FINITE-MEASUREMENT TRANSPORT OF PERTURBATION RELIABILITY RANKINGS

Anonymous authors
Paper under double-blind review

ABSTRACT

Predictions validated in one biological context are often considered for use in another, yet perturbations with low prediction error in one context need not retain their relative accuracy after a context change. We study this ordering of perturbation-level realized error while the predictor is held fixed. Because finite cell counts can themselves induce apparent rank changes, we define a tie-aware discrepancy that compares pairwise ordering distributions and subtracts expected within-context ordering disagreement. The resulting finite-measurement estimand is zero when the two pair-state laws agree, while its unbiased estimator may be negative at finite sample size. In the Frangieh Perturb-CITE-seq dataset (Frangieh et al., 2021), reliability rankings change across control, co-culture and IFNγ conditions under three complementary risk metrics. The changes exceed same-context sampling baselines, become increasingly resolvable with cell budget, and are reproduced in independent HepG2–Jurkat and iPSC–neuron raw-cell analyses (Nadig et al., 2025; Tian et al., 2019). Matched source-only RBF, direct-MLP, latent-MLP and strict source-frozen GEARS sensitivities preserve positive transport signs on their respective eligible Frangieh surfaces, although their within-metric profiles are not identical. Reliability reordering also changes finite top-k experimental shortlists. In the tested controls, however, D_{adj} did not improve held-out regret prediction beyond metric identity and mean-risk shift. A known-truth comparator audit shows why the finite-measurement correction matters: uncorrected rank statistics can declare transport under a same-context null, whereas matched-null corrections restore calibration. The framework therefore supports measurement-aware auditing of cross-context changes in model-error ordering.

1 INTRODUCTION

Single-cell perturbation screens make it possible to measure many interventions in parallel, but exhaustively measuring them across biological contexts is often impractical. Pooled Perturb-seq established a route from genetic perturbations to high-dimensional single-cell readouts (Dixit et al., 2016; Adamson et al., 2016); genome-scale screens and harmonized perturbation resources have since expanded both the intervention universe and the range of measurement depths (Replogle et al., 2022; Peidli et al., 2024). A computational predictor can therefore help determine which source-validated predictions may be reused after a context change.

That narrower reuse decision is an ordering problem. Existing perturbation-prediction studies usually ask whether a model reconstructs an unseen response or generalizes to an unseen perturbation. Representative approaches span latent-variable, optimal-transport, graph-based and foundation-model methods, including scGen, CPA, CellOT, GEARS, scGPT and TxPert (Lotfollahi et al., 2019; 2023; Bunne et al., 2023; Roohani et al., 2024; Cui et al., 2024; Wenkel et al., 2026). Benchmarks broaden this comparison across datasets and metrics (Wei et al., 2026), while direct audits show that complex models need not outperform simple linear baselines under a fixed protocol (Ahlmann-Eltze et al., 2025; Wong et al., 2025). These studies evaluate response fidelity; they do not directly ask whether the identity of reliably predicted perturbations survives a biological context shift.

A ranking shift is not automatically a biological-context effect. With finite numbers of cells, perturbation risks fluctuate, and two observed orders can differ even when the underlying order is not resolved. The number of cells per perturbation–context pseudoreplicate is therefore part of the in-

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.

ferential problem, not merely technical details (Peidli et al., 2024; Squair et al., 2021). Work on ranking under noisy measurements makes the same point for uncertain and tied orders (Zuk et al., 2007; Sankaran et al., 2025). Our question is complementary to calibration and selective prediction: after accounting for this finite-cell ordering floor, does the relative ordering of realized model risk change between biological contexts?

We answer this question with a source-frozen protocol. Predictions are fixed before target outcomes are observed; context-specific risks are then compared through a tie-aware, measurement-adjusted discrepancy. For each perturbation pair, the estimator compares the distributions of lower, tied and higher risk states across contexts and subtracts expected within-context ordering disagreement using an unbiased order-2 U-statistic (Hoeffding, 1948). The primary analysis uses the Frangieh context surface, same-context pseudoreplicate controls, independent HepG2–Jurkat and iPSC–neuron raw-cell replications, matched-cell-budget resampling and a finite top-k decision layer.

We make three contributions. First, we formalize reliability transport as a finite-measurement comparison of perturbation-pair ordering distributions rather than mean response fidelity. Second, we show that context-associated reordering exceeds a matched sampling floor, has an explicit measurement boundary, and recurs across independent biological systems and matched nonlinear-predictor sensitivities. Third, we quantify its consequence for finite experimental shortlists while keeping post-hoc transport measurement separate from prospective source-only prediction.

2 Problem setting and measurement-adjusted ordering

Let $X_{p,e}^{(r)}$ denote the raw cells sampled for perturbation p in context e and Monte Carlo replicate r. Each cell is library-size normalized to 10^4, transformed by $\log(1+x)$, and aggregated into the declared gene-panel response $y_{p,e}^{(r)}$ (a perturbation-versus-control delta for delta-based metrics). The primary predictor is an out-of-fold map: for perturbation p, $f_{s,-\mathrm{fold}(p)}$ is fit only to labels outside p's fixed fold in the source training context s, using a ten-component PCA representation, ridge penalty $\alpha = 0.1$, and three independently seeded members. It produces a fixed held-out vector $\hat{y}_{p,s}$ before any target outcome is examined. For metric m, the realized risk is

$$R_{p,e}^{(m,r)} = g_m(\hat{y}_{p,s}, y_{p,e}^{(r)}), \tag{1}$$

where larger values mean worse prediction and the direction $s \to e$ names the training context and evaluation context, respectively. We use one minus cosine similarity (delta cosine), one minus Systema centroid accuracy (Vinas Torne et al., 2026), and one minus absolute-effect rank agreement; complete metric definitions, comparator universes, and tie tolerances are given in Appendix B. These risks are evaluated after target outcomes are observed; source-only uncertainty is kept separate from the ordering estimand.

For a pair $a = (p, q)$, define the tie-aware pair state

$$Z_{a,e}^{(m,r)} = \mathrm{sign}\Big(R_{p,e}^{(m,r)} - R_{q,e}^{(m,r)}\Big) \in \{-1, 0, +1\}. \tag{2}$$

The resampling distribution gives $\pi_{a,e}^{(m)} = [\Pr(Z=-1), \Pr(Z=0), \Pr(Z=+1)]$. For two contexts e_1, e_2,

$$D_{\mathrm{cross}} = \mathbb{E}_{a\sim w}[1 - \pi_{a,e_1}^{\mathsf{T}} \pi_{a,e_2}], \qquad D_{\mathrm{within,e}} = \mathbb{E}_{a\sim w}[1 - \|\pi_{a,e}\|_2^2]. \tag{3}$$

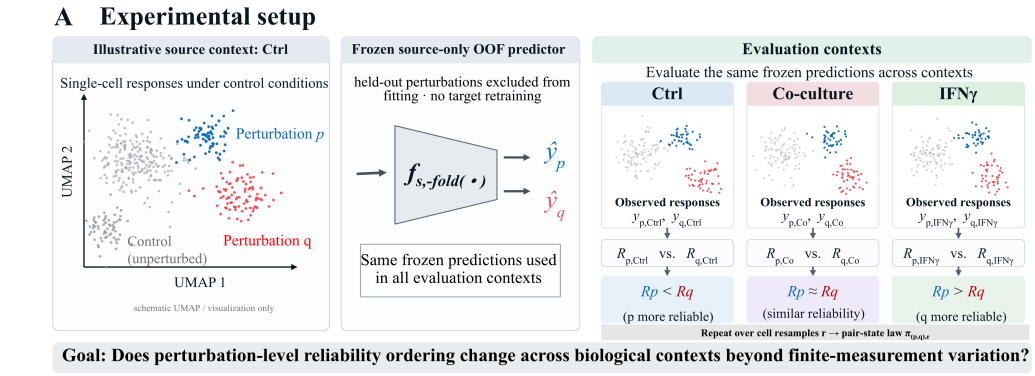
The measurement-adjusted discrepancy, equivalently one half of the squared delta-kernel MMD (Gretton et al., 2012), is

$$D_{\mathrm{adj}} = D_{\mathrm{cross}} - \frac{D_{\mathrm{within,e_1}} + D_{\mathrm{within,e_2}}}{2} = \frac{1}{2}\mathbb{E}_a \|\pi_{a,e_1} - \pi_{a,e_2}\|_2^2 \tag{4}$$

Here w is uniform over the declared unordered perturbation-pair universe and the delta kernel $k(z, z') = \mathbf{1}\{z = z'\}$ compares the three-valued pair states. Thus D_{adj} is the squared separation between the two ordering distributions after removing their within-context coincidence terms. It is a transport quantity for relative reliability, rather than an average change in prediction accuracy. For interpretability, write $t_e = \pi_{a,e}(0)$ and $b_e = \pi_{a,e}(+1) - \pi_{a,e}(-1)$. The pairwise squared separation has the exact decomposition

$$\begin{aligned} D_{\mathrm{adj}}(a) = D_{\mathrm{tie}}(a) + D_{\mathrm{strength}}(a) + D_{\mathrm{flip}}(a), \quad & D_{\mathrm{tie}}(a) = \tfrac{3}{4}(t_{e_1} - t_{e_2})^2, \\ & D_{\mathrm{strength}}(a) = \tfrac{1}{4}(|b_{e_1}| - |b_{e_2}|)^2, \\ & D_{\mathrm{flip}}(a) = |b_{e_1} b_{e_2}|\, \mathbf{1}\{b_{e_1} b_{e_2} < 0\}. \end{aligned} \tag{5}$$

Fig. 1 Reliability ordering under biological context shift



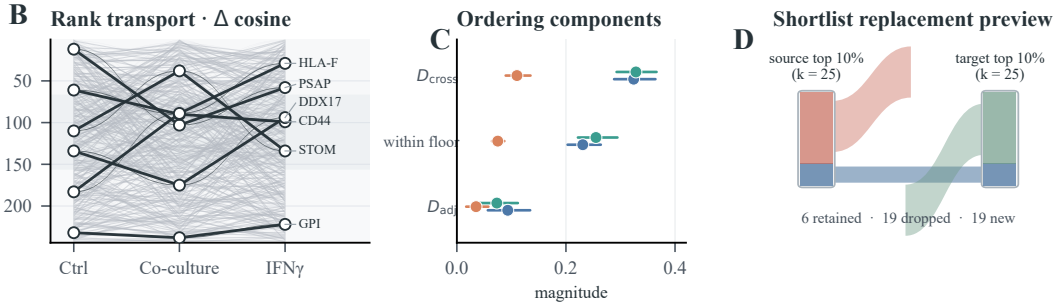


Figure 1: **Reliability ordering under biological context shift.** The source predictor is frozen before evaluation in the target context. The rank braid retains all 243 background rank trajectories and highlights a deterministic set of source-rank quantile exemplars for legibility. The compact shortlist preview reports 25 perturbations per context: 6 retained, 19 dropped and 19 newly entering the target shortlist. The full 243-label accounting remains in the Fig. 1 source table, while the detailed replacement flow is reserved for Fig. 4A. The decomposition uses the same source-frozen predictor and prespecified evaluation protocol on its declared eligible transfer set. Metric colors are fixed throughout the paper.

For finite measurement, the reported components use a cross-product of independent 15-seed halves while retaining both measurement replicates. The empirical report is therefore $D_{\text{tie}}^{\text{CF}}$ plus a directional-bias component; the nonlinear strength/flip split is retained as population algebra, not presented as an unbiased finite-sample decomposition. The canonical per-seed U-corrected D_{adj} is reconstructed from the merged member-level risks. Finite-measurement component calibration and the comparison between cross-fit and canonical estimates are reported in the Supplement; they are not additional population identities.

Estimand properties. Because each $\pi_{a,e}$ is a probability vector, $0 \leq D_{\text{adj}} \leq 1$. The discrepancy is symmetric in the two contexts and equals zero if and only if the pair-state distributions agree almost everywhere over the declared pair universe. It is therefore a squared divergence; we do not claim a triangle inequality and use "discrepancy" rather than "distance". The bound applies to the population estimand. Its unbiased finite-sample U-corrected estimate $\widehat{D}_{\text{adj}}$ is signed and is not clipped to $[0, 1]$; a negative estimate can occur when the observed pair-state discrepancy is smaller than the finite-replicate correction. In the deterministic tie-free limit, each pair law collapses to a point mass and D_{adj} reduces exactly to the normalized Kendall inversion fraction, $(1 - \tau)/2$. More generally, it equals cross-context pair disagreement minus one half of each within-context pair

disagreement, a population quantity that deterministic Kendall correlation does not define under repeated noisy measurement. Its distinct role is finite-measurement transport, where pair states are uncertain and the matched within-context component must be estimated rather than ignored. This distinction is not cosmetic. Consider one perturbation pair whose strict ordering probabilities are $(0.9, 0, 0.1)$ in the source and $(0.6, 0, 0.4)$ in the target. Both contexts have the same modal strict order, so Kendall after collapsing each pair law to its modal order records no inversion, whereas $D_{\mathrm{adj}}(a) = \frac{1}{2}[(0.9 - 0.6)^2 + (0.1 - 0.4)^2] = 0.09$. The estimand therefore detects transport of ordering reliability even when the modal order does not flip. Conversely, an expression-space MMD can be positive when every risk ordering is unchanged; it tests a different population object. Hence D_{adj} is neither renamed Kendall nor a generic distribution-shift score: it is the declared average separation of finite-measurement pair-state laws. For finite resampling sets, the within-context term uses the unbiased order-2 U-statistic (Hoeffding, 1948). A strict direction is declared stable when the 95% equal-tail Beta(1,1) posterior credible interval excludes 0.5, with at least eight non-tie Monte Carlo resampling outcomes supporting the strict comparison. This threshold diagnoses state-estimation stability and is not a count of biological replicates; otherwise the pair remains unresolved. This categorical state rule is separate from the 90% intervals reported for effect estimates. At the no-transport boundary the U-statistic is degenerate, so an ordinary percentile bootstrap is not used as a zero-effect test. Detectability is instead assessed from the empirical variance of the 30 independent seed-level U-corrected contributions using a Student-t interval. A complete-seed block bootstrap is retained as a sensitivity analysis. This preserves the registered seed as the independent unit and does not resample the two context arms independently.

