A Pre-Registered Direction Instability Atlas Across Three Perturbation Modalities

Elliot Tower

Independent researcher
elliot@elliottower.ai

Abstract

Whether a perturbation produces consistent effects across cell types determines if laboratory findings transfer to patients—yet no existing metric separates inconsistent effects from merely weak ones. Selectivity metrics count how many targets a drug hits; direction instability (DI) measures whether downstream effects depend on cell type. Across ∼55,000 drugs and genetic perturbations, DI reveals that mechanism of action structures transferability. Drugs targeting core cellular machinery—mTOR inhibitors, HDAC inhibitors—act the same way regardless of cell type (low DI), while receptor-signaling drugs—serotonin agonists, histamine agonists—produce cell-type-dependent effects (high DI). A highly selective drug can still have high DI if its target pathway is wired differently across cell types—selectivity alone does not predict transferability. Essential genes produce uniform disruption across cell types, extending the pattern from drugs to genetic perturbations. The atlas provides a lookup resource for whether a perturbation's effects will transfer to a new context.

1 Introduction

High-throughput perturbation screens now profile thousands of drugs and genetic perturbations across panels of cell lines, producing multi-dimensional signatures in transcriptomic [10, 17], morphological [2], and single-cell modalities [19]. Whether a perturbation's effects transfer across cellular contexts has practical consequences for drug development and genetic screening [11, 20].

Direction instability (DI) quantifies this context-dependence as $\mathrm{DI} = 1 - \overline{\cos(\mathbf{s}_i, \mathbf{s}_j)}$, the mean pairwise cosine dissimilarity among a perturbation's signatures across contexts. Because cosine normalizes each signature to unit length before comparison, DI is immune to the first-order magnitude confound—systematic covariation between effect size and metric value—that dominates alternative geometric measures such as Grassmannian geodesic distance, profile variability, and optimal transport cost [20]. A residual second-order correlation persists near zero (noise-dominated signatures inflate pairwise cosine dissimilarity), which we remove by magnitude correction (Section 2). A companion study [20] established this immunity and showed that DI predicts cross-cell-line mechanism transport (pre-registered, construction-neutral AUROC = 0.945 on 8,949 LINCS drugs). Because DI is defined through cosine geometry, that study validated it through three independent biological channels that do not involve cosine similarity. These include Jaccard overlap of perturbed gene sets (AUROC = 0.957), breadth of target gene expression across tissues ($\rho = -0.17$, $p < 0.001$), and PRISM cell-viability correlation ($\rho = 0.261$, $p < 0.001$). A separate construct-validity study [21] extended these results to Perturb-seq and JUMP-CP morphology, showing that domain-specific validity conditions preserve global DI rankings ($\rho > 0.91$ in all three domains) while upgrading

inferential warrant at the item level. The present study takes DI's validity as established and asks a different question: what does DI capture, and what does it fail to capture, across modalities and perturbation types?

We constructed a DI atlas spanning three measurement technologies and four perturbation classes: drug perturbations measured by L1000 transcriptomics (LINCS, 8,949 compounds) and single-cell transcriptomics (Tahoe-100M, 379 compounds); genetic knockouts measured by Cell Painting morphology (JUMP-CP CRISPR, 7,946 genes); and genetic overexpressions measured by Cell Painting morphology (JUMP-CP ORF, 12,590 genes). For each perturbation, the atlas provides magnitude-corrected DI with bootstrap confidence intervals, enabling downstream analyses that require a calibrated measure of context-dependence.

To characterize this resource, we pre-registered eleven primary hypotheses across three batches before computing DI on new datasets (Batch 1 SHA `0d07a01`; Batch 2A SHA `a17c125`; Batch 3 SHA `4c34699`). The hypotheses test whether DI is concordant across modalities, whether it correlates with gene essentiality, whether it is an intrinsic property of genes, whether expression variance, drug selectivity, or paralog buffering can explain it, and whether DI is structured by drug mechanism-of-action class. Batch 3 was prompted by peer review of an earlier draft and adds three experiments: MoA-stratified DI characterization, cross-cell-line CRISPR replication of the essentiality sign-flip, and convergent evidence from DepMap dependency-score variability. Of the eleven primary tests, three pass their pre-registered criteria and eight fail. A twelfth test—PRISM viability concordance—is exploratory.

