WHEN TRUST RANKINGS GENERALIZE: RELIABILITY AND INVARIANCE IN PERTURBATION TRANSFER

Anonymous authors
Paper under double-blind review

ABSTRACT

Perturbation-response models are usually evaluated for average predictive accuracy, but deploying them across biological contexts also requires knowing when a prediction is likely to fail. We show that two pre-response signals – baseline atypicality and intervention-response geometric coherence – carry information about relative transfer error before a target context's response is observed. Under a strict, leakage-safe evaluation that holds out entire contexts and recomputes every feature per fold, a preregistered trust ranking combining these signals improves selective prediction over random and magnitude-only baselines for normalized and directional error in two independent datasets, Tahoe-100M and MIX-Seq; raw and intervention-scaled error do not satisfy the same criterion, and a third dataset, sci-Plex3, is structurally non-evaluable under this protocol. We then ask whether a related but separately operationalized context-level transfer-difficulty statistic behaves as an invariant property: it is metric- and estimator-dependent, a substantial share of its structure is associated with technical and quality-control covariates, and a permuted-baseline control satisfies the same internal-reliability criterion as the real estimator. Predictive structure should therefore not be read as invariant or intrinsic biological structure. The trust ranking supports retrospective ranking and experimental triage, not calibrated uncertainty or causal inference.

1 INTRODUCTION

Models of cellular perturbation response are usually evaluated for average predictive accuracy on held-out interventions, combinations, or contexts (Section 2). Average accuracy, however, does not answer a different, practical question: for one specific, not-yet-observed context, should a particular predicted transfer be trusted? A deployed pipeline must decide, before the true response is known, which predictions to act on and which to defer, validate, or measure directly.

We treat this as a prediction problem in its own right: can transfer reliability itself be predicted before a target context's response is observed, under held-out biological *contexts* rather than held-out samples of an already-seen context? This differs from a model's internal uncertainty (Section 2): it is structure in the transfer itself, not in any model's confidence. We evaluate this under a leakage-safe protocol that holds out entire contexts, recomputes features per fold, and reveals the true response only after a prediction is committed.

Two natural candidate signals follow from this framing: consistency in how an intervention's effect is expressed across contexts, and typicality of a context's own baseline state. Two claims recur and must stay separate: that a signal predicts which transfers succeed (*predictive structure*) does not establish that it is stable under a different metric or estimator (*invariant structure*). We evaluate a fixed, non-learned mean-transfer reference, not a learned architecture — a generalization and invariance problem, not a representation-learning one (Ahlmann-Eltze et al., 2025).

We first establish, conservatively, what pre-response structure exists. Two signals computed without the held-out context's perturbation response — the atypicality of a target context's baseline profile, and the directional agreement of an intervention's response across other observed contexts — are associated with relative transfer error in Tahoe-100M (Zhang et al., 2025) and MIX-Seq (McFarland et al., 2020). Combined into a single, preregistered trust ranking (Section 5), this survives a strict, leakage-safe context-holdout test for normalized and directional error, but not raw or intervention-scaled error; a third dataset, sci-Plex3 (Srivatsan et al., 2020), lacks enough contexts to be evaluated

under this holdout design at all — a structural limit of the design, not a failed replication. Effect sizes throughout are modest, and the intended use is ranking transfers for triage, not calibrated absolute-error prediction.

A positive predictive result invites a further, easy-to-conflate claim: that a context's transfer difficulty is itself stable. We test this, not assume it, with a targeted falsification analysis (Figure 1, bottom row) — the natural next question once a predictive result exists, not a second study — on a context-level difficulty statistic defined independently of the two predictors above, so it neither confirms nor refutes their result. The result is a caution: normalized and cosine-based difficulty rankings do not survive an estimator swap, much of what does survive is associated with technical covariates rather than confirmed as biological, and a permuted-baseline estimator matches the real one on the same internal-reliability criterion — so internal reliability alone is not sufficient evidence of real signal.

Taken together, these results support a single organizing claim: predictive structure for perturbation-transfer reliability exists and replicates under a strict test, but it should not be promoted, implicitly or explicitly, into a claim of estimator-invariant or biologically intrinsic structure. Predictive structure need not be invariant structure — a distinction we draw across two related but non-identical constructs (Section 3), not by testing one statistic's invariance twice. We treat this as the paper's central contribution rather than as a limitation appended at the end.

Our contributions are threefold.

1. We formulate perturbation-transfer reliability as a leakage-safe, pre-response prediction problem and provide preregistered evidence, replicated across two datasets, that a trust ranking built from two pre-response signals predicts relative transfer error under strict context holdout — reported together with the exact metrics, dataset, and holdout regime under which this does not hold.
2. We conduct a falsification-oriented robustness analysis of a separately operationalized context-level difficulty statistic, showing that it fails an estimator swap for the same two metrics that carry the predictive result above, is substantially associated with technical measurement covariates, and can be matched in internal-reliability terms by a permuted negative control.
3. We introduce, and use throughout, an explicit distinction between predictive, reproducible, estimator-invariant, and biologically intrinsic structure, showing how two related but non-identical empirical constructs can be evaluated together without collapsing into a single overclaimed notion of context difficulty.

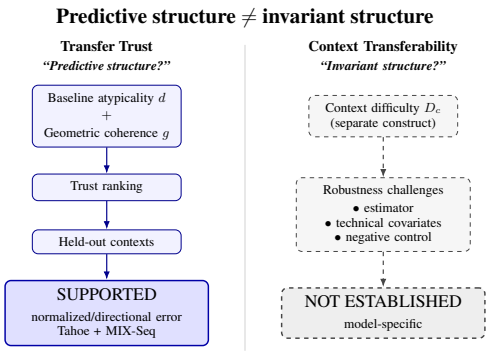


Figure 1: Conceptual overview: two independent columns, not a shared pipeline ($d = d_{c,-c}$, $g = g_{i,-c}$, Section 3.1; $D_c = D_c^{(m,S)}$, Section 3.2). Left: Transfer Trust's predictors feed a trust ranking tested under strict held-out contexts (Sections 4– 5). Right: Context Transferability's D_c is audited separately (Section 6) against an estimator swap, a technical-covariate adjustment, and a permuted negative control, supporting a model-specific rather than estimator-invariant interpretation. No arrow or shared box connects the columns: D_c is never computed from, or tested against, d or g (Table 1).

