primary; positive
IFN-gamma	S2 all pairs; E1	205 genes	0.001921 [0.001081, 0.002814]	primary; positive
IFN-gamma	S3 drop overlap flag; E1	203 genes	0.002002 [0.001014, 0.003070]	primary; positive
IFN-gamma	S4 drop hcrispri coincident; E1	199 genes	0.001981 [0.000978, 0.003051]	primary; positive
IFN-gamma	S5 drop top 1 out; E1	204 genes	0.001771 [0.000883, 0.002701]	primary; positive
IFN-gamma	S5 drop top 2 out; E1	203 genes	0.001624 [0.000762, 0.002498]	primary; positive
IFN-gamma	S5 drop top 3 out; E1	202 genes	0.001526 [0.000684, 0.002391]	primary; positive
Control	S1 selected pair $\geq$50 cells; E1	24 genes	0.004665 [0.001104, 0.008849]	secondary; positive
Control	S2 all pairs; E1	182 genes	0.002855 [0.001439, 0.004379]	secondary; positive
Control	S3 drop overlap flag; E1	179 genes	0.003613 [0.001901, 0.005454]	secondary; positive
Control	S4 drop hcrispri coincident; E1	179 genes	0.003637 [0.001899, 0.005487]	secondary; positive
Control	S5 drop top 1 out; E1	181 genes	0.003341 [0.001724, 0.004980]	secondary; positive
Control	S5 drop top 2 out; E1	180 genes	0.003171 [0.001597, 0.004770]	secondary; positive
Control	S5 drop top 3 out; E1	179 genes	0.003027 [0.001467, 0.004581]	secondary; positive
Co-culture	S1 selected pair $\geq$50 cells; E1	128 genes	0.003040 [0.002179, 0.003907]	descriptive; positive
Co-culture	S2 all pairs; E1	195 genes	0.002925 [0.002213, 0.003663]	descriptive; positive
Co-culture	S3 drop overlap flag; E1	193 genes	0.002772 [0.001978, 0.003585]	descriptive; positive
Co-culture	S4 drop hcrispri coincident; E1	189 genes	0.002778 [0.001974, 0.003600]	descriptive; positive
Co-culture	S5 drop top 1 out; E1	194 genes	0.002701 [0.001909, 0.003490]	descriptive; positive
Co-culture	S5 drop top 2 out; E1	193 genes	0.002628 [0.001862, 0.003404]	descriptive; positive
Co-culture	S5 drop top 3 out; E1	192 genes	0.002556 [0.001789, 0.003342]	descriptive; positive

Table A21: Drepanos registered sensitivities (continued reporting of Table A20). All saved records are retained; unavailable intervals are not recomputed.