The primary Frangieh surface contains 243 shared perturbation labels across three contexts. Fig. 1B selects source-rank quantile exemplars only to keep the complete braid legible, whereas its background ribbons and the Fig. 1 source table retain the full 243-label source-frozen surface. Fig. 1D is a compact decision preview; the detailed shortlist flow appears only in Fig. 4A. Aggregate transport estimates use their declared full eligible tables. The predictor uses a ten-component PCA basis, ridge penalty $\alpha = 0.1$, one global five-fold perturbation split and three independent model members. Matched raw-cell pseudoreplicates use 30 resampling seeds and a minimum of 20 cells per perturbation–context. The independent Nadig analysis uses the exact shared HepG2–Jurkat perturbation set and is not pooled with Frangieh.

2.1 Decision consequences

For this descriptive decision analysis, $\bar{R}^{(m)}_{p,e}$ is the mean full-depth risk over the frozen model members and measurement replicates for metric m. We set $k = \max(1, \lfloor q|\mathcal{P}| \rceil)$ for budget fraction q and break exact boundary ties by the fixed perturbation identifier. For a source context e_s and selection budget k, let

$$S_k(e_s) = \operatorname*{arg\,min}_{S:\,|S|=k} \sum_{p \in S} \bar{R}^{(m)}_{p,e_s} \tag{6}$$

be the source-ranked shortlist. Its target-context loss is $L_{e_t}(S) = k^{-1} \sum_{p \in S} \bar{R}^{(m)}_{p,e_t}$, and normalized selection regret compares $L_{e_t}(S_k(e_s))$ with the target-oracle shortlist $S_k(e_t)$. Thus the decision layer measures the consequence of transporting a source reliability order, rather than evaluating a second predictor.

For the decision layer, the target regret at budget k is defined as

$$\mathrm{Regret}_k(e_s \to e_t) = L_{e_t}(S_k(e_s)) - L_{e_t}(S_k(e_t))\,, \tag{7}$$

where the second term is the target-oracle loss. Let $S^{\mathrm{worst}}_k(e_t)$ be the deterministic k-item set with largest target risk and let $L_{e_t}(\mathcal{U})$ be the mean target risk over the eligible universe. We report two explicitly distinct normalizations,

$$\mathrm{Regret}^{\mathrm{rand}}_k = \frac{\mathrm{Regret}_k}{L_{e_t}(\mathcal{U}) - L_{e_t}(S_k(e_t))}, \qquad \mathrm{Regret}^{\mathrm{worst}}_k = \frac{\mathrm{Regret}_k}{L_{e_t}(S^{\mathrm{worst}}_k) - L_{e_t}(S_k(e_t))}. \tag{8}$$

The first preserves the random-reference scale used in the primary result; the second is bounded by the deterministic worst shortlist. Rows with a non-positive normalization denominator are undefined and excluded; no such row occurred in the reported surface. Both target-oracle quantities are descriptive benchmarks computed from the same target table, not prospective policies.

Fig. 2 Context-dependent reliability transport and independent replication

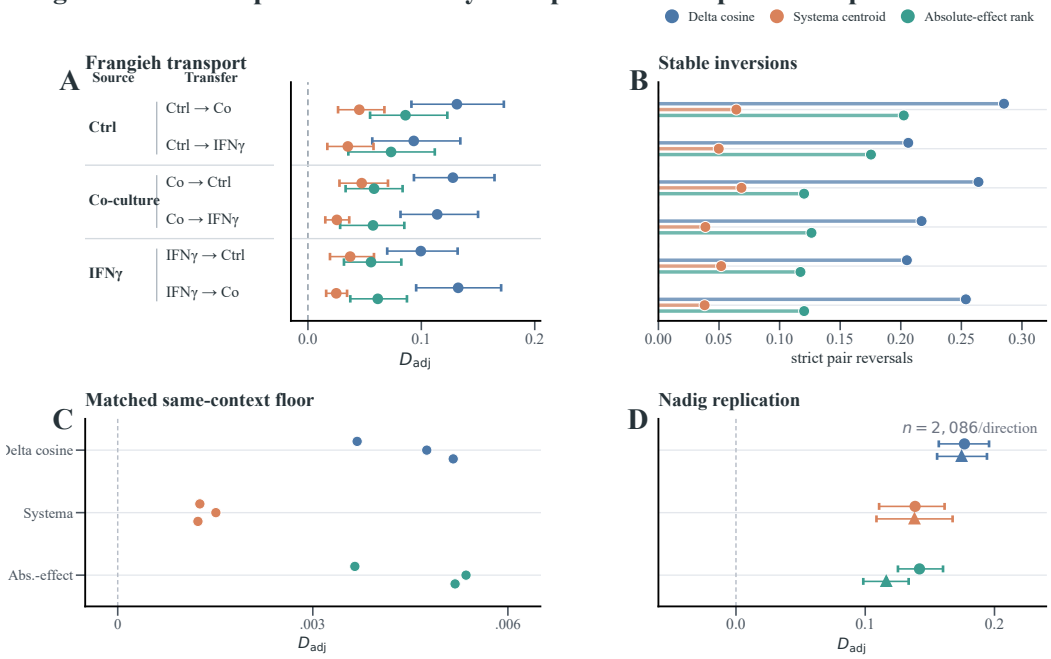


Figure 2: **Context-dependent reliability transport.** (A) Six directed Frangieh transfers with three metric estimates and 90% intervals. (B) Stable-pair inversion fraction: fraction of pairs stable in both contexts whose strict ordering reverses. (C) Same-context independent pseudoreplicate control, shown as point estimates because it is a floor diagnostic rather than a biological transfer. (D) Independent Nadig HepG2–Jurkat replication in both source directions. Metric color is fixed; in panel D, marker shape identifies the source direction and context is otherwise carried by position and labels.

3 Results

3.1 Reliability ordering changes across biological contexts

Figure 1 introduces the object of interest: predictions are held fixed while target biology changes, so movement in the perturbation order cannot be attributed to refitting the predictor. The full Frangieh atlas contains 243 shared labels; Fig. 1B retains all 243 background trajectories and highlights a deterministic set of source-rank quantile exemplars so that individual ribbons remain readable. Preprocessing and replication eligibility are summarized in Appendix A; the complete candidate registry is summarized in the supplementary material. Fig. 1D previews the fixed top-10% shortlist replacement example; the complete label-level risk accounting is retained in the figure source table. The same-context resampling control gives $D_{\mathrm{adj}} = 0.0012 - -0.0054$ across its nine context–metric cells, whereas the matched biological transfers produce larger measurement-adjusted values.

The full-size Ctrl-to-IFNγ transfer illustrates metric dependence. For delta cosine, $D_{\mathrm{cross}} = 0.324175$ and $D_{\mathrm{adj}} = 0.093331$ ($[0.056748, 0.134389]$); for Systema centroid, the corresponding values are 0.110196 and 0.035181 ($[0.017110, 0.057823]$); for absolute-effect rank they are 0.328296 and 0.073300 ($[0.035638, 0.111859]$). These generated values are read from the same prespecified analysis table used by the figure builder. The conservative joint-floor stress test is retained separately, because it is not an additive component of the measurement-adjusted estimand. The atlas in Fig. 2 retains metric and transfer heterogeneity rather than averaging it away. The finite-sample estimator and tie rule are given in Appendix B.

3.2 CALIBRATED INFERENCE DISTINGUISHES TRANSPORT FROM FINITE MEASUREMENT

The estimator addresses a finite-measurement target, so a useful comparison is not only against raw rank distances but also against their strongest simple matched-null variants. In a registered 120-condition known-truth simulation, raw Kendall, Spearman and top-10 Jaccard declared transport in 0.952, 0.995 and 1.000 of null conditions, respectively. Label bootstrapping alone did not remove this finite-sample effect. Subtracting a matched within-context null reduced the Kendall false-transport rate to 0.003 and the Jaccard rate to 0.000, while the U-corrected D_{adj} rate was 0.017. At non-zero inversion fractions, detection power was 1.000 for D_{adj} and 0.997 for matched-null Kendall. Thus the result is not that D_{adj} universally dominates every rank statistic: precise measurements make deterministic rank comparators reasonable, whereas D_{adj} supplies one coherent pair-state population estimand with an explicit finite-replicate correction. For the interval itself, direct variance estimation over the 30 independent seed-level U-corrected contributions yielded a null false-transport rate of 0.040 with 0.960 coverage in 2,880 known-truth trials, and 0.015 with 0.985 coverage in 3,000 Frangieh semi-synthetic null trials. These intervals are mildly conservative and replace the anti-conservative hierarchical percentile intervals in every detectability result. The full calibration table, including coverage, bias and the Jaccard comparison, is provided in the Supplement.

3.3 INDEPENDENT RAW-CELL SYSTEMS REPRODUCE ORDERING TRANSPORT

The independent Nadig replication provides a different biological system and context pair and a larger shared perturbation universe. The adjusted discrepancy D_{adj} is 0.176 for delta cosine, 0.138 for Systema centroid and 0.129 for absolute-effect rank, with positive 90% intervals in both source directions. This replication supports the ordering-level phenomenon without being used to tune the Frangieh decision association. Matched-universe sensitivity for this replication is reported in Appendix C, while the complete two-direction full-size surface remains in the figure source table.

A second raw-cell replication reconstructs the complete hierarchy in the Tian iPSC and day-7 neuron CRISPR screens (Tian et al., 2019), rather than using an aggregate rank audit. We froze 26 shared single-gene perturbations, a 1,536-gene panel chosen from source controls only, and a 40-cell budget before evaluating either direction. Five of six direction–metric intervals were positive. From iPSC to neuron, D_{adj} was 0.032 [0.005, 0.060], 0.010 [-0.010, 0.030], and 0.016 [0.004, 0.028] for delta cosine, Systema and absolute-effect rank, respectively; the reverse direction gave 0.049 [0.018, 0.079], 0.042 [0.025, 0.060], and 0.013 [0.001, 0.025]. The direction-specific estimates support context-dependent ordering transport; their asymmetry is reported descriptively rather than treated as a separate biological conclusion. The complete six-row ledger and source-frozen construction are in Appendix C.

3.4 SOURCE-FROZEN PREDICTOR FAMILIES PRESERVE THE QUALITATIVE FINDING

As one matched predictor-family sensitivity, we also evaluated a source-only RBF kernel-ridge family on the same Frangieh surface. Its source-only penalty selection required an expanded grid. The matched transport profile had Spearman $\rho = 0.983$ with the primary bilinear-ridge profile (within-metric profile correlations 0.829–1.000; mean row-wise prediction-vector cosine 0.733 in the Supplement), and adjusted-discrepancy point and interval signs agreed on all 18 transfer–metric rows. This is descriptive evidence that the profile is not an immediate artifact of one parameterization, not predictor- family equivalence or broader generalization.

We then ran the official GEARS graph predictor under the same strict information contract. For each Frangieh source context, five label-disjoint folds and all graph construction, fitting and tuning used source cells only; 13 labels absent from the official interaction graph were excluded before splitting, leaving 230 perturbations and 8,229 evaluation genes. The frozen out-of-fold vectors were evaluated by the same 30-seed, two-pseudoreplicate raw-cell bridge. All 18 directed transfer–metric rows retained positive adjusted-discrepancy intervals. Relative to the bilinear profile, GEARS had perfect point-sign agreement but metric-dependent concordance (mean within-metric Spearman $\rho = 0.371$: -0.086, 1.000 and 0.200 for delta cosine, Systema centroid and absolute-effect rank). Delta cosine was effectively uncorrelated. Thus a modern graph model reproduces the existence of transport while showing that its detailed geometry is predictor- and metric-dependent; it does not establish family equivalence or superiority. The complete surface is reported in Supplementary Appendix H.