2 Related Work

Perturbation-response prediction. Models predicting a cell population's response to a perturbation are typically trained and evaluated on held-out interventions, combinations, or contexts within one benchmark (Lotfollahi et al., 2019; 2023; Roohani et al., 2024; Hetzel et al., 2022; Bunne et al., 2023), including Tahoe-100M (Zhang et al., 2025), MIX-Seq (McFarland et al., 2020), and sci-Plex3 (Srivatsan et al., 2020). Our question differs from response-prediction accuracy: whether a specific predicted transfer should be trusted before its true response is observed. Consistent with evidence that learned perturbation-response models do not yet reliably outperform simple baselines (Ahlmann-Eltze et al., 2025), we fix a simple, non-learned mean-transfer reference and study reliability in isolation from response-prediction accuracy.

Reliability under distribution shift. Model confidence degrades under distribution shift (Ovadia et al., 2019), motivating uncertainty methods internal to a fitted model. We instead ask whether transfer reliability is predictable *before* a target context's response exists, under held-out-context generalization rather than i.i.d. held-out samples.

Representation geometry and invariance. A separate line of work asks whether representations learned by different models converge toward a shared geometry (Huh et al., 2024). We do not test representation convergence or the Platonic Representation Hypothesis: geometric coherence is a specific, preregistered summary statistic of response directions, not a learned embedding. Our question is narrower: whether a related but separately operationalized context-level construct, motivated by the same idea of predictive usefulness, is thereby estimator-invariant — and we find that it is not (Section 6).

3 Problem Formulation

This section fixes notation and states, up front, why two similarly-named constructs must not be treated as the same quantity. Every definition below reproduces, without modification, the frozen definition used in the corresponding analysis.

3.1 Pair-level prediction of transfer error

Let i index interventions and c index biological contexts, and let $\boldsymbol{y}_{ic}$ be the perturbation response (a pseudobulk delta relative to the matched unperturbed baseline). The fixed mean-transfer reference and realized transfer error are

$$\hat{\boldsymbol{y}}_{ic} = \frac{1}{|\mathcal{C}_{i,-c}|} \sum_{c' \in \mathcal{C}_{i,-c}} \boldsymbol{y}_{ic'}, \qquad e_{ic} = \frac{\|\boldsymbol{y}_{ic} - \hat{\boldsymbol{y}}_{ic}\|_2}{\|\boldsymbol{y}_{ic}\|_2 + 10^{-8}}, \tag{1}$$

where $\mathcal{C}_{i,-c}$ is the set of contexts, other than c, in which i's response is observed; raw, intervention-scaled, and cosine variants replace only the numerator/normalization (Section 4). Computing e_{ic} requires $\boldsymbol{y}_{ic}$, revealed only to score a prediction: e_{ic} **is the prediction target**, not a predictor.

Two signals predict e_{ic} without observing $\boldsymbol{y}_{ic}$, but differ in what else they require. *Baseline atypicality* uses only the mean control profile $\boldsymbol{b}_c$,

$$d_{c,-c} = 1 - \cos\Big(\boldsymbol{b}_c,\ \tfrac{1}{|\mathcal{C}_{-c}|} \textstyle\sum_{c' \neq c} \boldsymbol{b}_{c'}\Big), \tag{2}$$

and requires no perturbation response anywhere. *Geometric coherence*,

$$g_{i,-c} = \Big\| \tfrac{1}{|\mathcal{C}_{i,-c}|} \textstyle\sum_{c' \in \mathcal{C}_{i,-c}} \boldsymbol{y}_{ic'} / \|\boldsymbol{y}_{ic'}\|_2 \Big\|_2, \tag{3}$$

is narrower: it requires i's response to already be observed in some other context, excluding only the held-out target's own response $\boldsymbol{y}_{ic}$. Response magnitude $m_{i,-c}$ is a nuisance covariate wherever $d_{c,-c}$ or $g_{i,-c}$ is related to e_{ic}; reported associations are magnitude-adjusted partial correlations unless stated otherwise.

Table 1: The constructs defined in this section, and what each does and does not use.

Construct	Level	Uses target context's own response?	Role
Transfer error e_{ic}	pair (i, c)	Yes — required to compute it; revealed only after a prediction is committed	Transfer Trust prediction *target*
Baseline atypicality $d_{c,-c}$	context c	No — uses only control profiles $\boldsymbol{b}_c$	Transfer Trust *predictor*
Geometric coherence $g_{i,-c}$	pair (i, c)	No for context c; yes for intervention i's response in other observed contexts	Transfer Trust *predictor*
Context-level difficulty $D_c^{(m,S)}$	context c, aggregated over S	Yes, for every intervention averaged into it	Context Transferability's own *audited statistic* — not a Transfer Trust predictor, not the Transfer Trust prediction target

3.2 A SEPARATELY OPERATIONALIZED CONTEXT-LEVEL STATISTIC

A second construct, defined independently of the above, is not a predictor of anything but a statistic whose own reliability is audited across estimators, metrics, and technical covariates (Section 6). Let $\Delta_{ic} \in \mathbb{R}^G$ be a treated-minus-control response vector and T_{ic} the other contexts observed for i. A leave-one-context-out reference $\widehat{\Delta}_{ic} = |T_{ic}|^{-1} \sum_{c' \in T_{ic}} \Delta_{ic'}$ gives four per-pair errors, denoted $\check{e}_{ic}^{(m)}$ for $m \in \{\mathrm{norm}, \mathrm{raw}, \mathrm{scaled}, \cos\}$ (the check accent avoids confusion with e_{ic} above: the two source analyses independently name their own per-pair error identically). Context-level difficulty averages these over an eligible intervention subset S_c:

$$D_c^{(m,S)} = |S_c|^{-1} \sum_{i \in S_c} \check{e}_{ic}^{(m)}. \tag{4}$$

Its own source manuscript states explicitly that $D_c^{(m,S)}$ is *not a latent context attribute*.

3.3 BASELINE ATYPICALITY IS NOT D_c

$d_{c,-c}$ and $D_c^{(m,S)}$ sound alike but are not the same quantity (Table 1). $d_{c,-c}$ uses only control profiles and is a *predictor* evaluated against a separately defined outcome; $D_c^{(m,S)}$ requires the target's own response and is itself the *audited object*, not a predictor. They also belong to separate frozen analyses: $d_{c,-c}$, $g_{i,-c}$, and e_{ic} to Sections 4–5; $D_c^{(m,S)}$ to Section 6, which never computes or uses the other three. Section 6 therefore neither validates nor invalidates $d_{c,-c}$ or $g_{i,-c}$: it tests whether the broader idea of context-level transfer difficulty that motivates $D_c^{(m,S)}$ survives a stronger evidentiary standard than either predictor has been asked to meet. The two constructs share a question, not a statistic.