The results sharply constrain what DI measures. Cross-modal concordance confirms that DI reflects biology rather than measurement artifact: compounds whose transcriptomic effects vary across cell lines also produce variable morphological profiles. Essential genes show *low* DI, because functional constraint produces stereotyped disruption regardless of cellular identity—a sign-flip that directionally replicates on true cross-cell-line CRISPR data ($\rho = -0.26$, cpg0000 pilot, $K = 2$ cell lines). Knockout and overexpression DI are unrelated, meaning DI characterizes how a specific perturbation mechanism interacts with context rather than indexing a stable property of the target gene. Expression variance, paralog count, and paralog expression breadth produce correlations too small to matter biologically (partial $\rho < 0.07$ after confound control), ruling them out as explanations for DI variation. DI varies systematically by drug mechanism-of-action class: core-machinery inhibitors (mTOR, HDAC) show low DI while receptor-signaling drugs (serotonin agonists, histamine agonists) show high DI, demonstrating that the atlas captures biologically interpretable structure.

Taken together, these results constrain DI to a dimension of perturbation biology that existing gene annotations do not index, and the atlas makes that measurement available at scale.

The remainder of the paper is organized as follows. Section 2 describes the atlas as a resource: datasets, DI computation, and magnitude correction. Section 3 presents the construct validity results, with PRISM viability as independent corroboration. Section 4 reports the informative null and sign-flipped results that delimit what DI measures. Section 5 presents the reviewer-prompted Batch 3 extensions: MoA stratification, cross-cell-line CRISPR replication, and the DepMap convergent-evidence failure. Section 6 describes the pre-registration protocol and deviations. Section 7 synthesizes the results into a characterization of what DI measures.

2 The DI atlas as a resource

2.1 Datasets

The atlas draws on five perturbation datasets spanning three measurement technologies (Table 1). Two drug datasets were previously analyzed in the companion confound study [20]: LINCS L1000 transcriptomics (978 landmark genes, 8,949 compounds across variable numbers of cell lines; 17) and JUMP-CP Cell Painting compound morphology (3,180 features, 25,254 compounds; 3). Three datasets are new to this study: Tahoe-100M single-cell transcriptomics (379 drugs across 47–50 cell lines per drug; 19), JUMP-CP CRISPR knockouts (7,946 genes, 259 post-sphering PCA features, plate replicates as contexts; 4), and JUMP-CP ORF overexpressions (12,590 genes, 722 features, plate replicates as contexts; 4). Together, the atlas covers ∼55,000 perturbations (8,949 LINCS drugs, 379 Tahoe drugs, 25,254 JUMP-CP compounds, 7,946 CRISPR knockouts, 12,590 ORF overexpressions; drug-dataset overlaps are small, $n \leq 378$).

An important asymmetry: LINCS and Tahoe compute DI over cell lines, measuring biological context-dependence, while JUMP-CP CRISPR and ORF compute DI over plate replicates within a single cell line, measuring technical reproducibility. The two concordance tests (Section 3) compare datasets that share the same context type (cell lines), but the essentiality, mechanism-specificity, and proxy tests (Section 4) use plate-replicate DI. Section 7 discusses how this distinction affects interpretation.

Table 1: Datasets in the DI atlas. "Contexts" denotes the unit over which DI is computed: cell lines for drug datasets, plate replicates for genetic perturbation datasets. Only perturbations with ≥ 5 contexts are included. The pre-registration estimated 15,121 ORF genes; the actual count after quality filtering is 12,590.

Dataset	Modality	Perturbations	Features	Contexts	Source
LINCS L1000	Transcriptomic	8,949 drugs	978	Cell lines	Pre-computed
Tahoe-100M	Transcriptomic	379 drugs	∼62,700	Cell lines	HuggingFace
JUMP-CP compounds	Morphological	25,254 compounds	3,180	Varied[†]	Pre-computed
JUMP-CP CRISPR	Morphological	7,946 genes	259	Plate reps	CPG
JUMP-CP ORF	Morphological	12,590 genes	722	Plate reps	CPG

CPG = Cell Painting Gallery (cpg0016). [†]Context definitions vary by compound; see companion study [20].

A sixth dataset, PRISM repurposing viability [5], is included as a complementary validation source. PRISM measures scalar cell viability (log-fold-change) rather than high-dimensional profiles, so cosine-based DI is undefined. We compute the coefficient of variation (CV) of viability across cell lines as a scalar analogue of context-dependence. PRISM results are labeled exploratory throughout.