Fig. 3 Cell budget determines when reordering is resolvable

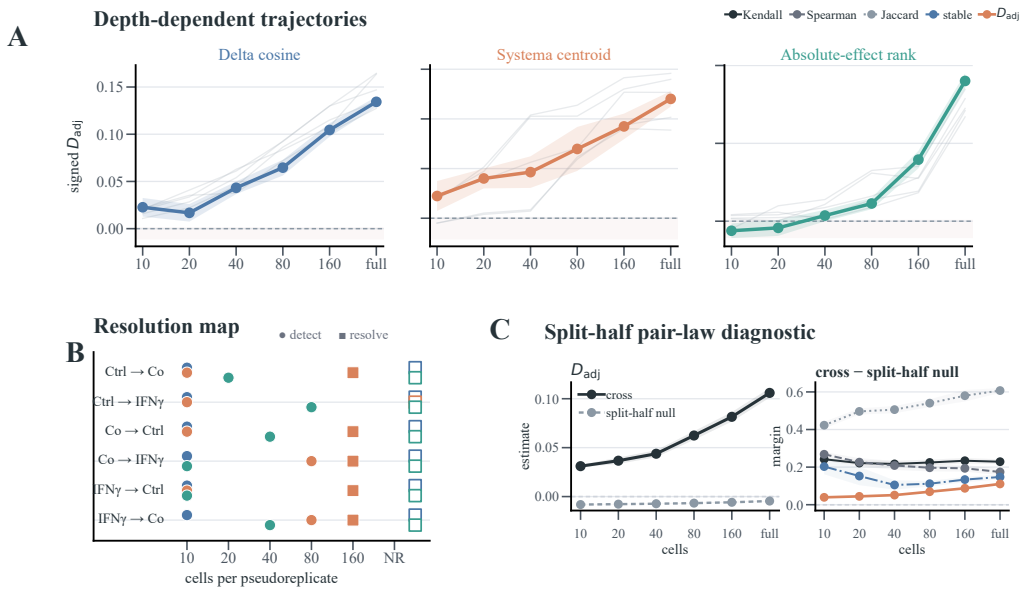


Figure 3: **Cell budget defines a finite-measurement resolution boundary.** (A) Signed depth trajectories for the primary Ctrl–IFNγ contrast with other transfers as pale paths. (B) The six-transfer resolution map; circles mark detectability and squares mark the resolution budget, with open glyphs denoting not reached. (C) Cross-context transport and matched same-context split-half pair-law null margins across depth for the five prespecified comparator families. This is a finite-measurement diagnostic, not the canonical per-seed D_{adj} summary. The full-depth endpoint is the fixed-160 eligibility universe ($n = 139$), not the wider full-size minimum-20 universe used for Figs. 1–2. Signed finite-sample estimates are retained, so null diagnostics can be negative. Detectability and quantitative resolution are separate experimental requirements under the declared resampling protocol.

3.5 Cell budget determines when reordering is resolvable

We next hold the perturbation universe fixed and resample 10, 20, 40, 80 and 160 cells per perturbation–context pseudoreplicate, together with a matched full-depth endpoint. The signed $D_{\mathrm{adj}}(n)$ trajectory is retained. The first budget within ten percent of the full-depth value summarizes resolution; detectability requires a positive lower endpoint of the 90% interval. These are metric- and transfer-specific summaries, not a universal cell threshold. We additionally compare each transport statistic with a matched same-context pseudo-context null at every depth, because the canonical within-context measurement term is not itself a same-context D_{adj} estimate.

The trajectories rise at different rates for the three metrics, and several source–contrast cells remain unresolved at the available budgets. Increasing cell budget therefore improves resolution under this protocol without producing a single portable cutoff. This measurement boundary is a prerequisite for interpreting a context-associated ordering difference beyond the measured sampling floor. The matched-null analysis shows positive cross-minus-null margins for every comparator at every depth, but does not support presenting D_{adj} as a universally superior rank statistic; its role is the explicit tie-aware, measurement-adjusted estimand. The depth comparator audit is reported in the Supplementary finite-measurement audit. The depth definitions and all transfer-specific results are in Appendix D.

Fig. 4 From reordering to experimental regret

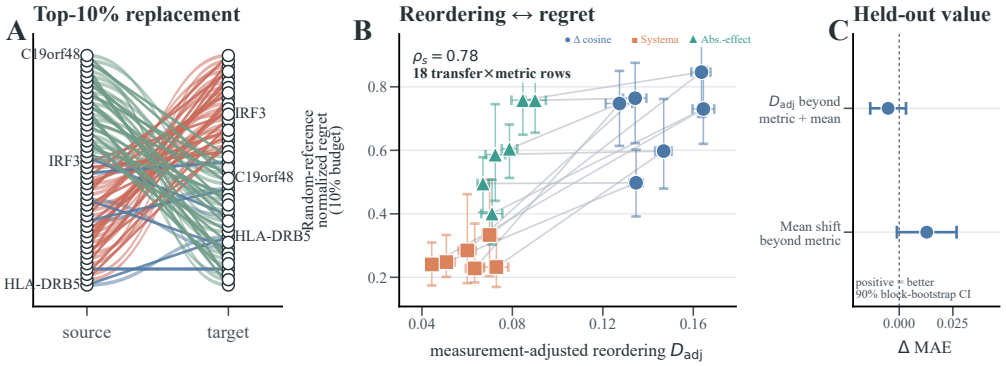


Figure 4: **Context-dependent ordering changes experimental priorities.** (A) Detailed example of delta-cosine top-k replacement under the Ctrl-to-IFNγ shift (6 of 25 shortlist positions are retained; 19 are dropped and 19 are new). (B) The 18 full-depth transfer–metric observations; marker shape identifies metric and pale connectors identify directed transfer. The pointwise error bars are canonical 90% intervals; the descriptive association and its unordered-context-pair sensitivity are reported in the text, and no fitted line is implied. (C) Unordered-context-pair-held-out incremental prediction of normalized regret beyond metric identity; intervals are block-bootstrap diagnostics.

3.6 Context-dependent reordering changes shortlist decisions

To connect the ordering to an experimental action, each source context selects the lowest-risk perturbations at budgets $k/n \in \{0.05, 0.10, 0.20, 0.50\}$ and the target context evaluates the selected set. We report target top-k retention and normalized regret under both a random-reference and a bounded worst-oracle normalization.

The primary full-depth link contains six directed source–target surfaces and three metrics, hence 18 descriptive rows. Across all rows, adjusted reordering and normalized regret have pooled Spearman $\rho = 0.78$; after ranking within metric, the association is $\rho = 0.26$. The three unordered context-pair and source-context sensitivities are reported separately. The association is a decision-level summary of the declared surface, not a claim that every predictor or biological system will have the same slope.

An explicit unordered-context-pair leave-one-out comparison separates metric identity, mean-fidelity shift, and ordering transport. Adding the matched mean-risk shift to metric identity reduces OOF MAE by 0.013 ($[-0.001, 0.027]$). Adding D_{adj} increased OOF MAE from 0.087 to 0.092, corresponding to an improvement score of -0.005 ($[-0.014, 0.003]$); the two nested contrasts are shown separately in Fig. 4. Thus the pooled association is not interpreted as a unique causal contribution of one scalar ordering summary. The corresponding unordered-pair block-bootstrap table is reported in the Supplement; no headline permutation p-value is used for only three blocks.

The shortlist view makes the mechanism concrete: perturbations that occupy the source top-k need not occupy the target top-k, so a source ranking can lose retention and incur regret even when the predictor itself has not changed. The budget-wise association summaries show that the link is not confined to one selection fraction, while its magnitude depends on the transfer and metric. The four-budget, two-normalization robustness summary is reported in Appendix D. A metric-matched mean-risk control, using the same eligible labels and frozen predictions, shows that context-average fidelity and reliability- order transport are empirically non-equivalent (Supplementary Fig. S3).

Additional target-informed and prospective source-only analyses are reported in the Supplement. They are kept separate from the post-hoc transport claim: the target-informed audit uses target outcomes, whereas the prospective analysis uses source-observable features only and reserves target outcomes exclusively for evaluation.

4 DISCUSSION

The central result is that perturbation reliability is an environment-dependent ordering, not a fixed property of a perturbation. With predictions held fixed, matched biological context shifts reordered realized risk beyond a same-context sampling floor, and the direction recurs in both directions of an independent HepG2–Jurkat context pair. The result is therefore an empirical reliability-transport object: it does not claim that every predictor or biological system must reorder in the same way.

Unlike response-prediction benchmarks (Lotfollahi et al., 2019; 2023; Bunne et al., 2023; Roohani et al., 2024; Cui et al., 2024; Wei et al., 2026), our object asks whether perturbation-level error ordering remains useful after biology changes. It is also distinct from experimental perturbation reproducibility (Wang et al., 2026): predictor, metric and context are part of the estimand.

Cell budget separates detectability from resolution of transport magnitude. The tie-aware support rule and matched same-context control make this distinction explicit. In the deterministic tie-free limit, D_{adj} reduces to the normalized Kendall inversion fraction; here it extends the same question to uncertain pair states under finite repeated measurement.

Source-ranked top-k choices lose target retention when reliability moves, but this is a diagnostic rather than a decision-focused learner (Donti et al., 2017; Wilder et al., 2019; Elmachtoub & Grigas, 2022). The 18 rows arise from only three unordered context-pair blocks. Metric-stratified associations are weaker than the pooled surface; mean-risk shift provides the only modest OOF improvement (0.013; $[-0.001, 0.027]$) and the additional D_{adj} contrast beyond mean shift does not add held-out value (-0.005; $[-0.014, 0.003]$). We therefore do not claim that one scalar universally predicts regret.

Scope and testable extensions. The primary decision surface is limited to one bilinear-ridge predictor, three metrics and three unordered Frangieh context pairs. Matched RBF, neural and GEARS analyses test predictor sensitivity but are not pooled with it; the Nadig and Tian reconstructions replicate biological transport but do not enlarge the decision sample. TxPert lacked a source-matched Frangieh checkpoint, and the outcome-blind Feng aggregate audit was neutral, so neither is promoted as a positive replication. The next test is a prospectively frozen, label-disjoint analysis across additional contexts and modalities. This scope is important as context diversity becomes central to virtual-cell evaluation (Dibaeinia et al., 2026).

Reproducibility. The anonymous core release provides fixed splits, source tables and commands for the four main figures, Supplementary Figs. S1–S7 and compact tables. Appendices A, B and I specify the data, estimators and regeneration path; engineering artifacts are excluded from manuscript evidence.

Code and data availability. The anonymized code repository is available at `https://anonymous.4open.science/r/PTL-B377/`. It contains analysis and figure code, manuscript sources and compact result products. Raw matrices are not redistributed. Frangieh and Nadig data are available from the public scPerturb v1.4 Zenodo record; Feng data are available from its Cell Genomics record (Feng et al., 2026). Public accession records and regeneration commands are provided in the anonymous release.

AI USE STATEMENT

AI-assisted tools supported literature triage, mathematical and experimental brainstorming, implementation and debugging of analysis code, figure-script development, interpretation checks, and language editing. The authors designed the study, inspected all generated analyses, verified citations and numerical claims against source artifacts, and take responsibility for the manuscript.

REFERENCES

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

A multiplexed single-cell CRISPR screening platform enables systematic dissection of the unfolded protein response. *Cell*, 167(7):1867–1882.e21, 2016. doi: 10.1016/j.cell.2016.11.048.

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

Charlotte Bunne, Stefan G. Stark, Gabriele Gut, Jacobo Sarabia del Castillo, Mitch Levesque, Kjong-Van Lehmann, Lucas Pelkmans, Andreas Krause, and Gunnar Raetsch. Learning single-cell perturbation responses using neural optimal transport. *Nature Methods*, 20(11):1759–1768, 2023. doi: 10.1038/s41592-023-01969-x.

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.

Payam Dibaeinia, Sudarshan Babu, Mei Knudson, Ali ElSheikh, Yibo Wen, Han Liu, Jason Perera, and Aly A. Khan. Virtual cells need context, not just scale. *bioRxiv*, 2026. doi: 10.64898/2026.02.04.703804.

Atray Dixit, Oren Parnas, Biyu Li, Jenny Chen, Charles P. Fulco, Livnat Jerby-Arnon, Nemanja D. Marjanovic, Danielle Dionne, Tyler Burks, Raktima Raychowdhury, Britt Adamson, Thomas M. Norman, Eric S. Lander, Jonathan S. Weissman, Nir Friedman, and Aviv Regev. Perturb-Seq: Dissecting molecular circuits with scalable single-cell RNA profiling of pooled genetic screens. *Cell*, 167(7):1853–1866.e17, 2016. doi: 10.1016/j.cell.2016.11.038.

Priya L. Donti, Brandon Amos, and J. Zico Kolter. Task-based end-to-end model learning in stochastic optimization. In *Advances in Neural Information Processing Systems*, volume 30, pp. 5484–5494, 2017. URL https://papers.nips.cc/paper/2017/hash/3fc2c60b5782f641f76bcefc39fb2392-Abstract.html.

Adam N. Elmachtoub and Paul Grigas. Smart "predict, then optimize". *Management Science*, 68 (1):9–26, 2022. doi: 10.1287/mnsc.2020.3922.

Claudia Feng, Elin Madli Peets, Yan Zhou, et al. A genome-scale single-cell CRISPRi map of trans gene regulation across human pluripotent stem cell lines. *Cell Genomics*, 6(2):101076, 2026. doi: 10.1016/j.xgen.2025.101076. URL https://pmc.ncbi.nlm.nih.gov/articles/PMC12903452/. Public multi-line, multi-donor targeted-screen resource accessed for an outcome-blind external replication audit.

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

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

Wassily Hoeffding. A class of statistics with asymptotically normal distribution. *The Annals of Mathematical Statistics*, 19(3):293–325, 1948. doi: 10.1214/aoms/1177730196.