Condition	Registered sensitivity / endpoint	Remaining population	Estimate [saved interval]	Role / saved status / note
K562 D14 KO	S1 all pairs; XCD	23 genes	0.001409 interval unavailable	primary; n-eligible; point only
K562 D14 KO	S2 ST20 drop; XCD	23 genes	0.001321 interval unavailable	primary; n-eligible; point only; no-op (0 flagged)
K562 D14 KO	S3 hCRISPRi drop; XCD	23 genes	0.001321 interval unavailable	primary; n-eligible; point only; no-op (0 flagged)
K562 D14 KO	S4 Brunello drop; XCD	11 genes	0.001431 interval unavailable	primary; descriptive ($n < 20$); point only
K562 D14 KO	S5 drop top 1; XCD	22 genes	0.000864 interval unavailable	primary; n-eligible; point only
K562 D14 KO	S5 drop top 2; XCD	21 genes	0.000658 interval unavailable	primary; n-eligible; point only
K562 D14 KO	S5 drop top 3; XCD	20 genes	0.000564 interval unavailable	primary; n-eligible; point only
K562 D14 CRISPRi	S1 all pairs; XCD	22 genes	0.006281 interval unavailable	primary; n-eligible; point only
K562 D14 CRISPRi	S2 ST20 drop; XCD	22 genes	0.005898 interval unavailable	primary; n-eligible; point only; no-op (0 flagged)
K562 D14 CRISPRi	S3 hCRISPRi drop; XCD	16 genes	0.005540 interval unavailable	primary; descriptive ($n < 20$); point only
K562 D14 CRISPRi	S4 Brunello drop; XCD	22 genes	0.005898 interval unavailable	primary; n-eligible; point only; no-op (0 flagged)
K562 D14 CRISPRi	S5 drop top 1; XCD	21 genes	0.004069 interval unavailable	primary; n-eligible; point only
K562 D14 CRISPRi	S5 drop top 2; XCD	20 genes	0.003361 interval unavailable	primary; n-eligible; point only
K562 D14 CRISPRi	S5 drop top 3; XCD	19 genes	0.002940 interval unavailable	primary; descriptive ($n < 20$); point only
A549 D7 KO	S1 all pairs; XCD	24 genes	0.003485 interval unavailable	secondary; n-eligible; point only
A549 D7 KO	S2 ST20 drop; XCD	24 genes	0.001590 interval unavailable	secondary; n-eligible; point only; no-op (0 flagged)
A549 D7 KO	S3 hCRISPRi drop; XCD	24 genes	0.001590 interval unavailable	secondary; n-eligible; point only; no-op (0 flagged)
A549 D7 KO	S4 Brunello drop; XCD	12 genes	0.002066 interval unavailable	secondary; descriptive ($n < 20$); point only
A549 D7 KO	S5 drop top 1; XCD	23 genes	0.001286 interval unavailable	secondary; n-eligible; point only
A549 D7 KO	S5 drop top 2; XCD	22 genes	0.001065 interval unavailable	secondary; n-eligible; point only
A549 D7 KO	S5 drop top 3; XCD	21 genes	0.000848 interval unavailable	secondary; n-eligible; point only
A549 D7 CRISPRi	S1 all pairs; XCD	24 genes	0.017834 interval unavailable	secondary; n-eligible; point only
A549 D7 CRISPRi	S2 ST20 drop; XCD	24 genes	0.014703 interval unavailable	secondary; n-eligible; point only; no-op (0 flagged)
A549 D7 CRISPRi	S3 hCRISPRi drop; XCD	19 genes	0.013785 interval unavailable	secondary; descriptive ($n < 20$); point only

Table A21 continued

Condition	Registered sensitivity / endpoint	Remaining population	Estimate [saved interval]	Role / saved status / note
A549 D7 CRISPRi	S4 Brunello drop; XCD	24 genes	0.014703 interval unavailable	secondary; n-eligible; point only; no-op (0 flagged)
A549 D7 CRISPRi	S5 drop top 1; XCD	23 genes	0.012072 interval unavailable	secondary; n-eligible; point only
A549 D7 CRISPRi	S5 drop top 2; XCD	22 genes	0.009493 interval unavailable	secondary; n-eligible; point only
A549 D7 CRISPRi	S5 drop top 3; XCD	21 genes	0.006704 interval unavailable	secondary; n-eligible; point only

Table A22: Jiang registered sensitivities (continued reporting of Table A20). CGD denotes targeting excess and CGF the targeting-minus-control contrast. All saved records are retained; unavailable intervals are not recomputed.

Condition	Registered sensitivity / endpoint	Remaining population	Estimate [saved interval]	Role / saved status / note
HT29/TNFA	S1 all pairs; CGD	44 genes	0.003289 [0.002284, 0.004411]	primary; positive
HT29/TNFA	S1 all pairs; CGF	44 genes; 14 guides	0.002199 [0.000727, 0.003794]	secondary; positive
HT29/TNFA	S2 ST20 drop; CGD	44 genes	0.003429 [0.002299, 0.004657]	primary; positive; no-op (0 flagged)
HT29/TNFA	S2 ST20 drop; CGF	44 genes; 14 guides	0.002340 [0.000759, 0.004040]	secondary; positive; no-op (0 flagged)
HT29/TNFA	S3 hCRISPRi drop; CGD	35 genes	0.003897 [0.002562, 0.005359]	primary; positive
HT29/TNFA	S3 hCRISPRi drop; CGF	35 genes; 14 guides	0.002808 [0.000981, 0.004769]	secondary; positive
HT29/TNFA	S4 drop top 1; CGD	43 genes	0.003159 [0.002127, 0.004284]	primary; positive
HT29/TNFA	S4 drop top 1; CGF	43 genes; 14 guides	0.002069 [0.000570, 0.003646]	secondary; positive
HT29/TNFA	S4 drop top 2; CGD	42 genes	0.002881 [0.001964, 0.003863]	primary; positive
HT29/TNFA	S4 drop top 2; CGF	42 genes; 14 guides	0.001792 [0.000398, 0.003240]	secondary; positive
HT29/TNFA	S4 drop top 3; CGD	41 genes	0.002699 [0.001840, 0.003627]	primary; positive
HT29/TNFA	S4 drop top 3; CGF	41 genes; 14 guides	0.001609 [0.000227, 0.002973]	secondary; positive
HAP1/IFNB	S1 all pairs; CGD	51 genes	0.005268 [0.002735, 0.008667]	secondary; positive
HAP1/IFNB	S2 ST20 drop; CGD	51 genes	0.005246 [0.002713, 0.008638]	secondary; positive; no-op (0 flagged)
HAP1/IFNB	S3 hCRISPRi drop; CGD	34 genes	0.006586 [0.002982, 0.011310]	secondary; positive
HAP1/IFNB	S4 drop top 1; CGD	50 genes	0.004353 [0.002456, 0.006705]	secondary; positive
HAP1/IFNB	S4 drop top 2; CGD	49 genes	0.003854 [0.002256, 0.005856]	secondary; positive
HAP1/IFNB	S4 drop top 3; CGD	48 genes	0.003375 [0.002086, 0.004927]	secondary; positive
HT29/IFNB	S1 all pairs; CGD	51 genes	0.005796 [0.003094, 0.009347]	secondary; positive