2.2 DI computation

For a perturbation with $K \geq 5$ context-specific signatures $\mathbf{s}_1, \ldots, \mathbf{s}_K \in \mathbb{R}^d$, direction instability is

$$\mathrm{DI} = 1 - \frac{2}{K(K-1)} \sum_{i<j} \frac{\mathbf{s}_i \cdot \mathbf{s}_j}{\|\mathbf{s}_i\| \, \|\mathbf{s}_j\|}. \tag{1}$$

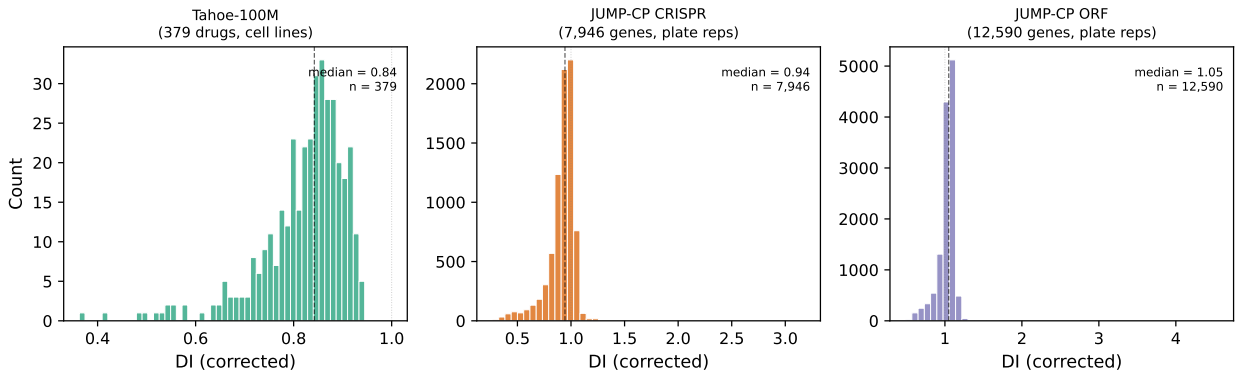

Figure 1: Distribution of magnitude-corrected DI across the three new atlas datasets. Dashed lines mark medians; dotted lines mark DI = 1 (orthogonal signatures). DI can exceed 1 when signatures are anti-correlated across contexts. Tahoe-100M drug DI (median = 0.84) spans a wider range than the genetic perturbation datasets (CRISPR median = 0.94, ORF median = 1.05), consistent with drugs engaging more diverse mechanisms. CRISPR and ORF distributions are concentrated near 1 (near-orthogonal signatures across plate replicates), reflecting the high baseline noise of morphological profiling at plate-replicate resolution.

DI = 0 when all signatures point in the same direction (perfectly consistent perturbation), equals 1 when signatures are orthogonal, and exceeds 1 when signatures are anti-correlated. By construction, DI is a reparameterization of mean pairwise cosine similarity (Pearson $r = -1.0$ between DI and $\overline{\cos}$).

2.3 Magnitude immunity and correction

The first-order magnitude confound arises when a metric's expected value changes with effect size: stronger perturbations mechanically produce larger distances (or smaller ones, depending on the metric). Cosine similarity is ratio-scale—$\cos(\alpha\mathbf{s}_i, \alpha\mathbf{s}_j) = \cos(\mathbf{s}_i, \mathbf{s}_j)$ for any $\alpha > 0$—so DI is invariant to uniform rescaling of all signatures. This is the sense in which DI is "immune by construction": the dominant confound pathway (stronger effect → larger norm → larger distance) is algebraically cancelled. The companion study [20] verified this empirically across three modalities (LINCS, DepMap CRISPR, Perturb-seq), showing that alternative metrics exhibit Cohen's $d = 0.55$–2.65 magnitude confounding while DI shows none after correction.

A residual second-order correlation remains. Weakly-acting perturbations produce noisier signature estimates, and noise inflates the average pairwise cosine dissimilarity even after unit-length normalization. We remove this residual confound by OLS regression of raw DI on mean signature L_2 norm:

$$\mathrm{DI}_{\mathrm{corrected}} = \mathrm{DI}_{\mathrm{raw}} - \hat{f}(\overline{\|\mathbf{s}\|}) + \hat{\beta}_0, \tag{2}$$

where $\hat{f}$ is the OLS prediction and $\hat{\beta}_0$ is the intercept. This correction is applied per dataset. $\mathrm{DI}_{\mathrm{corrected}}$ is the primary metric throughout; $\mathrm{DI}_{\mathrm{raw}}$ is reported for transparency.