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

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

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*, 57:1228–1237, 2025. doi: 10.1038/s41588-025-02169-3.

Stefan Peidli, Tessa D. Green, and Ciyue Shen. 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.

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.

Aravind Sankaran, Lars Karlsson, and Paolo Bientinesi. Ranking with ties based on noisy performance data. *International Journal of Data Science and Analytics*, 20:4363–4384, 2025. doi: 10.1007/s41060-025-00722-1.

Jordan W. Squair, Michael A. Skinnider, Matthieu Gautier, Leonard J. Foster, and Grégoire Courtine. Prioritization of cell types responsive to biological perturbations in single-cell data with Augur. *Nature Protocols*, 16(8):3836–3873, 2021. doi: 10.1038/s41596-021-00561-x.

Ruilin Tian, Mariam A. Gachechiladze, Connor H. Ludwig, Matthew T. Laurie, Jason Y. Hong, Diane Nathaniel, Anika V. Prabhu, Michael S. Fernandopulle, Rajan Patel, Mehrnoosh Abshari, Michael E. Ward, and Martin Kampmann. Crispr interference-based platform for multimodal genetic screens in human ipsc-derived neurons. *Neuron*, 104(2):239–255.e12, 2019. doi: 10.1016/j.neuron.2019.07.014.

Ramon Vinas Torne, Maciej Wiatrak, Zoe Piran, Shuyang Fan, Liangze Jiang, Sarah A. Teichmann, Mor Nitzan, and Maria Brbic. 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.

Zhiting Wei, Yiheng Wang, Yicheng Gao, Shuguang Wang, Ping Li, et al. Benchmarking algorithms for generalizable single-cell perturbation response prediction. *Nature Methods*, 23:451–464, 2026. doi: 10.1038/s41592-025-02980-0. URL `https://www.nature.com/articles/s41592-025-02980-0`.

Frederik Wenkel, Wilson Tu, Cassandra Masschelein, Hamed Shirzad, Liam Hodgson, Ihab Bendidi, Cian Eastwood, Shawn T. Whitfield, Craig Russell, Yassir El Mesbahi, Jiarui Ding, Marta M. Fay, Berton Earnshaw, Emmanuel Noutahi, and Alisandra K. Denton. TxPert: using multiple knowledge graphs for prediction of transcriptomic perturbation effects. *Nature Biotechnology*, 2026. doi: 10.1038/s41587-026-03113-4.

Bryan Wilder, Bistra Dilkina, and Milind Tambe. Melding the data-decisions pipeline: Decision-focused learning for combinatorial optimization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pp. 1658–1665, 2019. doi: 10.1609/aaai.v33i01.33011658.

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

Or Zuk, Liat Ein-Dor, and Eytan Domany. Ranking under uncertainty. In *Proceedings of the 23rd Conference on Uncertainty in Artificial Intelligence*, pp. 466–473, 2007.

Supplementary Material

Finite-Measurement Transport of Perturbation Reliability Rankings

Organization. This appendix supplies the data and source-frozen protocol (A), estimators and uncertainty decomposition (B), extended transport and independent replication evidence (C), measurement and decision robustness (D), source-only prospective analyses (E), additional robustness and diagnostics (F), calibrated inference (G), external predictors and biological audits (H), and scientific regeneration (I). It complements rather than repeats the main-text sequence: observed reordering $\rightarrow$ measurement-adjusted transport $\rightarrow$ independent replication $\rightarrow$ measurement boundary $\rightarrow$ decision consequence.

A Data and source-frozen evaluation protocol

The primary surface is the Frangieh Perturb-CITE-seq melanoma data set (Frangieh et al., 2021). Using raw-count metadata, we first retain cells with at least 700 detected genes, 1,500 counts, and at most 25% mitochondrial counts; retained cells are then normalized to a total of 10,000 and transformed by $\log(1 + x)$. After cell QC, the analysis uses the fixed 8,229-gene evaluation panel. Control, co-culture, and IFNγ form the three biological contexts. A source-frozen predictor is fit at the source context and evaluated without re-training in the target context. Fig. 1B retains all 243 background rank trajectories and highlights a deterministic set of source-rank quantile exemplars for legibility; the full shortlist accounting retains all 243 labels.

The independent HepG2–Jurkat replication uses the Nadig data set (Nadig et al., 2025). Its target-sum-10,000/$\log(1 + x)$ preprocessing is fixed before evaluation. The primary analysis requires 20 cells per condition, uses five label-disjoint folds, three frozen ensemble members, and 30 resampling seeds. These are evaluation settings, not fitted outcomes.

Supplementary Table S1: Eligibility conditions for the independent replication. Rejected and incomplete candidates are summarized in the provenance-audit sections.

Criterion	Requirement	Nadig evidence
Shared perturbations	matched labels across contexts	2,392
Cell support	minimum 20 cells per condition	87.2% eligible
Matched controls	controls available in both contexts	4,976 / 12,013 cells
Technical comparability	same library, RNA readout, and harvest design	supported
Major context confounding	no obvious context-linked batch confound	none identified

B Measurement-adjusted reliability transport

Pairwise risk comparisons retain ties as an explicit ordering state rather than coercing them into a strict direction. Separately, exact/tolerance ties are excluded from the strict-sign Beta posterior used for the stable-direction diagnostic. A strict direction is declared stable when the 95% equal-tail Beta(1,1) posterior credible interval excludes 0.5, with at least eight non-tie Monte Carlo resampling outcomes supporting the strict comparison; otherwise the pair remains unresolved. This threshold diagnoses state-estimation stability and is not a minimum number of biological replicates.

We report the finite-sampling floor, the measurement-adjusted discrepancy, and a conservative joint-floor stress test separately. Final detectability intervals use the empirical variance of complete seed-level U-corrected contributions and a Student-t reference distribution; label/member bootstrap intervals are retained only where explicitly marked as descriptive sensitivity analyses. The decision-link association is descriptive over 18 directed source–target surface rows; its dependence-aware sensitivity resamples three unordered context-pair blocks while preserving both directions and all metric rows. No target labels or target-derived features enter the source-frozen predictor.

For a finite set of M resampling replicates, let Z_r denote the tie-aware state for one perturbation pair in replicate r. The within-context coincidence term is estimated with the unbiased order-2 U-statistic

$$\widehat{W}_U = \frac{1}{M(M-1)} \sum_{r \neq r'} \mathbf{1}\{Z_r = Z_{r'}\}, \qquad \widehat{D}_{\text{within}} = 1 - \widehat{W}_U.$$

This correction removes the positive finite-replicate bias of the plug-in coincidence estimate under a null context pair. All exact comparisons use a fixed tolerance of 10^{-8}. The strict-direction diagnostic uses a Beta$(1, 1)$ prior after excluding exact/tolerance ties; its posterior support count is the number of remaining strict observations. The metric-dependence diagnostic uses 15 prespecified discovery seeds and 15 held-out validation seeds, while the primary effect estimates retain the declared 30-seed resampling contract.

The component analysis uses the empirical pair-order law rather than the posterior tail probability used for stable-state classification. For each directed transfer and metric, both measurement replicates are retained and independent 15-seed halves form a cross-product estimator. The finite-sample report contains a tie component and a directional-bias component, compared with the canonical per-seed U-corrected D_{adj} on the same eligible universe. The population strength/flip identity is retained algebraically, but its nonlinear empirical split is not treated as unbiased. At the population level $0 \leq D_{\text{adj}} \leq 1$, but the unbiased finite-sample estimate $\widehat{D}_{\text{adj}}$ is signed and is not clipped; negative values indicate that the observed pair-state discrepancy is smaller than the finite-replicate correction. In the deterministic tie-free limit, each pair law is a point mass and D_{adj} reduces exactly to the normalized Kendall inversion fraction $(1 - \tau)/2$. This is a limit check, not a substitute for the finite-measurement estimand.

B.1 Why the finite-measurement estimand is not interchangeable with Kendall or generic MMD

For each perturbation pair, D_{adj} compares the complete categorical laws on $\{-1, 0, +1\}$. For example, source and target laws $(0.9, 0, 0.1)$ and $(0.6, 0, 0.4)$ have the same modal strict order, so a deterministic Kendall comparison after collapsing both laws to their modal orders records no inversion. Nevertheless, their pair-state discrepancy is 0.09: the ordering has become substantially less reliable in the target. This is the population quantity that is absent from a point-ranking comparison. Conversely, a generic expression-space MMD can detect distributional changes that preserve every risk ordering. The exact identity with one half of squared MMD holds only for the declared delta kernel on pair states and does not turn D_{adj} into a generic expression two-sample statistic.

The U-corrected estimator is unbiased when complete measurement-seed blocks are independent across seeds and identically distributed within each context. The two pseudoreplicates and three predictor members inside a seed are dependent subunits and are aggregated before variance estimation; shared controls remain inside the same seed block. The frozen predictor and perturbation panel do not depend on target outcomes, and the two contexts are evaluated with the declared resampling design. At the null $\pi_{a,e_1} = \pi_{a,e_2}$, its first-order influence vanishes and the order-2 statistic is degenerate. Ordinary percentile bootstrap intervals are therefore not used to decide whether transport is detectable. The calibrated procedure treats the 30 seed-level U-corrected contributions as independent units, estimates their variance directly, and uses a Student-t interval for their mean. A complete-seed block bootstrap is retained as a sensitivity analysis; the two context arms are never resampled independently.

C Extended transport evidence and independent replication

The full-surface transport diagnostics contain the shared labels available to each matched context pair and are reported in the figure-source tables. Aggregate D_{adj} estimates are formed from globally merged eligible per-transfer sets; shard-local pairwise summaries are never averaged. Same-context controls are near zero (0.001–0.005), showing that the signal is not a generic artifact of the pairwise estimator.

Table S2 gives the matched-universe sensitivity of the independent Nadig replication at the two prespecified eligibility thresholds (minimum 20 versus 40 cells per condition). The six rows cover all three metrics at each threshold; the complete two-direction full-size surface remains in the figure

Supplementary Table S3: Independent Tian raw-cell replication. Intervals are calibrated 90% Student-t intervals over 30 independent seed-level U-corrected contributions. All rows use 26 shared perturbations, 1,536 source-selected genes and 40 cells per condition.

Source → target	Metric	D_{adj}	90% CI	Detected
iPSC → day-7 neuron	Delta cosine	0.032	[0.005, 0.060]	yes
iPSC → day-7 neuron	Systema centroid	0.010	[−0.010, 0.030]	no
iPSC → day-7 neuron	Absolute-effect rank	0.016	[0.004, 0.028]	yes
day-7 neuron → iPSC	Delta cosine	0.049	[0.018, 0.079]	yes
day-7 neuron → iPSC	Systema centroid	0.042	[0.025, 0.060]	yes
day-7 neuron → iPSC	Absolute-effect rank	0.013	[0.001, 0.025]	yes

source table. The values are presented as measurement-adjusted components rather than as conventional prediction accuracy, keeping the replication aligned with the main claim.

For universe accounting, the Nadig registry contains 2,392 shared labels; 2,086 labels per direction meet the primary minimum-20-cell eligibility rule, and the matched sensitivity universes contain 1,255 labels at minimum 20 and 345 labels at minimum 40 after intersecting the two directions and the fixed metric panel. These counts are registry properties, not model-selected subsets.

Supplementary Table S2: Nadig matched-universe sensitivity: measurement-adjusted ordering disagreement (D_{adj}; 90% bootstrap interval) at two minimum-support thresholds.

Universe	Metric	D_{adj} [90% CI]
min20 / 1255	Delta cosine	0.170 [0.152, 0.188]
min20 / 1255	Systema centroid	0.119 [0.092, 0.141]
min20 / 1255	Absolute-effect rank	0.108 [0.093, 0.124]
min40 / 345	Delta cosine	0.168 [0.128, 0.204]
min40 / 345	Systema centroid	0.085 [0.047, 0.124]
min40 / 345	Absolute-effect rank	0.073 [0.044, 0.103]

C.1 Second raw-cell replication: Tian iPSC to day-7 neuron

We additionally rebuilt the full source-frozen, cell-resampled hierarchy in the Tian CRISPR screens (Tian et al., 2019). Eligibility was frozen to the 26 single-gene perturbations represented by at least 80 cells in both systems. For each source direction, the 1,536-gene evaluation panel was chosen by variance in source control cells while excluding perturbation targets. Source effects were fitted on a disjoint half of source cells; evaluation then used 40 cells per perturbation and control in each of 30 seeds with two independent measurement replicates. Target cells were absent from panel selection and prediction construction. Table S3 reports the calibrated Student-t intervals over per-seed U-corrected contributions. Five of six intervals exclude zero; the iPSC-to-neuron Systema row does not. This is a complete raw-cell replication of the estimand and measurement floor, not an aggregate rank-distance audit. Its scope is nevertheless the 26 eligible perturbations, not the complete screened perturbation universe.

D Measurement and decision robustness

Depth analyses use matched pseudoreplicates at 10, 20, 40, 80, and 160 cells per perturbation–context condition, with the full-depth estimate as reference. Detectability is the first depth whose 90% interval lower bound exceeds zero; resolution is the first depth within 10% of the full-depth estimate.