We also distinguish, and use throughout, four non-interchangeable levels of evidence, each requiring more than the last: *predictive* (a signal ranks held-out error), *reproducible* (stable across resampling under one fixed estimator), *estimator-invariant* (ranking survives an estimator change), and *biologically intrinsic* (stable independent of measurement, estimator, and technical covariates) structure — none is implied by the last (full taxonomy in Appendix A.1). All associations reported in this paper are correlational: no analysis here is designed to identify a causal mechanism, and no claim in later sections should be read as one.

4 PREDICTIVE STRUCTURE OF PERTURBATION TRANSFER

We first ask the narrowest question: is transfer error predictably structured at all, using only information available before the held-out context's response is observed? We answer it here with plain associations; Section 5 then asks whether the same associations survive a strict, leakage-safe holdout of entire contexts.

4.1 BASELINE ATYPICALITY

Baseline atypicality $d_{c,-c}$ is positively associated with transfer error, individually significant, in both datasets with enough contexts for independent inference: Tahoe-100M ($r = +0.312$, 95% CI $[0.202, 0.414]$) and MIX-Seq ($r = +0.425$, 95% CI $[0.316, 0.532]$; full predictive-structure values in Appendix A.2). This correlation is not itself magnitude-adjusted; magnitude and control-profile-scale robustness checks (Appendix A.8) did not remove the association in either dataset. sci-Plex3 has only three contexts and does not support an independent context-level estimate for $d_{c,-c}$ at all. We refer to this signal by its measured name throughout: it is a Transfer Trust predictor (Table 1), not an intrinsic property of context.

4.2 GEOMETRIC COHERENCE

Geometric coherence $g_{i,-c}$, a magnitude-adjusted partial Spearman correlation with error, is more heterogeneous across datasets (Figure 2, per-dataset intervals). Random-effects pooling across all three excludes zero (pooled support, $I^2 = 0\%$), but the critical sensitivity check — pooling the two replication datasets alone, excluding Tahoe — crosses zero (dashed interval in the figure): a modest, directionally reproducible signal whose pooled support currently depends on including the discovery dataset, not an association confirmed by either replication dataset alone. We do not describe geometric coherence as replicating significantly across datasets.

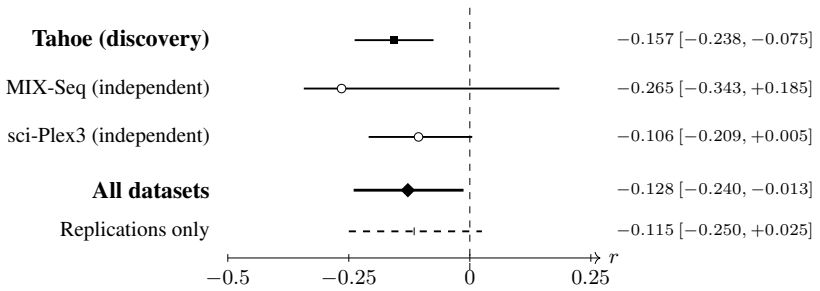


Figure 2: Frozen partial Spearman associations between geometric coherence $g_{i,-c}$ and transfer error (normalized L_2), adjusted for response magnitude; negative values favor coherence. Tahoe is the discovery dataset. The three-dataset interval excludes zero; the replication-only sensitivity (Tahoe excluded) does not. Reused verbatim from `geometry_forest.tex`; atypicality and raw-L_2 values in Appendix A.2.

4.3 METRIC DEPENDENCE

This directional signal is itself metric-dependent, anticipating the sharper metric dependence found under strict generalization (Section 5): under unnormalized (raw) L_2 error, geometry's three-dataset pooled estimate crosses zero with the opposite sign, and none of its three per-dataset raw-L_2 estimates is individually significant (exact values in Appendix A.2). Geometric coherence therefore anticipates relative or directional transfer discrepancy more clearly than raw residual magnitude; this result is not demonstrated for raw error.

4.4 GEOMETRY AND ATYPICALITY ARE NOT ONE AXIS

Baseline atypicality and geometric coherence are separately motivated signals, not components of one universal decomposition of transfer difficulty: their own correlation differs in sign across datasets ($r = +0.243$ Tahoe vs. -0.353 MIX-Seq), so we treat their relationship as dataset-dependent, not evidence for a shared underlying construct. The complete numeric predictive-structure summary is given in Appendix A.2.

Table 2: Context-grouped strict-generalization result: refit-bootstrap AURC difference (M3 vs. baseline), 95% CI, 200 replicates. Negative values favor M3.

Dataset	Metric	vs. random	vs. magnitude-only	Both exclude 0
Tahoe	`target_normalized_l2`	-0.070 $[-0.102, -0.022]$	-0.035 $[-0.087, -0.009]$	yes
Tahoe	`cosine_distance`	-0.080 $[-0.094, -0.067]$	-0.016 $[-0.024, -0.009]$	yes
Tahoe	`raw_l2`	-3.149 $[-4.894, +1.012]$	-0.510 $[-1.735, +0.826]$	no
Tahoe	`intervention_scaled_l2`	-0.068 $[-0.101, +0.006]$	-0.015 $[-0.041, +0.011]$	no
MIX-Seq	`target_normalized_l2`	-0.158 $[-0.337, -0.056]$	-0.106 $[-0.259, -0.024]$	yes
MIX-Seq	`cosine_distance`	-0.171 $[-0.237, -0.110]$	-0.135 $[-0.209, -0.064]$	yes
MIX-Seq	`raw_l2`	-90.825 $[-151.572, -29.826]$	-15.188 $[-45.623, +8.196]$	no*
MIX-Seq	`intervention_scaled_l2`	-0.079 $[-0.166, +0.010]$	-0.066 $[-0.166, +0.021]$	no

*MIX-Seq `raw_l2` excludes zero against the random baseline alone; it does not exclude zero against the magnitude-only baseline, so it does not meet the criterion, which requires both.

5 Strict Generalization Across Held-Out Contexts

The associations in Section 4 could in principle be an artifact of pair-row evaluation, where other rows from the same context or intervention are present in trust-model training. We now ask the sharper question: does a trust ranking built from these signals still rank error when an entire biological context is absent from trust-model training?