Table A22 continued

Condition	Registered sensitivity / endpoint	Remaining population	Estimate [saved interval]	Role / saved status / note
HT29/IFNB	S2 ST20 drop; CGD	51 genes	0.005537 [0.002861, 0.009021]	secondary; positive; no-op (0 flagged)
HT29/IFNB	S3 hCRISPRi drop; CGD	35 genes	0.006320 [0.002917, 0.011006]	secondary; positive
HT29/IFNB	S4 drop top 1; CGD	50 genes	0.004607 [0.002550, 0.007053]	secondary; positive
HT29/IFNB	S4 drop top 2; CGD	49 genes	0.004092 [0.002303, 0.006093]	secondary; positive
HT29/IFNB	S4 drop top 3; CGD	48 genes	0.003697 [0.002137, 0.005444]	secondary; positive
MCF7/IFNB	S1 all pairs; CGD	51 genes	0.031647 [0.023368, 0.041557]	descriptive; positive
MCF7/IFNB	S2 ST20 drop; CGD	51 genes	0.032035 [0.023367, 0.042347]	descriptive; positive; no-op (0 flagged)
MCF7/IFNB	S3 hCRISPRi drop; CGD	35 genes	0.028587 [0.021703, 0.036541]	descriptive; positive
MCF7/IFNB	S4 drop top 1; CGD	50 genes	0.029170 [0.021884, 0.038108]	descriptive; positive
MCF7/IFNB	S4 drop top 2; CGD	49 genes	0.026268 [0.020694, 0.032533]	descriptive; positive
MCF7/IFNB	S4 drop top 3; CGD	48 genes	0.024754 [0.019914, 0.030203]	descriptive; positive

Table A23: Per-gene distributions qualify the aggregate target-disagreement results. Means, medians, signs and contribution shares refer to noisy per-gene estimates on each study's own response space. Positive-estimate fractions are not affected-gene prevalence. Frangieh's primary targeting excess survives its registered influential-gene exclusions; Drepanos has a concentrated upper tail and no registered targeting-minus-intergenic (XCF) influence interval; Jiang's primary targeting-minus-control (CGF) contrast remains positive under the registered leave-top-three analysis. Descriptive quantities first supplied by verification are labeled separately from registered sensitivities. No cross-dataset magnitude ranking is intended.

System / condition	n	Mean / median	Positive fraction	Top gene / share	Top-three share	Registered influence
Frangieh IFN-gamma	205	0.002001 0.000830	0.556	PUF60 0.119	0.248	Drop 3 targeting: 0.001526 [0.000684, 0.002391]; no E3 sensitivity
Frangieh Control	182	0.003645 0.002470	0.626	PPIA 0.088	0.183	Drop 3 targeting: 0.003027 [0.001467, 0.004581]; no E3 sensitivity
Frangieh Co-culture	195	0.002779 0.002692	0.646	HNRNPA1 0.033	0.095	Drop 3 targeting: 0.002556 [0.001789, 0.003342]; no E3 sensitivity
Drepanos K562 D14 KO	23	0.001321 0.000532	1.000	NDUFV2 0.375	0.629	Drop 3 targeting: 0.000564 interval unavailable; no XCF interval
Drepanos K562 D14 CRISPRi	22	0.005898 0.001423	0.955	STK24 0.342	0.570	Drop 3 targeting: 0.002940 interval unavailable; no XCF interval
Drepanos A549 D7 KO	24	0.001590 0.000706	0.833	GNG5 0.225	0.534	Drop 3 targeting: 0.000848 interval unavailable; no XCF interval
Drepanos A549 D7 CRISPRi	24	0.014703 0.001857	0.833	IFT25 0.213	0.601	Drop 3 targeting: 0.006704 interval unavailable; no XCF interval