The decision link is robust over 5%, 10%, 20%, and 50% shortlist budgets, using the random-reference regret scale of the main analysis and a bounded worst-oracle scale as a secondary diagnostic. Table S4 reports the descriptive surface association and unordered-pair/source-context leave-out ranges. These rows are a compact counterpart to the main-text decision figure, not a second set of decision claims.

For the random-reference normalization, a uniformly sampled size-k subset of the eligible universe has expected mean target loss $\mathbb{E}[L_{e_t}(S_k^{\mathrm{random}})] = L_{e_t}(\mathcal{U})$. Thus the denominator is the expected random-shortlist loss minus the oracle loss, rather than the loss of one particular random draw.

Supplementary Table S4: Decision-link robustness across shortlist budgets and regret normalizations. The random-reference rows use the same canonical full-depth decision surface as Fig. 4B; the bounded worst-oracle rows are a secondary source-frozen diagnostic. The last column is the leave-one-unordered-context-pair-out range.

Budget	Regret scale	ρ	Unordered-pair LOPO range
5%	random-reference normalized regret	0.769	[0.748, 0.783]
5%	worst-oracle-normalized regret	0.701	[0.657, 0.776]
10%	random-reference normalized regret	0.775	[0.755, 0.811]
10%	worst-oracle-normalized regret	0.796	[0.755, 0.818]
20%	random-reference normalized regret	0.886	[0.860, 0.916]
20%	worst-oracle-normalized regret	0.806	[0.741, 0.839]
50%	random-reference normalized regret	0.948	[0.930, 0.958]
50%	worst-oracle-normalized regret	0.889	[0.832, 0.902]

D.1 Transport decomposition and rank comparators

The full-surface decomposition contains 18 directed transfer–metric rows on the matched-fixed universe (139 or 186 perturbations, hence 9,591 or 17,205 unordered pairs). Table S5 reports a finite-measurement cross-product for the tie and directional-bias components, and compares their sum with canonical per-seed U-corrected D_{adj}. The population strength/flip identity is retained algebraically, but its nonlinear empirical split is not presented as unbiased. Table S6 compares alternative rank summaries on the identical 18-row surface; these are descriptive comparator analyses and do not change the primary estimand. The depth-convergence ledger records each comparator, its same-context floor, and $|T(n) - T(\mathrm{full})|$ for every matched depth.

Supplementary Table S5: Finite-measurement cross-product decomposition on the matched-fixed full-depth surface. The empirical report contains the tie and directional-bias components; the population strength/flip identity is not treated as an unbiased finite-sample split.

Transfer	Metric	$D_{\mathrm{tie}}^{\mathrm{CF}}$	$D_{\mathrm{dir}}^{\mathrm{CF}}$	Sum	D_{adj}	Difference
Co → Ctrl	Abs.-effect rank	0.000	0.090	0.090	0.072	0.017
Co → Ctrl	Δ cosine	0.000	0.160	0.160	0.147	0.013
Co → Ctrl	Systema	0.008	0.083	0.091	0.073	0.018
Co → IFNγ	Abs.-effect rank	0.000	0.083	0.083	0.067	0.016
Co → IFNγ	Δ cosine	0.000	0.152	0.152	0.135	0.017
Co → IFNγ	Systema	0.008	0.053	0.061	0.051	0.010
Ctrl → Co	Abs.-effect rank	0.000	0.101	0.101	0.085	0.017
Ctrl → Co	Δ cosine	0.000	0.181	0.181	0.164	0.018
Ctrl → Co	Systema	0.008	0.080	0.088	0.070	0.018
Ctrl → IFNγ	Abs.-effect rank	0.000	0.109	0.109	0.090	0.019
Ctrl → IFNγ	Δ cosine	0.000	0.156	0.156	0.134	0.022
Ctrl → IFNγ	Systema	0.008	0.067	0.075	0.060	0.015
IFNγ → Co	Abs.-effect rank	0.000	0.085	0.085	0.071	0.014
IFNγ → Co	Δ cosine	0.000	0.181	0.181	0.164	0.016
IFNγ → Co	Systema	0.008	0.050	0.058	0.044	0.014
IFNγ → Ctrl	Abs.-effect rank	0.000	0.092	0.092	0.079	0.014
IFNγ → Ctrl	Δ cosine	0.000	0.146	0.146	0.127	0.019
IFNγ → Ctrl	Systema	0.008	0.072	0.080	0.063	0.017

Supplementary Table S6: Rank-comparator benchmark on the same 18 directed transfer–metric rows. The final column is the descriptive Spearman association with normalized regret.

Comparator	n	Spearman with regret
Kendall $(1-\tau_b)/2$	18	0.87
Spearman $(1-\rho)/2$	18	0.84
Top-10% Jaccard distance	18	0.84
Stable-pair inversion	18	0.85
D_{adj}	18	0.78

D.2 Finite-depth discrimination against a matched null

The reported measurement-floor quantity is the within-context term subtracted by the primary estimator; it is not itself a same-context D_{adj} shift. We therefore recomputed a matched pseudo-context null at each depth by splitting the 30 seeds into independent 15-seed halves, keeping both measurement replicates in each half. The cross-context estimate uses both held-out orientations, and the same-context null uses the two within-context half comparisons. This produces 30 macro rows and 540 surface rows across the five comparator families and six depths. Because the finite D_{adj} unit is consequential, we also compared this flattened 15-seed$\times$2-replicate construction with a seed-block-preserving audit that U-corrects each aligned two-replicate seed block before averaging. Across the 108 D_{adj} null rows, the current construction was negative in 108/108 rows (mean -0.00673), whereas the seed-block audit was negative in 63/108 (mean -0.000334); the maximum absolute difference was 0.0302 and rank correlation was 0.074. The constructions therefore do not estimate the same finite-sample quantity. The main-text Fig. 3C is explicitly labeled a split-half pair-law diagnostic, while the canonical primary D_{adj} remains the per-seed U-corrected summary. No negative estimate was clipped or selected away. All comparator margins were positive at every depth, but their magnitudes differ substantially; in particular, top-10 Jaccard has the largest full-depth margin whereas D_{adj} is the finite-measurement pair-state formulation, not a universally dominant rank summary. The stable-inversion comparator has lower evaluable coverage at low depth and is interpreted accordingly. In addition, we computed a separate matched-unit audit for all five comparator families. Each aligned two-replicate seed block is collapsed to one block-level vector for the rank comparators; D_{adj} is U-corrected within each aligned block and averaged, and the stable-direction fraction uses the same 15 block-level vectors with the frozen support rule. This audit has one row per surface, depth, and comparator (540 rows) and is not substituted for the primary flattened-partition estimate. All five matched-unit margins remain positive in 108/108 rows; for D_{adj}, 3/108 matched cross estimates and 63/108 matched same-context null estimates are negative, while the margin remains positive in every row. We therefore retain the signed estimates and do not interpret the finite-sample null sign as a biological effect.

Supplementary Table S7: Finite-measurement discrimination at full depth. Each comparator is evaluated on the same 18 directed transfer–metric surfaces. The margin is the cross-context estimate minus the matched same-context pseudo-context null; the interval is a 90% unordered-context-pair bootstrap interval. These results do not rank D_{adj} as the universally best comparator.

Comparator	Cross	Same-context null	Margin [90% CI]	Recovery
Kendall distance	0.256	0.028	0.229 [0.216, 0.242]	1.00
Spearman distance	0.177	0.003	0.174 [0.157, 0.191]	1.00
Top-10 Jaccard distance	0.738	0.131	0.607 [0.594, 0.620]	1.00
Stable inversion	0.147	0.000	0.147 [0.134, 0.160]	1.00
D_{adj}	0.106	-0.005	0.111 [0.105, 0.116]	1.00

D.3 Transport-aware adaptive measurement allocation

We evaluated the practical measurement-design proposal on the held-out Frangieh risk grid, starting from 10 cells per perturbation and comparing uniform, random, lowest-cell-support-first, highest-pair-state-entropy and transport-variance allocation. Acquisition scores were frozen at the pilot stage; the full-depth vectors were used only as evaluation references. The transport-variance policy reduced mean absolute error by only 1.28% on average relative to uniform allocation, with gains

Supplementary Table S8: Transport-aware measurement allocation on the held-out Frangieh risk surface. Values average across the registered transfer–metric surfaces; the adaptive policy is promoted to the main text only if its error reduction and resolution gains are consistent rather than driven by one surface.

Policy	Abs. error	CI width	Detection	Resolution	Error reduction
highest pair state entropy	0.057	0.020	1.000	0.078	0.009
lowest cell support first	0.058	0.019	1.000	0.078	-0.013
random	0.056	0.019	1.000	0.078	0.025
transport variance	0.057	0.019	1.000	0.078	0.013
uniform	0.058	0.019	1.000	0.078	0.000

that varied by budget and surface. Random allocation happened to yield a larger average reduction (2.46%), but this is not an acquisition method and was not tuned or promoted. Resolution rates were identical across policies at each budget (0% through 40 cells, 5.6% at 80 cells and 33.3% at 160 cells). Thus the registered adaptive policy did not deliver a material or consistent cell saving and is retained as a negative practical-utility result rather than being moved into the main narrative. Supplementary Table S8 and the complete 450-row ledger at the accompanying Supplementary Data archive record the comparison.

D.4 Incremental decision value and empirical target-audit replay

To distinguish the pooled decision association from a mean-fidelity scale effect, we fit five prespecified ordinary least-squares models to normalized target regret under leave-one-unordered-context-pair-out validation. The prespecified model family is evaluated under this outer split; the nested contrasts are M1–M0, M2–M0 and M3–M1, while M4 is retained as an exploratory decomposition-feature model because it is not a superset of M3. All continuous features are standardized within the training fold. Each model includes an intercept and metric indicators; M4 replaces the adjusted total with its tie and directional components, alongside absolute mean-risk shift. The mean-shift model gives a modest OOF MAE improvement of 0.013 with an unordered-pair block-bootstrap 90% interval $[-0.001, 0.027]$. The D_{adj} contrast beyond metric plus mean shift is -0.005 ($[-0.014, 0.003]$). These are uncertainty-aware diagnostics, not causal effects. Table S9 gives the complete OOF comparison and 2,000-draw unordered-pair block bootstrap; no headline permutation p-value is reported for three blocks. The three unordered context pairs are the validation blocks, not the 18 transfer-by-metric rows. This limited block count restricts power and the precision of generalization claims; failure to improve held-out prediction does not establish that the adjusted discrepancy is universally uninformative.

Supplementary Table S9: Prespecified model comparison under leave-one-unordered-context-pair-out validation. ΔMAE is positive when a model improves on the stated reference; intervals are 2,000 unordered-pair block-bootstrap diagnostics. M4 is exploratory because its decomposition features are not a superset of M3.

Model	MAE	RMSE	OOF ρ	$\Delta\mathrm{MAE}_{M0}$	$\Delta\mathrm{MAE}_{M1}$	90% interval	Validation
M0 metric	0.100	0.130	0.60	0.000	–	–	ref.
M1 metric plus mean	0.087	0.119	0.66	0.013	0.000	[-0.001, 0.027]	LOPO (3)
M2 metric plus D_{adj}	0.105	0.132	0.64	-0.005	–	[-0.019, 0.008]	LOPO (3)
M3 metric plus mean plus D_{adj}	0.092	0.121	0.62	0.008	-0.005	[-0.011, 0.026]	LOPO (3)
M4 metric plus mean plus decomposition (exploratory)	0.106	0.137	0.66	-0.007	-0.019	[-0.040, 0.027]	LOPO (3)

The shortlist audit uses lower/upper target-risk intervals and the rules REUSE if $\mathrm{UB}_{\mathrm{regret}} \leq \epsilon$, REVISE if $\mathrm{LB}_{\mathrm{regret}} > \epsilon$, and UNRESOLVED otherwise. The offline replay uses $\epsilon = 0.05$, gives every candidate a 10-cell target pilot, and spends an additional target-cell budget through the 20/40/80/160 depth ladder. Policies see only the audit pool; the disjoint reference pool is retained offline and the evaluation pool is used only for realized regret. This is an empirical decision replay, not a certification framework or a simultaneous confidence guarantee. Table S10 summarizes all 360 rows; the row-level ledger is retained in the repository.

Supplementary Table S10: Empirical target-audit replay. Every candidate receives a 10-cell pilot; the policy allocates additional target cells using only the audit pool. Values average the four decision budgets.

Policy	Mean upgraded	False REUSE	Correct REUSE	Correct REVISE
pilot only	0.0	0.000	0.333	0.028
ptl boundary	55.7	0.097	0.444	0.056
random extra depth	32.6	0.000	0.333	0.069
source uncertainty	33.0	0.000	0.333	0.125
uniform depth	144.9	0.000	0.333	0.056

D.5 Regret-bound-critical active-audit redesign

As a final prespecified attempt to improve the active audit, we allocated additional cells to candidates active in the regret upper bound: the highest upper-bound items inside the source shortlist and the lowest lower-bound items outside it, followed by interval width. At matched actual additional-cell cost, false reuse was eliminated relative to the legacy PTL boundary, but the policy did not exceed the source-uncertainty baseline in correct revision. The realized evaluation-pool regret is invariant under the frozen-shortlist replay by construction and is reported only as a policy-invariant reference. The prespecified comparison therefore retains this policy as a negative supplementary result rather than a main-text contribution.