5.1 Protocol

The evaluation holds out entire contexts using five-fold `GroupKFold` by context (one fold per context when fewer than five exist). Every feature (m, g, d) is recomputed inside each split: a held-out context is excluded from every training row's response and baseline panels, not only from evaluation. The fixed joint model is $\mathrm{M3} = m + g + d$ throughout (fitting procedure, standardization, and the frozen M3/M4 interaction-selection rule in Appendix A.5). A leakage audit logs, per fold and dataset, the intervention and context exclusion constraints; all 108 logged folds across four split regimes and three datasets satisfy both, with zero violations. The preregistered criterion (frozen before outcomes were computed): context-holdout risk–coverage improves *clearly* over both a random and a magnitude-only baseline, for at least two datasets and two outcomes, where "clearly" means the paired 95% refit-bootstrap CI for the AURC difference lies entirely below zero against *both* baselines.

5.2 Primary result

Table 2 reports the context-grouped, refit-bootstrap AURC differences that determine whether the preregistered criterion is met. Normalized L_2 and cosine distance clear both baselines, with intervals entirely below zero, in both Tahoe and MIX-Seq. Raw L_2 and intervention-scaled L_2 do not: in Tahoe, neither comparison excludes zero for either metric; in MIX-Seq, raw L_2 excludes zero against the random baseline but not against the magnitude-only baseline, and intervention-scaled L_2 excludes zero against neither. Two datasets and two outcomes (normalized L_2, cosine distance) meet the preregistered criterion, recorded in `results/run_status.json`. This is a metric-dependent result, not a uniform one: it does not extend to raw or intervention-scaled error, and we do not read it as evidence that the underlying pre-response signals predict absolute transfer magnitude.

5.3 sci-Plex3

Context-holdout and double-holdout folds were constructed and audited for sci-Plex3: three `GroupKFold` context folds (one per context, 376 training / 188 held-out rows each) and fifteen double-holdout folds, all passing the leakage audit. No row from these folds, however, survives the preregistered eligibility rule, which requires at least two eligible comparison contexts besides the held-out one; with only three contexts total, holding out one leaves two, and excluding the training row's own context from its comparison panel leaves at most one. Every row is therefore ineligible, and sci-Plex3 contributes zero rows to Table 2 or to any context-holdout metric. sci-Plex3 does have intervention-grouped and pair-row results, but these use a different holdout unit and do not

substitute for the context-holdout test. sci-Plex3 is **structurally not evaluable** under strict context holdout; this is a property of the eligibility rule applied to a three-context dataset, not a failed or null replication.

5.4 SCOPE: NO TECHNICAL-ONLY COMPARISON

The preregistered protocol includes a separate outcome, triggered only if a technical-only ranking meets the context-holdout criterion while the biological M3 model does not. Evaluating this requires an authorized, externally supplied `technical.csv`; none has ever existed for this experiment, and `technical_only` is never synthesized from a proxy such as raw cell counts. Consequently `technical_only` was unavailable, was never evaluated, and this outcome was mechanically unreachable in this run: the criterion-met result above was never tested against a technical-only baseline. This is a scope limitation on what the criterion establishes, not a property of the result itself: meeting the criterion should not be interpreted as demonstrating robustness to technical confounding.

5.5 CORROBORATING EVIDENCE: FIXED-SCORE BOOTSTRAP

The result above is determined by the preregistered 200-replicate refit bootstrap, which refits preprocessing and the trust model in every replicate. A separate 2,000-replicate fixed-score bootstrap, which conditions on already-computed out-of-fold scores without refitting, gives numerically close, corroborating results for both criterion-determining metrics in both datasets (Appendix A.4); it does not determine the criterion outcome and is not mixed into Table 2.

6 IS CONTEXT DIFFICULTY AN INVARIANT PROPERTY?

This section is not a technical-confound analysis of Transfer Trust's predictors; it tests the stronger claim Sections 4–5 deliberately avoid (Section 3.3): that context-level transfer difficulty is invariant, not merely predictable. Concretely, it asks how stable a separately operationalized context-level statistic, $D_c^{(m,S)}$ (Section 3.2), is under changes in estimator, metric, resampling, and technical adjustment. $D_c^{(m,S)}$ is not $d_{c,-c}$, not e_{ic}, and is not evaluated by, or used to evaluate, either Transfer Trust predictor (Section 3.3).

6.1 SCOPE

The evidence below comes from a single, separately preregistered confirmatory extension within MIX-Seq, reusing a fixed support of 958 eligible intervention–context pairs, 13 interventions, and 69 contexts, with three independent cell draws (seeds 0–2) and all $1{,}716$ possible 6-vs-7 intervention bipartitions. A model–metric pair is counted internally reliable only if every draw's median Spearman correlation with a held-out-half replicate of $D_c^{(m,S)}$ exceeds its matched-null 95th percentile, the median across the three draw medians exceeds 0.30, and the lowest draw-specific fifth percentile (over all $1{,}716$ partitions) is positive.

6.2 ESTIMATOR DEPENDENCE AND TECHNICAL COVARIATES

Table 3 reports internal reliability under two estimators — the leave-one-context-out mean-response construction of $\widehat{\Delta}_{ic}$, and an alternative context-conditioned ridge fit on other contexts' control profiles — with mean-vs-ridge consensus agreement before/after subtracting a technical-only out-of-fold prediction (full values in Appendix A.3). Raw and intervention-scaled L_2 are reliable under both estimators and stay above the 0.30 agreement threshold after adjustment. Normalized L_2 is reliable under mean-response but fails ridge's lower-tail rule, and its agreement collapses after adjustment. Cosine fails both estimators and its agreement reverses sign. A nested regression on the seven frozen technical/QC covariates achieves out-of-fold R^2 of 0.453–0.653 (mean-response) and 0.230–0.636 (ridge). $D_c^{(m,S)}$ is therefore metric- and estimator-dependent, and technical covariates explain a substantial share of the remaining cross-estimator agreement — evidence about $D_c^{(m,S)}$,

Table 3: Context Transferability: estimator and technical dependence (MIX-Seq, Phase 7). Consensus is mean-response/ridge Spearman agreement on $D_c^{(m,S)}$, raw $\rightarrow$ after subtracting a technical-only prediction.

Metric	Mean-resp. reliable	Ridge reliable	Consensus (raw $\rightarrow$ adj.)
Normalized L_2	yes	no	$0.638 \rightarrow 0.086$
Raw L_2	yes	yes	$0.776 \rightarrow 0.707$
Intervention-scaled L_2	yes	yes	$0.761 \rightarrow 0.571$
Cosine distance	no	no	$0.441 \rightarrow -0.118$

not that technical covariates explain Transfer Trust's predictors, which this analysis does not evaluate.