2.4 Bootstrap confidence intervals

For each perturbation, we resample the K contexts with replacement 1,000 times and recompute DI on each bootstrap sample. The 2.5th and 97.5th percentiles define the 95% bootstrap confidence

interval. These intervals quantify the stability of each DI estimate with respect to the particular set of contexts observed.

2.5 Statistical framework

All hypothesis tests use partial Spearman correlation with confound control. The partial correlation is computed by Pearson-residualizing both variables against the covariate matrix (with intercept column), then computing Spearman's ρ on the residuals. This method avoids the bias of rank-then-residualize approaches, which produce spurious partial correlations of $\rho \approx 0.05$–0.25 under the null at large n (verified in simulation; see pre-registration documents). Confidence intervals use the Fisher z-transform with degrees of freedom adjusted for the number of covariates.

Every primary test applies two decision criteria simultaneously: (i) $p < \alpha$ (Bonferroni-corrected: $\alpha = 0.0125$ for Batch 1's four primary tests, $\alpha = 0.0167$ for Batch 2A's three hypotheses and Batch 3's three), and (ii) the 95% CI lower bound exceeds ± 0.05 in the predicted direction. The CI floor is necessary because at $n > 5{,}000$, effects as small as $\rho = 0.03$ reach statistical significance. Without the floor, trivially small correlations would pass.

2.6 Data format

The atlas is distributed as CSV files containing, for each perturbation: $\mathrm{DI}_{\mathrm{raw}}$, $\mathrm{DI}_{\mathrm{corrected}}$, bootstrap CI bounds, magnitude CV, mean signature norm, and number of contexts. Full data provenance and repository locations are listed in the Data Availability section.

3 Construct validity: cross-modal concordance

The atlas is useful only if DI reflects a biological property of perturbations rather than a measurement artifact of any single platform. Two pre-registered concordance tests assess this by comparing DI values for the same compounds measured in different datasets. An exploratory analysis using PRISM viability provides independent corroboration from a scalar outcome.

3.1 Cross-modal concordance (LINCS ∩ JUMP-CP)

Compounds present in both LINCS L1000 (transcriptomic DI) and JUMP-CP Cell Painting (morphological DI) were matched by InChIKey connectivity prefix (14 characters). Of the matched set, $n = 378$ compounds have ≥ 5 contexts in both datasets.

Partial Spearman $\rho = 0.190$ (95% CI $[0.090, 0.285]$, $p = 2.1 \times 10^{-4}$), controlling for LINCS n_{contexts} as one covariate. The CI lower bound (0.090) exceeds the 0.05 floor, and $p < 0.0125$. This test passes both pre-registered criteria.

The effect size falls in the "weakly concordant" bin ($0.15 \leq \rho < 0.30$) defined in the pre-registration, consistent with the expectation that cross-modal comparisons capture only the fraction of biology shared between gene expression and cell morphology [8]. The concordance is weak but genuine: compounds whose transcriptomic signatures vary across cell lines also produce variable morphological profiles, and this relationship survives confound control.

3.2 Same-modality concordance (LINCS ∩ Tahoe)

Drugs present in both LINCS L1000 and Tahoe-100M were matched by InChIKey connectivity prefix, verified via PubChem CID cross-reference. This yielded $n = 76$ overlapping compounds (fewer than the pre-registered estimate of $n \approx 103$, because some Tahoe drugs lack PubChem CID

annotations mapping to LINCS InChIKeys). LINCS DI is computed in $\mathbb{R}^{978}$ (landmark genes) while Tahoe DI is computed in $\mathbb{R}^{\sim 62{,}000}$ (full transcriptome). Because cosine similarity is invariant to positive rescaling, DI values are comparable across feature spaces of different dimensionality.

Partial Spearman $\rho = 0.254$ (95% CI $[0.027, 0.456]$, $p = 0.029$), controlling for LINCS and Tahoe n_{contexts} (two covariates). This test **fails**: the CI lower bound (0.027) does not clear the 0.05 floor, and $p = 0.029$ exceeds the Bonferroni threshold ($\alpha = 0.0125$).