Supplementary Table S11: One-last active-audit redesign at matched actual additional-cell cost. The regret-bound-critical policy reduces false reuse relative to the legacy PTL boundary, but does not exceed the source-uncertainty baseline in correct revision at matched cost. Realized evaluation-pool regret is invariant under this frozen-shortlist replay by construction and is reported only as a policy-invariant reference. The prespecified decision rule therefore retains it as a negative supplementary result.

Policy	False reuse	Correct revise	True regret	Normalized regret
Regret-bound critical	0.000	0.083	0.040	0.378
Random extra depth	0.000	0.069	0.040	0.378
Uniform depth	0.000	0.056	0.040	0.378
Source uncertainty	0.000	0.125	0.040	0.378
Legacy PTL boundary	0.097	0.056	0.040	0.378

E Can reordering be anticipated from source-only information?

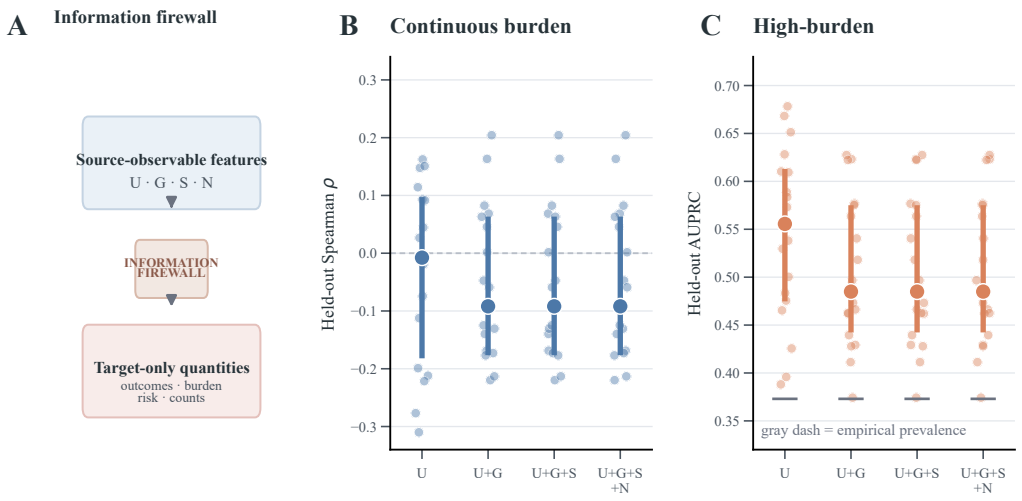
This analysis asks a narrower prospective question than the transport study: can a source-only feature ladder anticipate future reordering burden? The prespecified feature sets are cumulative uncertainty (U), geometry ($U+G$), support ($U+G+S$), and novelty ($U+G+S+N$). Target outcomes, target risk, and the measurement-adjusted transport value are unavailable to the predictor and excluded from the design matrix. The source-only task-specific usable universe contains approximately 28–44 labels; the 243-label universe belongs to the full-size shortlist surface.

The evaluation contains 18 transfer-by-metric tasks for each of continuous burden and high-burden classification. It uses ten repeats of five label-disjoint folds with nested ridge selection. Target-derived features are excluded from every source row. Supplementary Fig. S1 shows held-out task-level performance directly; it intentionally does not recycle the earlier deployment-gate analysis or an unreconstructible chance baseline. Across this declared feature ladder, held-out continuous correlations remain near zero and classification performance changes little, so the tested source-only descriptors do not reliably anticipate future reordering burden.

F Additional robustness and diagnostic analyses

Supplementary Fig. S2 checks whether the primary decision association is explained only by the different scales of the three metrics. Panel A shows the fixed 18-point surface; Panel B ranks D_{adj} and regret within each metric before pooling. The pooled within-metric association is descriptive

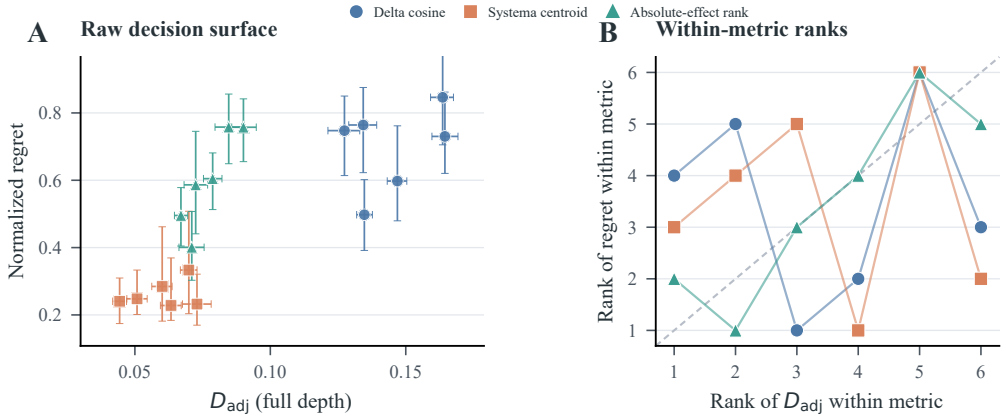
Supplementary Fig. S1 Source-only boundary



Supplementary Figure S1: Source-only prospective boundary. (A) The information boundary separates source-observable features from target outcomes reserved for evaluation. (B) Held-out Spearman correlation for continuous reordering burden. (C) Held-out AUPRC for high-burden classification. Each point is one transfer-by-metric task; thick marks summarize the median and interquartile range.

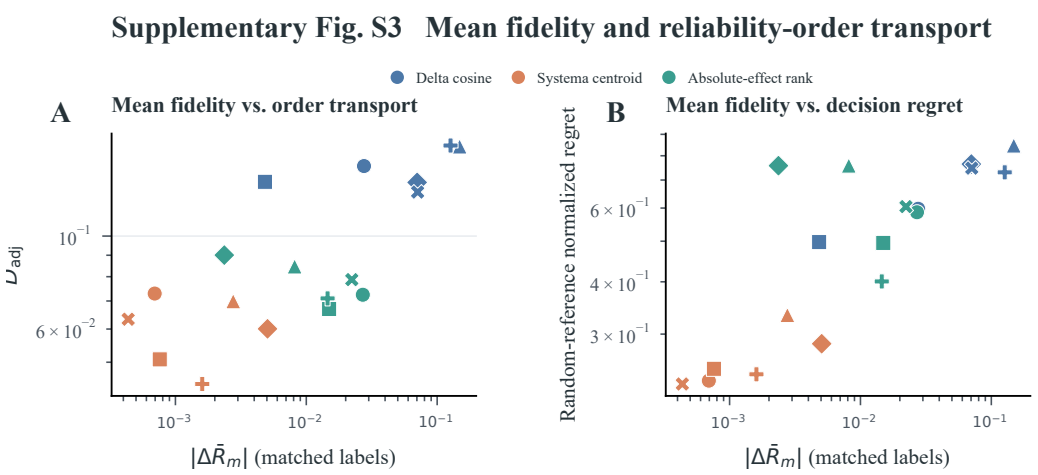
($\rho = 0.26$). With only three unordered context-pair blocks, the percentile 90% block-bootstrap interval is a diagnostic, $[0.02, 0.45]$ from 2,000 draws; leave-one-pair-out estimates range from 0.00 to 0.47. The six-point metric-specific summaries are shown without a headline significance claim.

Supplementary Fig. S2 Metric-stratified association



Supplementary Figure S2: Metric-stratified decision association. The 18 fixed full-depth, 10% random-reference decision rows are shown in their native scale (A) and after within-metric rank transformation (B). The pooled within-metric interval is reported in the text and source table rather than overlaid on the rank points. Colors identify metrics; no target-informed feature enters the source-frozen predictor.

Supplementary Fig. S3 computes the absolute change in mean metric risk on the exact matched perturbation universe used by D_{adj}. It is a metric-matched control for the statement that a change in average fidelity and a change in reliability ordering are distinct quantities; marker shape identifies directed transfer and color identifies the metric. Across the 18 rows, the mean-risk shift has Spearman $\rho = 0.69$ with D_{adj} and $\rho = 0.77$ with normalized regret; these are descriptive matched-universe associations, not a causal decomposition.



Supplementary Figure S3: Metric-matched mean-fidelity control. Each point is one of 18 directed transfer–metric rows. The horizontal coordinate is the absolute source-to-target mean-risk shift computed on the exact fixed label universe; the vertical coordinate is D_{adj} (A) or the corresponding normalized decision regret at 10% budget (B).

Supplementary Fig. S4 validates the finite-sample estimand in synthetic data with a known null and a known alternative. The stored trial-level distributions show the positive plug-in bias under the null and the centering of the U-corrected estimator near the population identity.

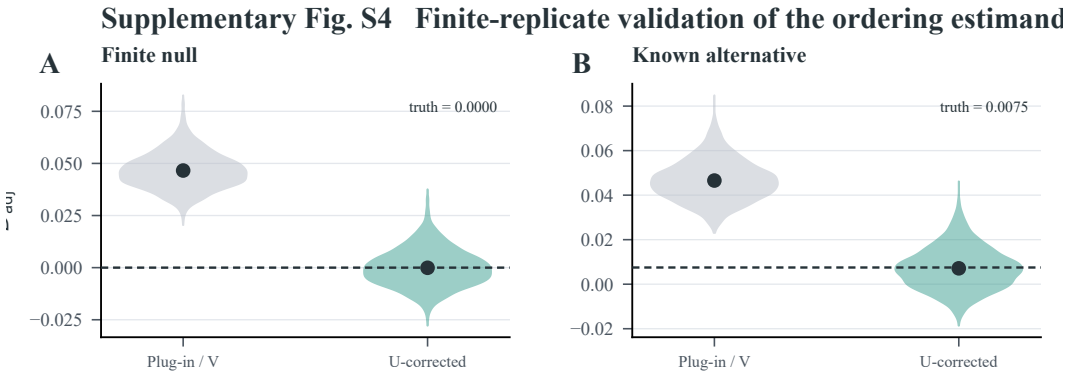
G STRONG COMPARATOR CALIBRATION

The finite-measurement claim was tested against the strongest registered traditional alternatives, not only against an uncorrected rank statistic. The benchmark contains two measurement laws, five depths, two resolution settings, two noise scales and three inversion fractions, for 120 conditions and 120 trials per condition. Every method uses the same perturbation-label resamples. The comparator set includes raw Kendall, Spearman and top-10 Jaccard, their label-bootstrap versions, matched-null-subtracted Kendall and Jaccard, stable inversion, and the U-corrected D_{adj}. The null target is zero for the measurement-adjusted quantities; the raw rank metrics are retained as descriptive sampling-disagreement comparators rather than being relabelled as transport estimands.

Raw Kendall, Spearman and Jaccard show high null false-transport rates because they compare two finite orderings without removing the same-context sampling component. Label bootstrap intervals quantify uncertainty around that comparison but do not change its target. Matched-null subtraction restores near-zero null false-transport for both Kendall and Jaccard and retains high power on the registered non-zero inversion surfaces. D_{adj} has the advantage of a single pair-state population interpretation and an explicit order-2 finite-replicate correction; the simulation does not support a claim that it universally dominates matched-null Kendall. Supplementary Table S12 reports false-transport rate, coverage and power side by side. The complete trial-level ledger is the accompanying Supplementary Data archive.

G.1 REAL-DATA SEMI-SYNTHETIC CALIBRATION

The known-truth simulation is complemented by a real-data semi-synthetic benchmark derived from the Frangieh raw-cell lineage. We used the canonical Control-context raw-cell pseudoreplicate risk ledger, selected the first 60 labels in the metadata-only eligible intersection, and generated two independent 12-replicate pseudo-contexts by resampling the observed replicate pool at 10, 20, 40, 80 and 160 cells. The same-context arm has a known transport null ($D_{\mathrm{adj}} = 0$). For the controlled alternatives, a single cyclic permutation of the first 5%, 10%, 20% or 40% of the frozen label panel



Supplementary Figure S4: Finite-sample validation of the ordering estimand. Trial-level sampling distributions compare the plug-in/V estimator with the order-2 U-corrected estimator under a finite-sample null and a known alternative; vertical references mark the population identity.

Supplementary Table S12: Strong comparator calibration in the registered 120-condition benchmark. Null false-transport and 90% coverage average over all zero-inversion conditions; alternative power and coverage average over the 0.1 and 0.4 inversion surfaces. MAE is to each method's own declared population target and is therefore not a universal ranking of estimators.