This does not contradict Section 5: that section tests whether pre-response signals predict realized transfer error under one fixed procedure (they do, for two metrics, under a strict context holdout); this one tests whether a context-level summary of difficulty keeps the same ranking when that procedure changes — not competing answers to the same question.

6.3 Permuted-baseline negative control

A permuted-baseline ridge estimator — identical to the ridge construction above, except that control (baseline) profiles are randomly reassigned to context labels before fitting, severing the mapping between context identity and baseline expression — satisfies the same three-part internal- reliability criterion for *all four* metrics. It does not prove that the surviving raw/scaled structure is purely technical; it shows that within-estimator split-half reliability can be reproduced even after the context–baseline mapping is destroyed, and therefore that internal reliability alone is insufficient evidence that a measured context-level statistic corresponds to a meaningful, biologically intrinsic property of context.

6.4 A model-specific, not estimator-invariant, result

Under the preregistered rule (full statement in Appendix A.12), evaluated in order: the criterion for terminating the analysis as unreliable is not met, because ridge retains reliability for two metrics; the criterion for a purely technical explanation does not apply, because it requires the technical-only pattern to hold for both estimator families, and it holds for all four mean-response metrics but only two ridge metrics; the criterion for estimator-invariant support does not apply, because ridge reliability and technical-residual agreement each clear their thresholds for only two of the four metrics. The results therefore support a model-specific interpretation rather than an estimator-invariant one, following a rule fixed before these outcomes were computed; the finding is the pattern of reliabilities, agreements, and the permuted-baseline result above, not any single label.

6.5 Synthesis

Sections 4–5 show exploitable pre-response structure for two signals, under two metrics, in two datasets, surviving a strict context holdout; this section shows that a related but separately operationalized context-level summary of transfer difficulty cannot automatically be promoted to an estimator-independent or intrinsic biological property. Predictive structure $\neq$ invariant structure.

7 Selective Prediction and Practical Triage

If the pre-response signals in Section 4 are modest and, per Section 6, not intrinsic biological structure, are they still useful? We ask narrowly whether the joint trust score — $\mathrm{M3} = m + g + d$ (Section 5), or its extension $\mathrm{M4} = \mathrm{M3} + (g{:}d)$ where a geometry–difficulty interaction meets the frozen selection rule — ranks intervention–context pairs by relative transfer risk well enough to support retrospective triage: which predicted transfers to use first, and which to defer or measure directly.

Retaining the top-ranked 10% of pairs by joint score monotonically reduces mean transfer error in both Tahoe and sci-Plex3 (Figure 3; exact values in Appendix A.7); on sci-Plex3, magnitude or geometry alone at 10% coverage is worse than full coverage and random, so triage benefits here from combining weak signals rather than any single one. MIX-Seq is *non-monotonic*: geometric coherence alone is slightly better at 10% coverage (teal dips below blue in the figure) — reported rather than smoothed over. This is retrospective, uncalibrated triage utility within the studied datasets, not a guarantee of monotonic risk reduction, causal evidence, or a claim of prospective or clinical deployment.

This risk-coverage behavior is the practical payoff of a modest ranking signal. Cross-validated AUROC for the frozen selected joint model is modest: 0.662 in Tahoe, 0.678 in MIX-Seq, and 0.668 in sci-Plex3 (Table 8); an intervention-grouped sensitivity leaves Tahoe and sci-Plex3 essentially unchanged (0.662, 0.656) but weakens MIX-Seq to 0.607 (13 interventions). Triage only needs ranking pairs better than chance, so a modest AUROC is compatible with the risk-coverage improvement above: retaining a top-ranked fraction depends on relative ordering, not absolute separability. We do not describe the joint score as a strong classifier.

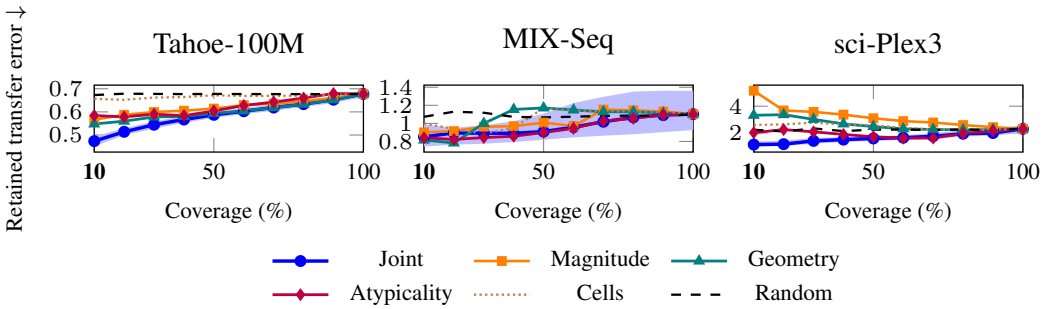


Figure 3: Selective prediction at ten prespecified retention levels; lower retained error is better; each panel has its own y-scale, so absolute error magnitudes are not comparable across datasets. All six rankings — joint (M3/M4), magnitude, geometry, atypicality, cell-count, random — keep a distinct color and marker; joint is only slightly thicker. The bold "10" tick marks the retention level discussed in the text. MIX-Seq is non-monotonic: geometry (teal triangles) is slightly better than the joint score at 10% coverage. Shading is a pointwise 95% percentile interval for the joint curve (2,000 resamples, conditional on fixed scores; not simultaneous). Reused verbatim (data paths adapted) from `risk_coverage.tex`; AUROC and exact risk values in Appendix A.6.

Full AUROC and coverage-risk values are in Appendix A.6.

8 Discussion and Limitations

Scientific meaning. The strongest defensible conclusion from Sections 4–6 is a distinction, not a single number: transfer error contains predictable structure — grounded in Tahoe and MIX-Seq under strict leakage-safe context holdout, beyond random and magnitude-only baselines, for normalized and directional error only, not raw or intervention-scaled, and not established in sci-Plex3 (Section 5) — but this is not, by itself, evidence for invariant or intrinsic biological structure. Baseline atypicality is the more consistently supported predictor; geometric coherence is comparatively modest, individually significant only in the discovery dataset. The qualifying half (Section 6) shows a separately operationalized context-level statistic failing an estimator swap for exactly the metrics carrying the predictive result, and substantially associated with technical covariates; reporting only the first half would invite the overinterpretation the second rules out.