Table A23 continued

System / condition	n	Mean / median	Positive fraction	Top gene / share	Top-three share	Registered influence
Jiang HT29/TNFA	44	0.003429 0.001581	0.977	CASP10 0.100	0.267	Drop 3 CGF: 0.001609 [0.000227, 0.002973]
Jiang HAP1/IFNB	51	0.005246 0.002481	0.922	IRF9 0.187	0.394	Drop 3 targeting: 0.003375 [0.002086, 0.004927]
Jiang HT29/IFNB	51	0.005537 0.002241	0.863	IRF9 0.184	0.371	Drop 3 targeting: 0.003697 [0.002137, 0.005444]
Jiang MCF7/IFNB	51	0.032035 0.022318	1.000	IFNAR1 0.107	0.273	Drop 3 targeting: 0.024754 [0.019914, 0.030203]

Shares and positive fractions are displayed as proportions.

I RELEASED STATE PREDICTION SENSITIVITY

Released State predictions (Phase A). This analysis scores only the released State ST-HVG zero-shot `eval_last` predictions Adduri et al. (2026) without retraining, checkpoint selection or global construct exclusion. Neither eval_best nor ST-SE is used. The released evaluation population determines gene availability; its efficacy filtering is not undone. Both selected constructs may be represented in the model's source information. Construct A/B are symmetric labels and do not encode whether a construct is present in source data or withheld from all source data or strict-transfer roles.

Cohort and frozen targets. For each gene and held-out panel, require a released gene prediction and at least two complete cassettes with at least 20 raw held-out cells each; select the lexicographically first unordered eligible pair. The cohorts contain 14 K562, 36 RPE1, 31 Jurkat and 31 HepG2 genes (Table A24). The predeclared rule is inferential for at least 20 genes, descriptive for 10–19 and unavailable below ten; it makes K562 descriptive and the other three inferential. There are 112 gene-folds over 52 distinct genes, not 112 independent perturbations. Neither GenePT feature eligibility nor the strict-transfer source-coverage rule defines this cohort.

Table A24: Frozen cohorts for released State predictions (Phase A) and higher-cell-count subsets. Selection uses released prediction availability, complete construct identifiers and raw held-out cell counts, not construct-specific scores or source-status labels. The registered sample-size rule makes K562 descriptive in both scopes and the eighteen-gene HepG2 higher-cell-count subset descriptive. No replacement genes or efficacy-filter reversal is performed. Equal target cell counts do not imply equal library sizes.

Fold	Original n / role	$\geq$30-cell n / role	Half-depth min / median / max
K562	14; descriptive	13; descriptive	13 / 29.5 / 90
RPE1	36; inferential	35; inferential	12 / 42.5 / 130
Jurkat	31; inferential	29; inferential	12 / 36.0 / 81
HepG2	31; inferential	18; descriptive	10 / 17.0 / 49

Response axis: all 2,000 released coordinates. Prediction source: pinned ST-HVG zero-shot eval_last release described above.

For each of 20 frozen partitions (seeds 30001–30020), form disjoint within-construct halves A1/A2 and B1/B2, all at $m = \lfloor \min(n_A, n_B)/2 \rfloor$ cells. Sum counts before CP10K normalization on the core 6,642-gene axis, use the pooled held-out non-targeting control and pseudocount 0.05 to form $\log_2$ fold changes, then restrict to the 2,000 released output coordinates. All released coordinates map without dropping genes. Matching cell counts does not match total UMI counts.

One prediction per gene. The released predicted cells are normalized in log1p CP10K units over 8,749 raw genes. Invert log1p first, average every predicted row carrying the gene label, and form a $\log_2$ ratio with pseudocount 0.05 against the identically aggregated predicted non-targeting reference. The same gene-conditioned prediction is scored against all four realized targets. Predicted rows are

never sliced by the real cells' construct identity. The release and target normalization axes differ; Pearson is primary, while MAE is secondary and retains the residual scale/pseudocount caveat.