Method	Null FTR	Null cov.	Alt. power	Alt. cov.	Null MAE
D_{adj} (U-corrected)	0.017	0.921	1.000	1.000	0.001
Kendall	0.952	0.976	1.000	1.000	0.003
Kendall + label bootstrap	0.952	0.976	1.000	1.000	0.003
Kendall + matched null	0.003	0.965	0.997	1.000	0.003
Spearman	0.995	0.975	1.000	1.000	0.000
Top-10 Jaccard	1.000	0.783	1.000	0.987	0.029
Jaccard + label bootstrap	1.000	0.783	1.000	0.987	0.029
Jaccard + matched null	0.000	0.960	0.528	0.993	0.032
Stable inversion	0.000	0.500	0.996	0.999	0.000

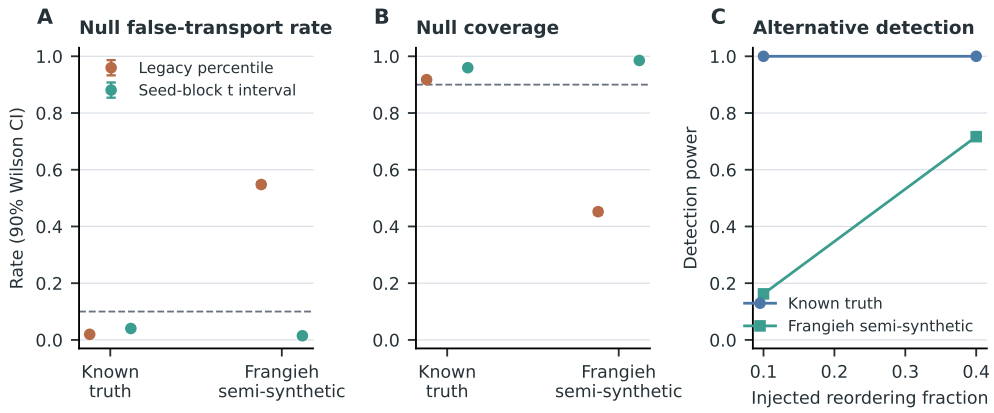
was applied to the target arm. This permutation was fixed before any estimates were inspected and is a response-order stress test, not a claim of a new biological replicate.

The original hierarchical percentile interval resampled the two arms independently. On this discrete ledger it produced a mean null false-transport rate of 0.548 and coverage of 0.452, exposing a nonregular boundary failure rather than a biological transport signal. We therefore replaced it with the inference specified in Sec. B: complete measurement-seed blocks are the independent units and the interval is based on the empirical variance of their U-corrected contributions.

The replacement was evaluated in 2,880 known-truth null trials and 3,000 Frangieh semi-synthetic null trials. The known-truth false-transport rate was 0.040 and null coverage was 0.960; the corresponding real-data semi-synthetic values were 0.015 and 0.985. Thus the replacement is mildly conservative, particularly on the discrete empirical ledger, rather than anti-conservative at the no-transport boundary. Detection power was 1.00 at the registered 10% and 40% simulated inversion fractions; in the harder Frangieh semi-synthetic benchmark it was 0.162 and 0.717, respectively. Supplementary Fig. S5 and Table S14 compare the failed interval and calibrated replacement. The old result remains in the audit ledger but is not used for any detectability conclusion.

Supplementary Table S13: Legacy real-data semi-synthetic comparator audit derived from the Frangieh raw-cell lineage. The same-context null resamples the Control-context canonical risk replicates; alternatives cyclically permute 5–40% of the 60-label panel (the horizontal levels denote the fraction of labels permuted, not the realized fraction of pair inversions). Values average over five cell budgets. The D_{adj} interval in this table is the failed hierarchical percentile procedure and is not used for any detectability conclusion; final seed-block inference is reported separately in Table S14.

Method	Null FTR	Null cov.	5% power	10% power	20% power	40% power
D_{adj} (legacy percentile interval; invalid)	0.548	0.452	0.854	0.909	0.980	0.999
Kendall	1.000	0.000	1.000	1.000	1.000	1.000
Kendall + matched null	0.038	0.961	0.150	0.254	0.500	0.771
Spearman	1.000	0.000	1.000	1.000	1.000	1.000
Top-10 Jaccard	1.000	0.132	1.000	1.000	1.000	1.000
Jaccard + matched null	0.006	0.994	0.009	0.011	0.025	0.024
Stable inversion	0.026	0.974	0.497	0.645	0.860	0.985



Supplementary Figure S5: Calibration of finite-measurement D_{adj} inference. (A) Null false-transport rate and (B) null coverage for the original percentile interval and the seed-block variance replacement; error bars are 90% Wilson intervals over trials and dashed lines mark nominal levels. (C) Detection power under controlled reordering alternatives.

H External predictor context under matched information constraints

To broaden the model-family context without changing the primary estimand, we also record the completed formal-v2 external GEARS contract (Roohani et al., 2024) on three public cohorts. These canonical runs use condition-disjoint holdout, zero perturbation overlap, a dataset-native frozen 4,096-gene panel, three seeds and 20 training epochs. The two larger cohorts include held-out validation predictions; the K562 GWPS cohort has only 64 test rows and is therefore retained as a test-only audit surface. Supplementary Table S15 reports the complete dataset-level summary. These results characterize conventional response prediction under external cohorts; they neither establish predictor-family invariance nor enter the Frangieh context-shift claim lock.

Supplementary Table S14: Calibration of the original percentile interval and the seed-block Student-t inference. Brackets are 90% Wilson intervals over null trials.

Benchmark	Inference	n	False transport	Null coverage
Frangieh semi-synthetic	Legacy percentile	7500	0.548 [0.538, 0.557]	0.452 [0.443, 0.462]
Frangieh semi-synthetic	Seed-block t interval	3000	0.015 [0.011, 0.019]	0.985 [0.981, 0.989]
Known truth	Legacy percentile	2880	0.020 [0.016, 0.025]	0.918 [0.909, 0.926]
Known truth	Seed-block t interval	2880	0.040 [0.035, 0.047]	0.960 [0.953, 0.965]

Supplementary Table S15: **Scope-limited external GEARS audit on public cohorts.** Each row aggregates three CUDA runs (20 epochs; seeds 0–2) under a random-condition-holdout split with zero perturbation overlap and a dataset-native frozen 4,096-gene panel. Cosine similarity is reported on non-control test cells (higher is better); MSE and MAE use the same cells (lower is better). "Validation" counts runs with held-out validation predictions. These conventional response-prediction results are not pooled with the Frangieh source-to-target reliability-transport estimand.

Dataset	Runs	Test cells	Cosine	MSE	MAE	Validation
Norman filtered	3	1,069	0.159 ± 0.058	0.01258	0.07552	3/3
Replogle RPE1	3	983	0.136 ± 0.018	0.02245	0.11218	3/3
Replogle K562 GWPS	3	64	0.069 ± 0.021	0.02716	0.12828	0/3[†]

[†]Test-only external audit surface; no validation vector was available.

We separately executed the stricter modern-predictor bridge required for the Frangieh claim lock. The official GEARS implementation was trained for 20 epochs in five label-disjoint folds independently within each source context. Graph construction, validation and fitting used source cells only. Thirteen labels absent from the official interaction graph were removed before splitting, leaving 230 perturbations and 8,229 evaluation genes. The frozen out-of-fold vector was then passed through the same 30-seed, two-pseudoreplicate raw-cell bridge as the other predictor families, producing the complete 18-row directed transfer–metric surface. TxPert was treated under the same information gate: the available checkpoint was trained on K562 rather than a source-matched Frangieh context, and using target basal expression would violate the frozen-source contract. It is therefore not included in the predictor-family concordance. These are deliberate contract outcomes, not missing values imputed from a different experiment; the machine-readable record is the accompanying predictor-contract summary.

An earlier one-epoch Frangieh GEARS pilot is retained only as engineering provenance and is not pooled as an additional replicate. The public-cohort GEARS audits above answer a different response-prediction question and are likewise not pooled with the strict Frangieh transport surface.

Using the same frozen panels and condition-disjoint split construction, we also completed five non-GEARS controls (45 runs total): control mean, global perturbation delta, perturbation mean delta, context-nearest-neighbour delta, and ridge regression. Together with GEARS this gives 54 comparable runs across six audited model/control families. Supplementary Table S16 reports equal-run aggregate metrics; the complete run-level records are retained so that the small-cohort sensitivity and model-specific variation remain auditable. Because the holdout is perturbation-disjoint, the context-kNN control falls back to the global mean on unseen test perturbations; we retain this row as an explicit protocol control rather than presenting it as successful neighbour transfer. These controls broaden the external benchmark context without changing the source-frozen Frangieh estimand or implying a legal second predictor family on that primary surface.

Supplementary Table S16: **Comparable external model-family audit on public cohorts.** Each entry is an equal-weight mean across nine runs (three cohorts × three seeds) on the corresponding cohort-specific frozen 4,096-gene panel and condition-disjoint holdout construction. Per-seed test-cell counts match across the audited model outputs. Cosine is higher-is-better; MSE and MAE are lower-is-better. The GEARS rows are the verified CUDA runs in Table S15; the remaining rows are deterministic or fitted baselines evaluated with the same split construction. These summaries are conventional response-prediction controls and are not pooled with the Frangieh reliability-transport claim.

Model family	Runs	Cohorts	Cosine	MSE	MAE
Ridge regression	9	3	0.233	0.01887	0.09970
Perturbation mean Δ	9	3	0.202	0.01705	0.09548
Global mean Δ	9	3	0.185	0.01737	0.09624
Context kNN Δ (fallback)	9	3	0.185	0.01737	0.09624
GEARS	9	3	0.122	0.02073	0.10533
Control mean	9	3	0.000	0.01823	0.09852

Equal-run means are not test-cell-weighted; the K562 GWPS cohort contributes only 64 non-control test cells per three-run cohort. Under the perturbation-disjoint holdout, the kNN control falls back to the global mean for unseen test perturbations; it is retained as a protocol control, not as evidence of context-neighbour transfer. Full run-level values are reproduced in the accompanying Supplementary Data archive.

The broader predictor-family plan was also audited against the same source-only contract. Supplementary Table S17 records the executable or semantic status of the additional families, including scGPT (Cui et al., 2024), TxPert (Wenkel et al., 2026), CPA (Lotfollahi et al., 2023), scGen (Lotfollahi et al., 2019) and CellOT (Bunne et al., 2023). These rows are deliberately not treated as performance comparisons: a package, checkpoint, or context-mismatched inference run is not evidence for the Frangieh transport claim. The corresponding machine-readable ledgers are the predictor-family protocol and feasibility summaries included in Supplementary Data.

Supplementary Table S17: **Predictor-family compatibility audit.** The audit records whether a public implementation can satisfy the manuscript's frozen source-only prediction contract: training uses source information only, target outcomes are unseen until evaluation, and the output is a held-out vector on the declared gene panel. Only the canonical formal-v2 GEARS cohort runs enter Supplementary Tables S15–S16; the other rows are feasibility or semantic-contract records, not performance results.

Family	Runnable	Contract	Context	Disposition
GEARS	yes	partial	external	Scope-limited cohort audit; not pooled with the Frangieh estimand.
State	yes	no	mismatched	Set-transition output lacks the required global held-out vector.
TxPert	yes	partial	K562 only	Reproducible inference, but no matched HepG2–Jurkat transfer.
scGPT	partial	no	mismatched	No prespecified held-out vector on the frozen Frangieh panel.
biolord	no	pending	—	Compatible source-only adapter and checkpoint unavailable.
PerturbNet	no	pending	—	Compatible source-only adapter and checkpoint unavailable.
Scouter	no	pending	—	Compatible source-only adapter and checkpoint unavailable.
CPA-ext / scGen / CellOT	partial	no	rescue	Information contract differs; no transport score claimed.

H.1 Source-only predictor-family sensitivity on the Frangieh surface

To test whether the transport profile is tied to the primary bilinear-ridge parameterization, we evaluated a source-only RBF kernel-ridge family on the same Frangieh labels, gene panel, three metrics,

Supplementary Table S19: **Canonical source-only predictor-family transport.** Each family is evaluated on the same six directed transfers, three metrics and full-size raw-cell measurement surface. The correlation and sign columns are relative to the bilinear-ridge reference and are descriptive; they do not establish family equivalence or superiority.

Family	Mean ρ	Min point sign	Min CI sign	Mean D_{adj}	Δ mean	Rows
Bilinear ridge	1.000	1.000	1.000	0.073	0.000	6
GEARS (strict source-frozen)	0.371	1.000	1.000	0.119	0.047	6
RBF kernel ridge	0.905	1.000	1.000	0.074	0.002	6
Latent-response MLP	0.886	1.000	1.000	0.093	0.020	6
Direct MLP	0.829	1.000	1.000	0.092	0.020	6

six directed transfers and full-size minimum-20 measurement surface. The target outcomes were loaded only after source-only prediction materialization; they were not visible during fit, penalty tuning or family selection. The initial penalty grid was extended because all 15 source-by-outer-fold choices selected its largest value. The resulting RBF sensitivity preserves the positive adjusted-discrepancy interval on all 18 matched rows and has descriptive D_{adj} profile Spearman $\rho = 0.983$ with the bilinear-ridge family; the within-metric profile range is 0.829–1.000 and the mean row-wise prediction-vector cosine is 0.733. Point and interval sign agreement holds on all 18 rows. This is a matched-surface sensitivity, not evidence of predictor-family equivalence, superiority, causality or broader generalization.