Predictive vs. invariant structure, and remaining limits. These claims carry different evidentiary bars (Table 4); treating one level as automatic evidence for a stronger level is the error this paper avoids. Two gaps follow. Transfer Trust's own predictors $d_{c,-c}$, $g_{i,-c}$ have never been evaluated for estimator invariance or against a technical-only baseline (Section 5), so Context Transferability's estimator-swap/technical-covariate finding for its own construct $D_c^{(m,S)}$ (Section 6) neither tests nor

extends to them. And none of the reported associations are causal, nor does Section 7's ranking utility — which needs only better-than-chance ranking of held-out pairs — depend on either predictor being estimator-invariant or intrinsic. Scope is bounded throughout: Context Transferability's decisive analysis is a single MIX-Seq extension, and all three datasets are transcriptomic, small-molecule, cancer-cell-line screens (Appendix A.13), with no claim beyond them — predictive structure need not imply invariant or intrinsic structure, and only the narrower claim is established here. The two most direct unresolved tests are a technical-covariate comparison for Transfer Trust's own predictors, and replication of Context Transferability's estimator-dependence audit beyond MIX-Seq. In short: there is exploitable pre-response structure in perturbation-transfer reliability, but its usefulness for prediction does not establish estimator-invariant or intrinsic biological structure.

AI USE STATEMENT

Generative-AI assistants were used during parts of the research and manuscript workflow, including code generation and editing, debugging, experiment-orchestration support, repository inspection, analysis support, LaTeX authoring, manuscript drafting and editing, and structured consistency, traceability, and claim-integrity audits. The human authors defined the scientific questions, determined the experimental and evaluation strategy, reviewed and accepted or rejected analyses, independently verified quantitative findings against the corresponding source artifacts, interpreted the results, determined which claims and results were scientifically supported and appropriate to report, and approved the final manuscript. AI-generated outputs were not treated as scientific evidence without human verification. The authors take responsibility for the final analyses, claims, text, and conclusions.

ETHICS STATEMENT

This work is a secondary, computational integration of previously published, preregistered analyses of existing single-cell perturbation datasets (Tahoe-100M, MIX-Seq, sci-Plex3), all derived from cancer cell lines rather than primary human tissue or living human subjects. No new data were collected, no human-subject intervention was conducted, and no IRB submission was made in connection with this work; ethical oversight of the original data generation is the responsibility of, and was addressed by, the respective original publications. We are not aware of concerns under the ICLR Code of Ethics raised by this work.

REPRODUCIBILITY STATEMENT

All scientific results reported in this manuscript are drawn from frozen, preregistered analyses completed prior to this manuscript's integration and never recomputed here (Appendix A.13). Transfer Trust's strict-generalization protocol fixes its preregistered criterion, split regimes, and outcome definitions before outcomes are computed, and is evaluated under a leakage-safe, fold-specific feature-recomputation scheme; every fold's intervention and context exclusion constraints are logged and audited ($108/108$ records satisfy both constraints, with zero violations). Context Transferability's estimator-dependence and technical-covariate analyses (Section 6) are likewise preregistered, with an internal-reliability criterion fixed before outcomes were computed. Every quantitative claim, table, and appendix value in this manuscript traces to a specific frozen result file or preregistered-criterion record (Appendix A.13); no value was recomputed or approximated for this manuscript. Specific repository locations and file paths are withheld from this anonymous version and will accompany a full artifact release at camera-ready.

ACKNOWLEDGMENTS

Omitted for anonymous review; will be added at camera-ready.

REFERENCES

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 Rätsch. Learning single-cell perturbation responses using neural optimal transport. *Nature Methods*, 20:1759–1768, 2023. doi: 10.1038/s41592-023-01969-x.

Leon Hetzel, Simon Böhm, Niki Kilbertus, Stephan Günnemann, Mohammad Lotfollahi, and Fabian J. Theis. Predicting cellular responses to novel drug perturbations at a single-cell resolution. In *Advances in Neural Information Processing Systems*, volume 35, pp. 26711–26722, 2022. doi: 10.52202/068431-1937.

Minyoung Huh, Brian Cheung, Tongzhou Wang, and Phillip Isola. Position: The platonic representation hypothesis. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, 2024.

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

Mohammad Lotfollahi, Anna Klimovskaia Susmelj, Carlo De Donno, Leon Hetzel, Yuge Ji, Ignacio L. Ibarra, Sanjay R. Srivatsan, et al. Predicting cellular responses to complex perturbations in high-throughput screens. *Molecular Systems Biology*, 19(6):e11517, 2023. doi: 10.15252/msb.202211517.

James M. McFarland, Brenton R. Paolella, Allison Warren, Kathryn Geiger-Schuller, Tsukasa Shibue, Michael Rothberg, Olena Kuksenko, William N. Colgan, Andrew Jones, Emily Chambers, et al. Multiplexed single-cell transcriptional response profiling to define cancer vulnerabilities and therapeutic mechanism of action. *Nature Communications*, 11(1):4296, 2020. doi: 10.1038/s41467-020-17440-w.

Yaniv Ovadia, Emily Fertig, Jie Ren, Zachary Nado, D. Sculley, Sebastian Nowozin, Joshua Dillon, Balaji Lakshminarayanan, and Jasper Snoek. Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

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.

Sanjay R. Srivatsan, José L. McFaline-Figueroa, Vijay Ramani, Lauren Saunders, Junyue Cao, Jonathan Packer, Hannah A. Pliner, Dana L. Jackson, Riza M. Daza, Lena Christiansen, et al. Massively multiplex chemical transcriptomics at single-cell resolution. *Science*, 367(6473):45–51, 2020. doi: 10.1126/science.aax6234.

Jesse Zhang, Airol A. Ubas, Richard de Borja, Valentine Svensson, Nicole Thomas, Neha Thakar, Ian Lai, Aidan Winters, Umair Khan, Matthew G. Jones, et al. Tahoe-100M: A giga-scale single-cell perturbation atlas for context-dependent gene function and cellular modeling. *bioRxiv*, 2025. doi: 10.1101/2025.02.20.639398.

A Appendix

A.1 Four kinds of structure: full taxonomy

Table 4 restates, with source pointers, the four-way distinction introduced in Section 3.3: no level implies the next, and each is established (or not) by a specific result in this paper.

A.2 Full predictive-structure table

Table 5 gives the complete numeric summary underlying Section 4's prose and Figure 2: both signals (baseline atypicality and geometric coherence), all three datasets, both the primary normalized-L_2 metric and the raw-L_2 sensitivity, with frozen 95% confidence intervals and status labels.