Score-disagreement estimator. For score function S and the frozen prediction $\hat{y}$, define $a_1 = S(\hat{y}, A_1)$, $a_2 = S(\hat{y}, A_2)$, $b_1 = S(\hat{y}, B_1)$ and $b_2 = S(\hat{y}, B_2)$. The partition statistic is $D^2_{\text{State}} = (a_1 - b_1)^2 - \{(a_1 - a_2)^2 + (b_1 - b_2)^2\}/2$. Average this statistic over partitions within gene, then equally over genes. It is excess squared score disagreement, not profile MSE, a signed accuracy penalty, an SB coefficient or the strict-transfer contrast D. No materiality threshold is attached to this analysis.

Inference and complete fold reporting. Paired-gene percentile bootstrap intervals use 20,000 draws and seed 1707. The three inferential State Pearson slots retain their primary family (tails 0.0083333/0.9916667). The six secondary slots comprise State MAE and the State-minus-context-comparator Pearson contrast in the three inferential folds (tails 0.0041667/0.9958333). K562, other comparator summaries and descriptive slots retain their registered treatment; family assignments are never inferred from a positive direction label alone. The primary State Pearson means are 0.012007 [0.002994, 0.024259] in RPE1, 0.006277 [0.000362, 0.016792] in Jurkat and 0.016423 [0.003985, 0.035947] in HepG2. Descriptive K562 is 0.000965 [-0.000201, 0.002814]. The three inferential means resolve positive excess, not three independent biological replications or a prevalence estimate (Table A25).

Table A25: Released State predictions (Phase A) have target-realization-dependent scores. The endpoint is excess squared score disagreement across two complete constructs after subtracting the average same-construct resampling term, averaged over twenty partitions within gene and equally over genes. It is not an ordinary Pearson-score change or a signed transfer penalty. RPE1, Jurkat and HepG2 Pearson endpoints form the three-slot primary family; State MAE belongs to the separate secondary family. K562 is descriptive. The aggregate is concentrated in a minority of genes; Jurkat loses resolution after removing its two most influential genes in a non-preregistered diagnostic. No global construct-exclusion contract is imposed on State.

Fold / scope	Quantity	n	Estimate [interval]	Family / cohort role	Direction
K562	State pearson	14	0.000965 [-0.000201, 0.002814]	descriptive; descriptive	uncertain
K562	State mae	14	0.001444 [0.000240, 0.003306]	descriptive; descriptive	positive
RPE1	State pearson	36	0.012007 [0.002994, 0.024259]	primary; inferential	positive
RPE1	State mae	36	0.006262 [0.000475, 0.017126]	secondary; inferential	positive
Jurkat	State pearson	31	0.006277 [0.000362, 0.016792]	primary; inferential	positive
Jurkat	State mae	31	0.000500 [0.000091, 0.001074]	secondary; inferential	positive
HepG2	State pearson	31	0.016423 [0.003985, 0.035947]	primary; inferential	positive
HepG2	State mae	31	0.004670 [0.000589, 0.011706]	secondary; inferential	positive

Distribution and influence. The fold means have concentrated magnitude (Table A26). The largest gene contributes 27.7% in RPE1, 49.7% in Jurkat, 39.6% in HepG2 and 86.2% in K562; medians are much smaller than means. Positive signs of noisy gene estimates do not establish how many genes truly shift. A non-preregistered descriptive influence diagnostic retains positive intervals in RPE1 and HepG2 after removing their top three contributors, but Jurkat loses resolution after removing its top two (Table A27). This diagnostic does not redefine the registered primary cohort or replace its result. Report its status and intervals separately from registered sensitivities.

Table A26: Concentration of State's Pearson-squared excess and non-preregistered influence diagnostics. The original registered endpoints are unchanged. Contribution shares and positive-estimate counts describe the observed noisy per-gene distribution, not prevalence. RPE1 and HepG2 retain positive intervals after removing their three largest contributors in the supplied diagnostic; Jurkat does not after removing its two largest. K562 remains descriptive. Reported diagnostic intervals reuse the stated tail convention but do not create a new confirmatory family. An unreported diagnostic interval remains unreported.