The failure is one of statistical power, not absence of concordance. The point estimate ($\rho = 0.254$) is larger than the cross-modal result ($\rho = 0.190$), and the pre-registered power analysis predicted this outcome: at $n = 76$, the study achieves 80% power only for true effects ≥ 0.35. We interpret this pattern as a sample-size artifact—the LINCS-Tahoe overlap ($n = 76$) is far smaller than LINCS-JUMP ($n = 378$)—rather than evidence that DI transfers better across modalities than within one. A deviation in the matching method is disclosed in Section 6.

3.3 PRISM viability (exploratory)

As an independent check on construct validity, we computed the Spearman correlation between LINCS DI and PRISM viability CV for $n = 75$ shared drugs (filtered to $|\text{mean LFC}| \geq 0.01$). Spearman $\rho = 0.359$ (95% CI $[0.144, 0.542]$, $p = 0.0016$).

This result is **exploratory**: n is small, the p-value is uncorrected, and no covariate adjustment was applied. It should carry weight only alongside the primary tests. That said, viability is a scalar outcome with a fundamentally different noise structure from high-dimensional expression or morphology profiles. The correlation suggests that compounds whose transcriptomic effects depend on cellular context also show variable cell-killing potency, consistent with the interpretation that DI captures a biological property of drug-context interactions.

4 What DI is not: informative nulls and the essentiality sign-flip

Five pre-registered experiments (Batches 1 and 2A) tested whether DI variation across genes and drugs can be explained by simpler biological properties. All five fail their pre-registered criteria, and each failure constrains the interpretation of DI in a specific way. Table 2 summarizes the full scorecard including the Batch 3 extensions (Section 5).

4.1 Functional constraint suppresses DI

The pre-registration predicted that broadly essential genes would show high morphological DI: genes required in many cell types should produce context-dependent phenotypic disruption. The observed correlation is large, precisely estimated, and in the opposite direction. Partial Spearman $\rho = -0.332$ (95% CI $[-0.352, -0.312]$, $p \approx 0$, $n = 7{,}653$), controlling for n_{contexts}. Essentiality breadth (fraction of DepMap 24Q4 cell lines with Chronos score < -0.5; 6, 7, 18) predicts *lower* DI, accounting for approximately 11% of variance.

The sign-flip admits a natural interpretation, which we emphasize is *post-hoc*: the pre-registration confused two senses of context-dependence. Essential genes affect many cell types, but they affect them in the same way. A gene required for mitosis, ribosome assembly, or DNA repair produces a stereotyped catastrophe regardless of cellular identity. Functional constraint makes the downstream phenotypic consequences uniform across contexts, yielding low DI. Genes with high DI tend to be selectively essential or non-essential—their knockout phenotypes interact with cell-type-specific pathways, producing variable consequences across the panel.

Table 2: Pre-registered experiment results. Batches 1–2A use partial Spearman ρ with confound control ($\alpha = 0.0125$ for Batch 1, 0.0167 for Batch 2A) except the KO vs. OE test (raw Spearman + permutation null). Batch 3 ($\alpha = 0.0167$, three tests) was prompted by peer review; Experiment F reports partial η^2 instead of ρ. "CI gate" indicates that the 95% CI lower bound did not clear the ± 0.05 threshold in the predicted direction.

Batch	**Hypothesis**	n	**Effect**	p	**Verdict**
1	Cross-modal concordance	378	$\rho = 0.190$	2.1×10^{-4}	**PASS**
1	Same-modality concordance	76	$\rho = 0.254$	0.029	FAIL (CI gate)
1	Essentiality $\rightarrow$ high DI	7,653	$\rho = -0.332$	$\sim$0	FAIL (wrong sign)
1	KO $\approx$ OE DI	5,220	$\rho = 0.011$	0.43	FAIL (null)
2A	Expression variance	7,802	$\rho = 0.060$	1.1×10^{-7}	FAIL (CI gate)
2A	Drug selectivity	75	$\rho = 0.144$	0.22	FAIL (underpowered)
2A	Paralog count	6,894	$\rho = -0.007$	0.57	FAIL (null)
2A	Paralog breadth	6,882	$\rho = -0.029$	0.017	FAIL (CI gate + p)
3	MoA stratification (F)	1,031	$\eta^2 = 0.201$	$<$0.001	**PASS**
3	Cross-cell-line ess. (G)	128	$\rho = -0.264$	0.003	**PASS**
3	DepMap convergent (H)	17,916	$\rho = +0.811$	$\sim$0	FAIL (wrong sign)
Exploratory					
1	Viability concordance	75	$\rho = 0.359$	0.0016	PASS (exploratory)

This interpretation generates a falsifiable prediction: among non-essential genes, those with tissue-restricted expression (active in few cell types) should show *higher* DI than ubiquitously expressed non-essential genes, because tissue-restricted function implies cell-type-specific downstream consequences. We flag this as a candidate for prospective pre-registration.