Table 4: Terminology used throughout the paper.

Term	Meaning as used in this paper
Predictive structure	A pre-response signal ($d_{c,-c}$ or $g_{i,-c}$) is associated with transfer error (e_{ic}) for a specific held-out pair (Section 4); combined into a joint trust ranking, it survives strict context holdout for two metrics, two datasets (Section 5).
Reproducible structure	A statistic ($D_c^{(m,S)}$) is stable across independent cell draws and independent control pools, under one fixed estimator. Part of the evidence in Section 6.
Estimator-invariant structure	A statistic's context ranking is stable when the estimator used to construct it changes (e.g., mean-response vs. ridge). Shown in Section 6 to hold for raw/scaled metrics and to fail for normalized/cosine metrics; *not established for $d_{c,-c}$ or $g_{i,-c}$ at all*.
Biologically intrinsic structure	A property of context that would hold independent of measurement, estimator, and technical covariates. Not established by any result in this paper.

Table 5: Predictive-structure summary. All associations are Spearman; geometric coherence is magnitude-adjusted (partial correlation conditional on response magnitude); baseline atypicality is a direct context-level correlation. "Pooled" is a three-dataset random-effects (DerSimonian–Laird) synthesis; "replication-only" excludes the Tahoe discovery dataset.

Signal	Dataset	Metric	r	95% CI	Status
Baseline atypicality	Tahoe	normalized L_2	$+0.312$	$[0.202, 0.414]$	individually significant
Baseline atypicality	MIX-Seq	normalized L_2	$+0.425$	$[0.316, 0.532]$	individually significant
Geometric coherence	Tahoe	normalized L_2	-0.157	$[-0.238, -0.075]$	individually significant
Geometric coherence	MIX-Seq	normalized L_2	-0.265	$[-0.343, +0.185]$	directionally consistent
Geometric coherence	sci-Plex3	normalized L_2	-0.106	$[-0.209, +0.005]$	directionally consistent
Geometric coherence	pooled (3 datasets)	normalized L_2	-0.128	$[-0.240, -0.013]$	pooled support
Geometric coherence	replication-only	normalized L_2	-0.115	$[-0.250, +0.025]$	underpowered replication
Geometric coherence	pooled (3 datasets)	raw L_2	$+0.025$	$[-0.091, +0.140]$	not demonstrated

A.3 Full Context Transferability reliability table

Table 6 gives the complete per-metric, per-estimator median Spearman correlation and lowest draw-specific fifth percentile underlying the pass/fail and consensus summary in Table 3.

Table 6: Frozen Phase 7 estimator-dependence results (MIX-Seq). Median ρ is the median of three draw-specific medians; "lowest 5th" is the minimum draw-specific fifth percentile over all 1,716 partitions.

Estimator	Metric	Median ρ	Lowest 5th	Reliable
Mean response	normalized L_2	0.682	0.104	yes
Mean response	raw L_2	0.621	0.377	yes
Mean response	intervention-scaled L_2	0.787	0.584	yes
Mean response	cosine distance	0.733	-0.033	no
Context ridge	normalized L_2	0.326	-0.018	no
Context ridge	raw L_2	0.641	0.403	yes
Context ridge	intervention-scaled L_2	0.798	0.684	yes
Context ridge	cosine distance	0.199	-0.145	no

A.4 Fixed-score bootstrap (corroborating evidence)

Table 7 reports the 2,000-replicate fixed-score bootstrap referenced in Section 5, for the two metrics that determine the preregistered criterion in both datasets. It corroborates, but does not determine, the criterion outcome in Table 2.

Table 7: Context-grouped fixed-score bootstrap AURC difference (M3 vs. baseline), 95% CI, 2,000 replicates.

Dataset	Metric	vs. random	vs. magnitude-only
Tahoe	`target_normalized_l2`	-0.086 $[-0.105, -0.063]$	-0.035 $[-0.057, -0.015]$
Tahoe	`cosine_distance`	-0.084 $[-0.097, -0.072]$	-0.016 $[-0.022, -0.011]$
MIX-Seq	`target_normalized_l2`	-0.156 $[-0.328, -0.046]$	-0.090 $[-0.201, -0.023]$
MIX-Seq	`cosine_distance`	-0.172 $[-0.237, -0.114]$	-0.146 $[-0.249, -0.056]$

A.5 Trust ranking model: M3/M4 fitting procedure

M3 and M4 are both ordinary-least-squares linear regressions (scikit-learn `LinearRegression`, no penalty term) predicting the realized error outcome directly from the feature set, inside a pipeline that standardizes features (zero mean, unit variance) immediately beforehand. M3 uses $\{m_{i,-c}, g_{i,-c}, d_{c,-c}\}$ (magnitude, geometry, atypicality); M4 adds one engineered feature, the product $g_{i,-c} \times d_{c,-c}$. Whether a given fold–outcome combination uses M4 instead of M3 is decided by a separate ordinary-least-squares fit on that fold's training rows only (the outcome regressed on magnitude, geometry, atypicality, and their geometry$\times$atypicality interaction); M4 is selected only if that interaction term's p-value is finite and below 0.05 (folds with fewer than 10 training rows default to M3). Both the interaction test and the final M3/M4 fit — including the standardization step — are refit independently inside every held-out-context fold, using only that fold's training rows; no statistic is carried across folds. The resulting held-out (out-of-fold) predicted error for every pair is what ranks pairs for risk–coverage: retaining the top-$k\%$ coverage means retaining the $k\%$ of pairs with the *lowest* predicted error, and the reported risk is the mean *realized* error among the retained pairs (Section 7, Figure 3).

A.6 Full selective-prediction table

Table 8 gives the complete numeric summary underlying Section 7's prose and Figure 3: global and intervention-grouped AUROC, and mean normalized L_2 risk at full versus 10% coverage, for all three datasets.

Table 8: Selective-prediction ranking utility (retrospective, out-of-fold). AUROC (grouped) holds out complete interventions. Risk is mean normalized L_2 error at full vs. 10% coverage for the frozen selected joint model (Tahoe/sci-Plex3: M4; MIX-Seq: M3).

Dataset	AUROC (global)	AUROC (grouped)	Risk: full	Risk: 10%
Tahoe	0.662	0.662	0.678	0.474
MIX-Seq*	0.678	0.607	1.102	0.851
sci-Plex3	0.668	0.656	2.307	1.137

*Non-monotonic: geometric coherence alone reaches 0.815 at 10% coverage, slightly below the joint score.