Fold	n	Positive estimates	Median / mean	Top gene (estimate)	Share
K562	14	6 (43%)	-0.000253 / 0.000965	PMPCB (0.0117)	86.2%
RPE1	36	20 (56%)	+0.001575 / 0.012007	ADAM10 (0.1195)	27.7%
Jurkat	31	19 (61%)	+0.000200 / 0.006277	ESF1 (0.0967)	49.7%
HepG2	31	19 (61%)	+0.003245 / 0.016423	LSM5 (0.2019)	39.6%

Table A27: Non-preregistered descriptive influence diagnostics accompanying Table A26. Jurkat loses resolution after dropping its top two contributors; K562 remains descriptive.

Fold	Drop top 1	Drop top 2	Drop top 3
K562	0.000143 [-0.000294, 0.000697]	-0.000075 interval not reported	-0.000198 interval not reported
RPE1	0.008935 [0.002059, 0.018029]	0.006714 [0.001319, 0.013965]	0.004896 [0.000725, 0.010211]
Jurkat	0.003261 [0.000137, 0.009722]	0.001138 [-0.000042, 0.002667]	0.000731 [-0.000173, 0.001816]
HepG2	0.010242 [0.002954, 0.019623]	0.008142 [0.002009, 0.015936]	0.006211 [0.001320, 0.012456]

Registered sensitivity and comparator scope. Raising the per-construct cell-count requirement to 30 cells retains 35 RPE1 and 29 Jurkat genes as inferential subsets, while the 18-gene HepG2 and 13-gene K562 subsets are descriptive. The source-context response comparator averages available same-gene source responses; 36 of 112 gene-folds have two source panels rather than three. It is not the globally excluded Perturb Mean of §5. In HepG2, State has less Pearson score disagreement than the source-context response comparator (Table A28). The comparator uses available source responses and differs from the globally excluded Perturb Mean in §5.

Table A28: Complete released State prediction (Phase A) outcomes and registered higher-cell-count sensitivity results. All model, metric, contrast and cohort-scope records are retained, including the source-context response comparator and State-minus-comparator contrasts. The three inferential State Pearson slots use primary tails 0.0083333/0.9916667; the six secondary slots use 0.0041667/0.9958333; descriptive endpoints use 0.025/0.975. Sensitivities retain their registered status, including descriptive K562 and the eighteen-gene HepG2 subset. MAE carries the release-to-target normalization and pseudocount-scale qualification. The context comparator uses two rather than three source panels for thirty-six gene-folds and is not the strict-transfer Perturb Mean. No core-paper materiality margin is applied.

Fold / scope	Quantity	n	Estimate [interval]	Family / cohort role	Direction
K562; original	State pearson	14	0.000965 [-0.000201, 0.002814]	descriptive; descriptive	uncertain
K562; original	State mae	14	0.001444 [0.000240, 0.003306]	descriptive; descriptive	positive
K562; original	Context pearson	14	0.032650 [0.007172, 0.067891]	descriptive; descriptive	positive
K562; original	Context mae	14	0.001059 [0.000061, 0.002733]	descriptive; descriptive	positive
K562; original	State - context pearson	14	-0.031685 [-0.066244, -0.007222]	descriptive; descriptive	resolved negative