The result connects DI to established concepts in functional genomics: the phenotypic landscape of essential genes is structured by functional constraint [7, 9, 12, 15], and DI indexes the complement of that constraint.

4.2 DI is mechanism-specific

If context-dependence were an intrinsic property of genes, then a gene's knockout DI and its overexpression DI should correlate. Spearman $\rho = 0.011$ (95% CI $[-0.016, 0.038]$, $p = 0.43$, $n = 5{,}220$ genes with ≥ 5 replicates in both CRISPR and ORF). A mandatory permutation null (1,000 label shuffles preserving plate structure; Appendix B) yielded a null distribution with mean = 0.0003, SD = 0.014, 95th percentile = 0.023. The observed $\rho = 0.011$ falls within the null (permutation $p = 0.215$).

At $n = 5{,}220$, the study had power to detect effects as small as $\rho = 0.04$ (SE ≈ 0.014). The null is well-powered: losing a protein and gaining excess protein engage different molecular mechanisms—loss-of-function versus gain-of-function—and those mechanisms interact with cellular context independently [1, 3].

This result constrains the theoretical scope of DI. A gene is context-dependent in a particular way (knockout, overexpression, pharmacological inhibition), and DI values from one perturbation type carry no information about another. A DI atlas must be stratified by perturbation mechanism.

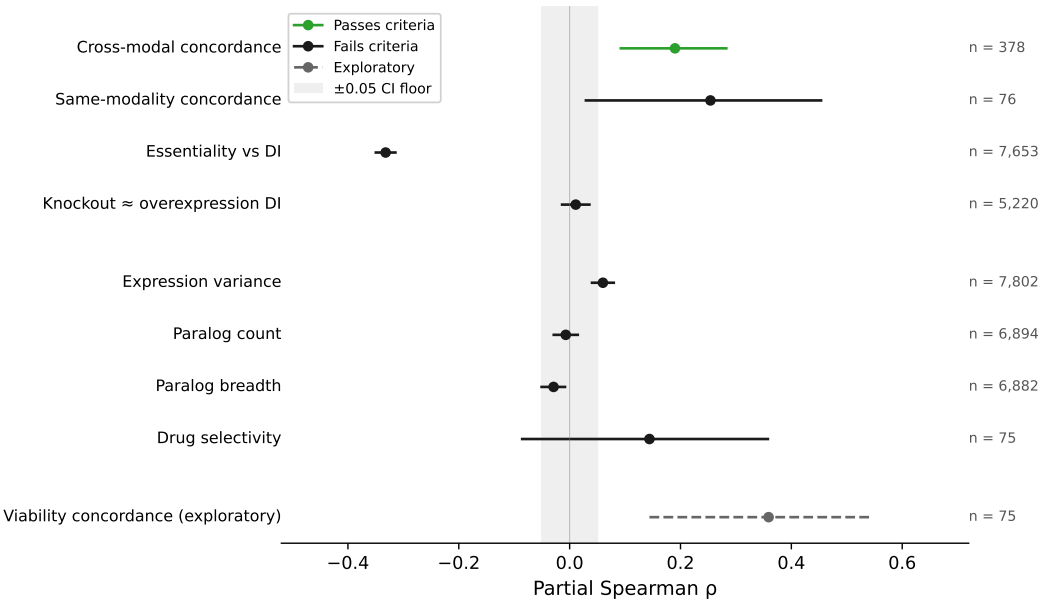

Figure 2: Pre-registered experiment results (Batches 1–2A). Each point shows the partial Spearman ρ (or raw Spearman for the knockout vs. overexpression test) with 95% confidence intervals. The gray band marks the ± 0.05 CI floor: a test passes only if its CI lower bound clears this band in the predicted direction. Same-modality concordance has a larger point estimate ($\rho = 0.25$) than cross-modal ($\rho = 0.19$) but fails because its CI lower bound (0.027) falls inside the floor at $n = 76$. Only cross-modal concordance passes both pre-registered criteria; the seven primary failures delimit what DI measures. Batch 3 results (Experiments F, G, H) are reported in Section 5 and Table 2.