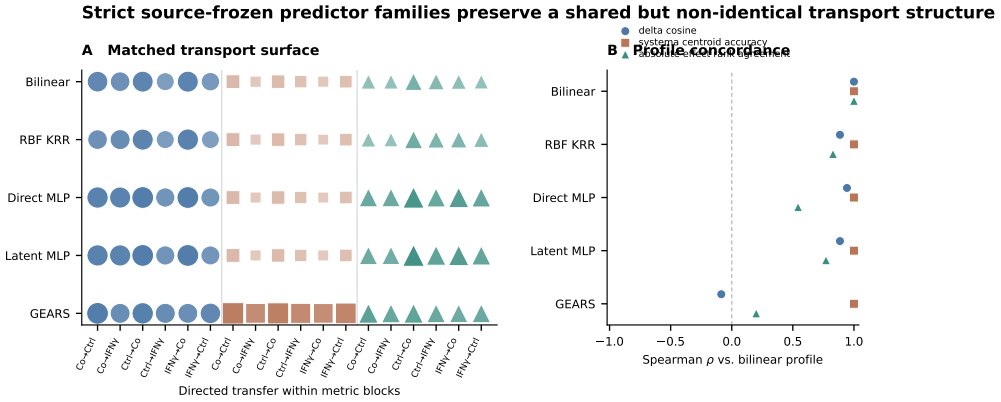
Supplementary Table S18: **Source-only predictor-family sensitivity on the matched Frangieh surface.** The RBF kernel-ridge family was tuned using source-only inner cross-validation; because all 15 initial source-by-outer-fold choices hit the initial largest penalty, the grid was extended to $\{30, 100, 300, 1000\}$. The comparison uses the same full-size minimum-20 label universes, frozen metrics, transfer directions and measurement-floor estimator. Source OOF fidelity is shown as the macro mean for delta cosine, Systema centroid accuracy and absolute-effect rank agreement, respectively. Concordance is descriptive and does not establish predictor-family equivalence or superiority. For the RBF family, the matched profile concordance is $\rho = 0.983$, with within-metric range 0.829–1.000 and mean row-wise prediction-vector cosine 0.733.

Family	Source-only tuning	Surfaces	Source OOF fidelity (Δ / Sys / rank)	Matched transport concordance
Bilinear ridge	Registered source-inner CV	18	0.240 / 0.499 / 0.299	Reference family
RBF kernel ridge	15/15 hit $\alpha = 10$; extended to $\alpha = 30$	18	0.257 / 0.498 / 0.310	D_{adj} profile $\rho = 0.983$; point and CI sign agreement 18/18

H.2 Neural source-only families and finite-measurement simulation

We extended the source-only sensitivity without changing the canonical bilinear-ridge estimand. Two additional families were fit on the frozen Frangieh surface: a direct MLP and a latent-response MLP. In every outer fold, the ten-dimensional input basis and, for the latent family, the 32-dimensional response basis were fit on source training conditions only. The models used the prespecified 256–512 GELU architecture, dropout 0.1, AdamW and a source-only inner selection of weight decay from $\{10^{-5}, 10^{-4}, 10^{-3}\}$. Across the three source contexts, the macro source-side OOF fidelity was 0.363 for the direct MLP and 0.366 for the latent-response MLP, compared with 0.346 for the bilinear reference and 0.355 for RBF. These are adequacy and profile diagnostics, not target- informed model selection; the target outcomes were not loaded during fitting.

The resulting predictor-family profiles, including strict source-frozen GEARS, use the same contexts, metrics and raw-cell bridge. Bilinear, RBF, direct-MLP and latent-MLP use the 243-label eligible panel; GEARS uses the verified 230-label intersection supported by its official interaction graph. Each family contributes 18 directed transfer–metric rows on its declared panel; it is a sensitivity analysis, not a predictor-superiority claim. Supplementary Fig. S7 shows this profile together with the corrected finite-measurement calibration: two generative laws, 960 registered conditions (six depths, four resolution settings, four noise scales and five inversion settings), and 500 trials per



Supplementary Figure S6: **Matched source-frozen predictor-family transport.** (A) Circle, square and triangle glyphs show D_{adj} for delta cosine, Systema centroid and absolute-effect rank; area and opacity encode magnitude. Every family uses the same six directed transfers and raw-cell bridge. (B) Within-metric profile Spearman correlation with the prespecified bilinear reference. Positive sign agreement does not imply identical transport geometry or predictor superiority.

Supplementary Table S20: **Corrected finite-measurement calibration.** Each primary condition uses a finite-depth pair-state population as the truth for its matching estimator. Values summarize the condition-level absolute error, signed bias, RMSE, and median absolute error over the registered primary grid; the grid contains 960 conditions and 500 trials per condition. The realized pair-tie rate is reported rather than equated with the resolution setting.

Estimator	Mean MAE	Mean bias	Mean RMSE	Median MAE
D_{adj}	0.002	-0.000	0.002	0.001
Kendall distance	0.011	-0.006	0.012	0.005
Spearman distance	0.002	0.001	0.003	0.002
Top-10% Jaccard distance	0.017	0.006	0.026	0.013
Stable-pair inversion	0.010	-0.010	0.010	0.004

Mean realized pair-tie rate across conditions: 0.063. The corrected D_{adj} truth is not a latent Kendall distance; the previous v2 artifact is retained only for provenance and is not pooled here.

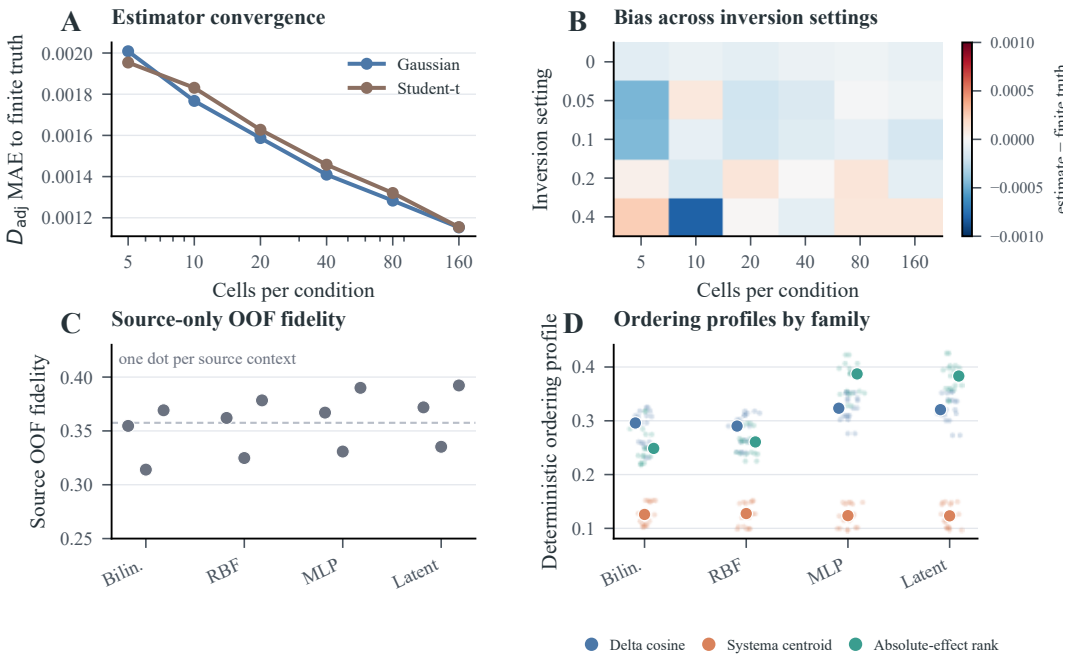
condition. Here the D_{adj} truth is the finite-depth population pair-state distance at the declared resolution, while Kendall, Spearman and top-10 Jaccard use their corresponding finite-depth population-mean ordering targets. Directional states use the exact Beta posterior with the frozen minimum strict support of eight; realized pair ties are retained as a third state. The simulation is a calibration and recovery screen; it does not justify a universal superiority claim for D_{adj} or for any predictor family. Full machine-readable outputs are the four-family transport, neural-profile, and finite-measurement tables in Supplementary Data. The earlier simulation series is retained only as provenance and is not pooled with the corrected grid.

H.3 Outcome-blind Feng multi-donor biological replication audit

We audited the public multi-donor human-pluripotent-stem-cell CRISPRi screen before inspecting any transport outcome (Feng et al., 2026). Eligibility was frozen from metadata support: 19 lines from 10 donors, with 135 targets supported in every line at the primary threshold of 20 cells. The response file is an aggregate target-by-line LFC table rather than raw cell-level outcomes. We therefore froze an outcome-blind 1,000-gene panel from the file schema and computed target-level RMS response risk, then compared target rankings across same-donor/different-line and different-donor pairs.

The audit is informative but not a positive replication of the full Frangieh D_{adj} hierarchy. Same-donor/different-line pairs had median Kendall distance 0.220, whereas different-donor pairs had 0.227; the corresponding Spearman distances were 0.124 and 0.131, and top-10 Jaccard distances

Supplementary Fig. S7 Calibration and source-only family sensitivity



Supplementary Figure S7: **Source-only family and finite-measurement sensitivity.** (A) Mean absolute error of the corrected ordering estimate relative to the finite- depth pair-state truth across cell depth for Gaussian-heteroscedastic and Student-t measurements. (B) Bias surface for the Gaussian primary slice; the horizontal axis is cell depth and the vertical axis is the true inversion fraction. (C) Source-side OOF fidelity for each source context and predictor family. (D) Deterministic ordering profiles for the three fixed metrics across the four source-only families. The neural family and simulation analyses are descriptive and were not used to tune target transport or alter the canonical main-text estimand.

Supplementary Table S21: Outcome-blind Feng multi-donor iPSC audit. Risks are RMS response magnitudes over the frozen 1,000-gene panel, so this is a biological-replication audit rather than a full raw-cell D_{adj} hierarchy. The same-donor and different-donor rank distances are similar; no cross-donor excess is promoted.

Comparison	Pairs	Targets	Kendall d	Spearman d	Top-10 Jaccard d
Different donor	162	135	0.227	0.131	0.700
Same donor, different line	9	135	0.220	0.124	0.818

were 0.818 and 0.700. The contrasts are small and metric-dependent: cross-donor distance was slightly higher for Kendall and Spearman but lower for top-10 Jaccard. We retain this as a neutral aggregate rank audit, not as a negative biological claim or a predictor-risk replication. A within-line cell resampling floor and hierarchical cell-level bootstrap could not be computed from the released aggregate LFC file; those analyses remain an explicit reopening requirement rather than an imputed result. Supplementary Table S21 and the accompanying Feng pair-level Supplementary Data table contain the complete pair ledger.

H.4 Provenance limits on a four-context expansion

We audited the exact provenance and support before adding a broader HepG2–Jurkat–K562–RPE1 transport map. HepG2 and Jurkat are both day-7 harvests, but the available Replogle K562 and RPE1

exports are day 6 and day 7, respectively, with no timepoint field in the local cell-level exports. The harvest difference is therefore inseparable from the intended context contrast. Although the four-way label intersection and minimum-support counts are large, the current 1,536-gene panel is missing 185 genes in at least one context. We consequently do not present a clean four-context biological map; the audit is a provenance boundary, not a negative biological result. The underlying public-cohort provenance follows the Nadig atlas and the Replogle genome-scale Perturb-seq resource (Nadig et al., 2025; Replogle et al., 2022).

Supplementary Table S22: Provenance audit for a possible four-context expansion. The HepG2–Jurkat pair is a supported independent context comparison, but the available K562 and RPE1 exports differ in harvest day and cannot support a clean four-context biological map under the current contract.

Criterion	Result	Consequence
HepG2–Jurkat harvest	Both day 7	Supported independent pair
K562–RPE1 harvest	Day 6 vs day 7	Context–time confounded
Four-way shared labels	2,053	Label support is adequate
Minimum-20 support	1,688 labels	Support is adequate for profiling
Current 1,536-gene panel	1,351 present in all four	Requires a newly frozen shared-gene subset
Four-context map	Blocked	No clean biological claim

H.5 Outcome-blind scPerturb3 context audit

As a separate provenance audit, we examined the local Srivatsan–Trapnell scPerturb3 file using observation and variable metadata only. The three candidate cell-line contexts (A549, MCF7 and K562) share 189 perturbations and 753 matched dose–time units per pair, with a median of at least 213 cells per matched unit. However, all 52 plates are unique to one cell line, so every cell is on a context-unique plate. This fails the prespecified no-unique-plate-confounding requirement. The source-only predictor was therefore not run on these contexts, and this audit is reported as a provenance boundary rather than as a negative biological result. The complete outcome-blind audit and the evidence required to reopen this analysis are reproduced in the Supplementary Data archive.

I Reproducibility and scientific regeneration

The scientific regeneration path starts from the canonical analysis tables, fixed splits and resampling settings, and rebuilds the four main figures, Supplementary Figs. S1–S7, and the compact supplementary tables. Comparator, Feng and adaptive-allocation records are summarized above and do not alter the canonical main figures. The anonymous release documents executable commands, software versions, public accession records and data lineage. Large raw single-cell matrices remain external; compact Supplementary Data tables expose the values used in figures and tables.

Code and data availability. The anonymized core release is available at `https://anonymous.4open.science/r/PTL-B377/` and contains the source code, analysis configurations, manuscript sources, publication figures, and compact reviewable data products. Raw single-cell matrices are not redistributed. The Frangieh RNA screen and the Nadig HepG2–Jurkat screens are available from the public scPerturb v1.4 Zenodo record (Zenodo); the anonymous release documents the relevant public accessions and regeneration commands.

The scope-limited external GEARS audit and its matched baseline controls are summarized in Supplementary Tables S15 and S16; the broader predictor-family compatibility ledger is summarized in Table S17. Broader condition-holdout analyses, old failure-anatomy graphics, and the earlier broad source-only deployment gate are outside the evidentiary scope of this submission. Predictor families that do not satisfy the frozen source-only contract are listed in the compatibility table but are not presented as primary scientific evidence.