Table A28 continued

Fold / scope	Quantity	n	Estimate [interval]	Family / cohort role	Direction
K562; original	State - context mae	14	0.000386 [0.000086, 0.000718]	descriptive; descriptive	positive
K562; ≥30 cells	State pearson	13	0.001018 [-0.000236, 0.003012]	sensitivity; descriptive	uncertain
K562; ≥30 cells	State mae	13	0.001571 [0.000296, 0.003595]	sensitivity; descriptive	positive
K562; ≥30 cells	Context pearson	13	0.035195 [0.008183, 0.073228]	descriptive; descriptive	positive
K562; ≥30 cells	Context mae	13	0.001161 [0.000103, 0.002969]	descriptive; descriptive	positive
K562; ≥30 cells	State - context pearson	13	-0.034177 [-0.071855, -0.008274]	sensitivity; descriptive	resolved negative
K562; ≥30 cells	State - context mae	13	0.000410 [0.000092, 0.000764]	sensitivity; descriptive	positive
RPE1; original	State pearson	36	0.012007 [0.002994, 0.024259]	primary; inferential	positive
RPE1; original	State mae	36	0.006262 [0.000475, 0.017126]	secondary; inferential	positive
RPE1; original	Context pearson	36	0.013301 [0.006197, 0.022840]	descriptive; inferential	positive
RPE1; original	Context mae	36	0.006714 [0.001213, 0.014079]	descriptive; inferential	positive
RPE1; original	State - context pearson	36	-0.001294 [-0.007252, 0.005250]	secondary; inferential	uncertain
RPE1; original	State - context mae	36	-0.000452 [-0.001919, 0.000670]	descriptive; inferential	uncertain
RPE1; ≥30 cells	State pearson	35	0.012285 [0.003006, 0.024651]	sensitivity; inferential	positive
RPE1; ≥30 cells	State mae	35	0.003589 [0.000295, 0.010451]	sensitivity; inferential	positive
RPE1; ≥30 cells	Context pearson	35	0.013515 [0.006083, 0.023369]	descriptive; inferential	positive
RPE1; ≥30 cells	Context mae	35	0.004238 [0.000731, 0.009643]	descriptive; inferential	positive
RPE1; ≥30 cells	State - context pearson	35	-0.001230 [-0.007550, 0.005609]	sensitivity; inferential	uncertain
RPE1; ≥30 cells	State - context mae	35	-0.000649 [-0.002073, 0.000415]	sensitivity; inferential	uncertain
Jurkat; original	State pearson	31	0.006277 [0.000362, 0.016792]	primary; inferential	positive
Jurkat; original	State mae	31	0.000500 [0.000091, 0.001074]	secondary; inferential	positive
Jurkat; original	Context pearson	31	0.012235 [0.004262, 0.022758]	descriptive; inferential	positive
Jurkat; original	Context mae	31	0.000547 [0.000188, 0.000971]	descriptive; inferential	positive
Jurkat; original	State - context pearson	31	-0.005958 [-0.019069, 0.002251]	secondary; inferential	uncertain
Jurkat; original	State - context mae	31	-0.000047 [-0.000301, 0.000173]	descriptive; inferential	uncertain
Jurkat; ≥30 cells	State pearson	29	0.006701 [0.000387, 0.018056]	sensitivity; inferential	positive
Jurkat; ≥30 cells	State mae	29	0.000539 [0.000113, 0.001153]	sensitivity; inferential	positive
Jurkat; ≥30 cells	Context pearson	29	0.012172 [0.003734, 0.023500]	descriptive; inferential	positive
Jurkat; ≥30 cells	Context mae	29	0.000566 [0.000192, 0.001011]	descriptive; inferential	positive

Table A28 continued

Fold / scope	Quantity	n	Estimate [interval]	Family / cohort role	Direction
Jurkat; ≥30 cells	State - context pearson	29	-0.005472 [-0.019681, 0.003170]	sensitivity; inferential	uncertain
Jurkat; ≥30 cells	State - context mae	29	-0.000028 [-0.000294, 0.000198]	sensitivity; inferential	uncertain
HepG2; original	State pearson	31	0.016423 [0.003985, 0.035947]	primary; inferential	positive
HepG2; original	State mae	31	0.004670 [0.000589, 0.011706]	secondary; inferential	positive
HepG2; original	Context pearson	31	0.045385 [0.019628, 0.077020]	descriptive; inferential	positive
HepG2; original	Context mae	31	0.003828 [0.001097, 0.007550]	descriptive; inferential	positive
HepG2; original	State - context pearson	31	-0.028962 [-0.054829, -0.009140]	secondary; inferential	resolved negative
HepG2; original	State - context mae	31	0.000842 [-0.000128, 0.002110]	descriptive; inferential	uncertain
HepG2; ≥30 cells	State pearson	18	0.019596 [0.001696, 0.045290]	sensitivity; descriptive	positive
HepG2; ≥30 cells	State mae	18	0.001827 [-0.000071, 0.005189]	sensitivity; descriptive	uncertain
HepG2; ≥30 cells	Context pearson	18	0.044411 [0.004464, 0.096105]	descriptive; descriptive	positive
HepG2; ≥30 cells	Context mae	18	0.001379 [0.000014, 0.003812]	descriptive; descriptive	positive
HepG2; ≥30 cells	State - context pearson	18	-0.024815 [-0.054071, -0.001846]	sensitivity; descriptive	resolved negative
HepG2; ≥30 cells	State - context mae	18	0.000448 [-0.000197, 0.001451]	sensitivity; descriptive	uncertain

Original-cohort State rows duplicate Table A25; CSV components are auxiliary quantities of the same D^2_{State} analysis.