A.7 Selective-prediction detail: single-signal rankings and risk-coverage intervals

Ranking by magnitude alone or geometry alone at 10% coverage gives sci-Plex3 mean normalized error of 5.155 and 3.321 respectively, both worse than full-coverage error (2.307) and worse than a random ordering (2.242); the joint ranking (Table 8) is monotonic and better than all three. In MIX-Seq, geometric coherence alone reaches 0.815 at 10% coverage versus 0.851 for the joint score, the one case in which a single signal is competitive with the joint ranking. The 10%-coverage joint-score percentile-bootstrap intervals (2,000 resamples of complete interventions, conditional on fixed out-of-fold scores) are $[0.450, 0.493]$ (Tahoe), $[0.744, 0.987]$ (MIX-Seq), and $[0.996, 1.363]$ (sci-Plex3); they quantify intervention-sampling variation only, are not simultaneous, and do not include model-refitting uncertainty.

A.8 Robustness diagnostics and magnitude controls

Response magnitude was treated as a nuisance covariate throughout (Section 3.1). For geometric coherence, adjusting for magnitude attenuated but did not eliminate the raw coherence–error correlation in Tahoe ($-0.374 \rightarrow -0.232$) or in MIX-Seq ($-0.281 \rightarrow -0.265$); in sci-Plex3, magnitude alone was itself positively associated with error ($r = +0.147$, $p = 4.63 \times 10^{-4}$), and adjustment *strengthened* rather than attenuated the geometry association ($-0.070 \rightarrow -0.106$). Training-cell count had a negligible association with error in Tahoe ($r = -0.041$) and MIX-Seq ($r = -0.042$, $p = 0.151$), and a small association in sci-Plex3 ($r = +0.152$, $p = 2.91 \times 10^{-4}$); drug frequency was null in MIX-Seq ($r = +0.027$, $p = 0.360$). Between-plate variance was 0.26% in Tahoe and 14.1% in sci-Plex3. Baseline atypicality remained associated with error after separately controlling for raw control-profile scale in Tahoe ($r = 0.141$) and MIX-Seq ($r = 0.264$). These are robustness controls, not evidence of a causal mechanism.

A.9 Leave-one-intervention-out (LOIO) analyses

For sci-Plex3 ($n = 188$ interventions), leave-one-intervention-out re-estimates of the geometry–error association ranged over $[-0.118, -0.096]$ (SD 0.004), with $0/188$ sign flips; removing the five or ten most influential interventions strengthened the estimate to -0.137 and -0.167 respectively. These are diagnostics against a magnitude- or single-outlier-driven explanation, not evidence of causality.

A.10 PCA artifact exclusion

An independently refitted PCA analysis, in which each Tahoe plate's coordinates were computed from a separately fitted basis, produced an apparently large geometry effect that was an artifact of comparing incommensurable plate-specific coordinate systems; it is excluded from the evidence base rather than reported as support. Repeating the geometry synthesis with a single shared basis reproduced a stable result (consistent with the raw-representation estimate used throughout this manuscript, $I^2 = 0\%$), but the corresponding atypicality synthesis under that shared basis was heterogeneous ($I^2 = 81.7\%$) and reversed sign on one plate. The raw (non-PCA) representation used throughout this manuscript is the frozen primary representation and had $I^2 = 0\%$ for both signals across plates.

A.11 Estimator details

The mean-response estimator (Section 3.2) excludes the target context's response by construction. The context-conditioned ridge alternative instead maps each target context's own control pseudobulk profile to its perturbation response via ridge regression fit on other contexts only; the outer holdout excludes the target response from both fitting and regularization-strength selection, and every inner fold refits gene-wise scaling and the ridge coefficients independently. Ridge is evaluated only in MIX-Seq and is treated as one alternative estimator, not a claim of estimator generalization: extending it to Tahoe or sci-Plex3, or adding further estimator families, was outside the scope of the frozen analysis audited in Section 6.

A.12 Context Transferability preregistered evaluation rule

The preregistered rule, evaluated in priority order, is: terminate as unreliable if ridge is internally reliable for fewer than two of the four metrics; otherwise conclude a purely technical explanation if, for at least three metrics in *both* the mean-response and ridge estimators, technical-only out-of-fold R^2 reaches at least 90% of the combined (technical + baseline) R^2 and the baseline group's incremental R^2 over technical-only is at most 0.02; otherwise conclude estimator-invariant support if mean-response and ridge are each internally reliable for at least three metrics, their consensus Spearman is at least 0.50 for at least three metrics, and technical-residual Spearman is at least 0.30 for at least three metrics; otherwise conclude a model-specific result if the analysis is identifiable and mean-response and ridge are each reliable for at least two metrics. Applied to the frozen results (Table 3, Appendix A.3): ridge is reliable for exactly two metrics (raw, scaled), so the termination criterion is not met; the technical-only pattern holds for all four mean-response metrics but only two

ridge metrics, short of the technical-explanation criterion's joint requirement; ridge reliability and technical-residual agreement each clear their thresholds for only two of the four metrics, short of the estimator-invariant-support criterion. The result is a model-specific outcome.

A.13 PROVENANCE AND REPRODUCIBILITY

Every scientific result in this manuscript originates from one of three independently frozen, preregistered analyses, referenced here by role only; specific repository locations, checkpoint identifiers, and file paths are withheld from this anonymous version to avoid identifying the authors, and will be disclosed alongside a full artifact release at camera-ready.

- **Transfer Trust**: the pre-response predictors ($d_{c,-c}$, $g_{i,-c}$), the strict-generalization protocol and its results (Table 2, Appendix A.4), and the selective-prediction results (Table 8, Appendix A.7).
- **Context Transferability**: the separately operationalized context-level statistic $D_c^{(m,S)}$ and all estimator- and technical-dependence results (Table 3, Appendix A.3).
- **Intervention Geometry**: the source, under a recorded one-time authorized import, of the raw response and control-baseline vectors used to construct Transfer Trust's strict-generalization inputs.

Each analysis fixed its preregistered criterion, split regimes, and outcome definitions before outcomes were computed, and none was recomputed for this manuscript: every number reported here, in the main text and appendix alike, traces to a specific frozen result file or preregistered-criterion record within the corresponding analysis. This traceability is internal to the authors' own version-controlled record and is not itself a public data or code release; a full artifact-to-number mapping will accompany the camera-ready version.