Limits of the contrast. This released-prediction analysis measures target-dependent score disagreement with predictions held fixed; global construct exclusion and fitted-model comparisons are evaluated separately in §5 and Appendix J. The primary estimator is invariant to exchanging A/B labels and measures the fold mean under the specified sampled targets.

J STATE ARCHITECTURE UNDER GLOBAL CONSTRUCT EXCLUSION

Phase A scores released predictions against alternative targets (Appendix I); Phase B instead trains State under global construct exclusion with a separate inferential family.

State architecture trained from scratch (Phase B). This experiment uses the released ST-HVG architecture without released weights, with one seed (42) and one model per held-out context. Each model is trained for 100,000 optimizer steps. The held-out context and its controls are absent from training and prediction. The complete withheld construct is excluded globally before normalization, feature selection, vocabulary construction and caches; each evaluation gene remains represented through its selected construct. The four folds reuse genes and constructs.

The source-only bridge uses source-context basal cells and a predicted non-targeting reference to form a log-fold-change vector; source contexts are averaged equally. This is separate from native held-out-context zero-shot inference. All four prediction freezes preceded held-out target access. The full-output State decoder and evaluation use the same 6,642-gene response axis.

State Pearson penalties exceed the predeclared +0.02 materiality margin in RPE1 and Jurkat. K562 and HepG2 remain unresolved because their adjusted intervals include zero.

Table A29: State target-realization sensitivity under global construct exclusion. State was trained from scratch under global construct exclusion. Represented and withheld columns report mean Pearson scores on the full-depth targets; the penalty is represented minus withheld, so positive values indicate lower scores for constructs withheld from all source data. RPE1 and Jurkat exceed the predeclared $+0.02$ materiality margin; K562 and HepG2 intervals include zero. Adjusted intervals use 20,000 paired-gene bootstrap draws (seed 1707; tail 0.05/48).

Context	n	Represented Pearson	Withheld Pearson	State Pearson penalty	Adjusted interval
K562	57	0.075690	0.011812	0.063878	[-0.006339, 0.139961]
RPE1	60	0.176886	0.047065	0.129821	[0.035025, 0.225598]
Jurkat	49	0.106359	-0.000558	0.106916	[0.047568, 0.178545]
HepG2	31	0.115975	0.010057	0.105918	[-0.022549, 0.251521]

REPRODUCIBILITY AND PROVENANCE.

Cross-appendix records for Appendices E, H, I and J.

Model implementation and RNG. The adapted fMLP core uses commit `15361aec8392fb45a75c8c9da88ca5e6510fca97`: `https://github.com/shiftbioscience/Perturbation-Models-Outperform-Baselines/blob/15361aec8392fb45a75c8c9da88ca5e6510fca97/docker/fmlp/fmlp_wrapper.py`. Three float32 prediction vectors are averaged with float64 accumulation before scoring; training uses an A100 GPU, deterministic algorithms and disabled mixed precision/TF32. Partition seeds are 19001–19020 in Appendix E, keyed by fold and gene, and 40001–40020 in Appendix H. Replay preserves each external design's frozen RNG ordering and guide/pair resampling algorithm.

Released State artifacts (Phase A). The source is `arcinstitute/ST-HVG-Replogle`, revision `bb6a9562cbbf1fd152df14cc53b4cc7517c77175`, path `zeroshot/{fold}/eval_last.ckpt/adata_pred.h5ad`; only eval_last is used, not eval_best or ST-SE. Half-depth receipt medians average the two central values for even-sized samples; the preregistration uses the upper median, a documentation convention rather than different cohorts. Not every per-partition score or raw prediction transformation is persisted for independent replay.

External-screen sources and verification. Frangieh: `FrangiehIzar2021_RNA.h5ad`, scPerturb 1.4 Peidli et al. (2024), Zenodo 13350497, `PREREG_FRANGIEH_V1`. Drepanos: Zenodo 20722064, `K562_Day14.Rds` and `A549_Day7.Rds`, `PREREG_DREPANOS_V1`. Jiang: Zenodo 14518762, GEO GSE281048, `PREREG_JIANG_V1`. The source manifest retains full source identities and binds frozen designs, coordinate lists and saved outputs. Frangieh's persisted pseudobulks support reconstruction of profile scores; Jiang persists per-partition components; Drepanos persists per-pair component aggregates. Distinctness and shared-spacer receipts document frozen eligibility, not additional endpoint tests. Drepanos top-three shares are descriptive verification quantities transcribed from the final verification table; other supplied fields come from persisted concentration records.