4.3 Proxy explanations fail

Three plausible proxies for context-dependence—expression variance across cell lines, paralog count, and paralog expression breadth—were tested in Batch 2A. All three fail after confound control, and the failures are informative about the structure of the confounds.

Expression variance. Partial Spearman $\rho = 0.060$ (95% CI $[0.038, 0.082]$, $p = 1.1 \times 10^{-7}$, $n = 7{,}802$), controlling for n_{contexts}. The correlation is statistically unambiguous but biologically negligible: $\rho = 0.060$ explains less than 0.4% of variance. The CI lower bound (0.038) does not clear the 0.05 floor. Expression variance captures a real but negligible relationship: genes with variable expression produce marginally more variable knockout phenotypes. The remaining DI variation reflects factors beyond expression variance. Expression data from DepMap 24Q4 (OmicsExpressionProteinCodingGenesTPMLogp1.csv, 1,673 cell lines, 19,193 genes).

Paralog count. Partial Spearman $\rho = -0.007$ (95% CI $[-0.031, 0.017]$, $p = 0.57$, $n = 6{,}894$), controlling for n_{contexts} and mean expression level. The raw Spearman correlation was $\rho = 0.038$ ($p = 0.002$)—a statistically significant positive association. The partial correlation is zero. The raw association existed because highly expressed genes have more paralogs (gene family expansion correlates with expression level) and independently have higher DI. Controlling for expression removes the confound entirely. The significant raw association ($p = 0.002$) evaporates under covariate control. Paralog data from Ensembl BioMart (GRCh38, Ensembl 113).

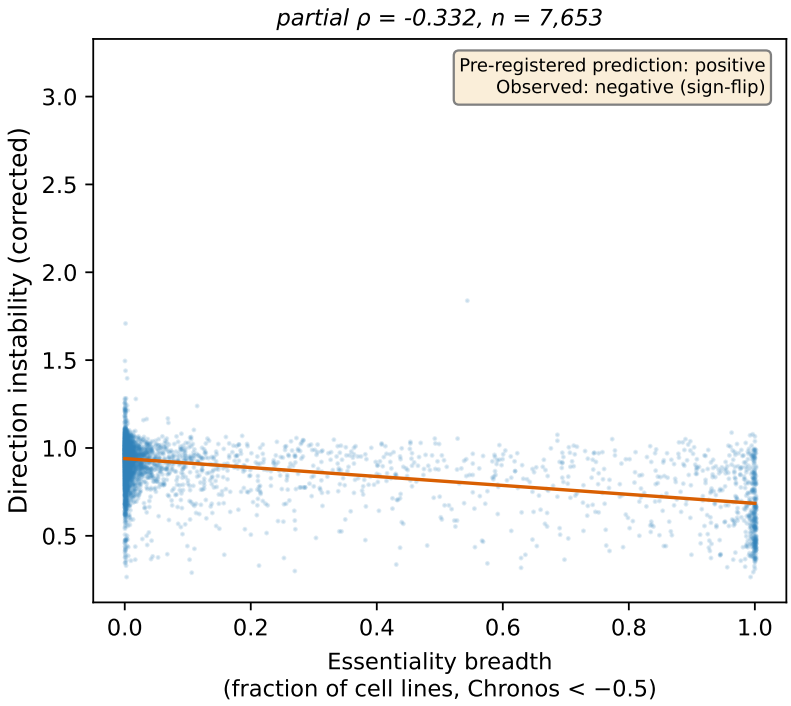

Figure 3: Essentiality breadth (fraction of DepMap cell lines with Chronos < -0.5) vs. magnitude-corrected DI for $n = 7{,}653$ JUMP-CP CRISPR genes. The pre-registration predicted a positive correlation; the observed relationship is negative (partial $\rho = -0.332$ controlling for n_{contexts}). Genes essential across many cell lines produce stereotyped disruption (low DI), while selectively essential or non-essential genes show more context-dependent phenotypes (high DI). Line: OLS fit.

Paralog expression breadth. Among genes with ≥ 1 expressed paralog ($n = 6{,}882$), partial Spearman $\rho = -0.029$ (95% CI $[-0.053, -0.005]$, $p = 0.017$), controlling for n_{contexts} and mean expression level. The sign is correct (broader paralog compensation predicts lower DI, as hypothesized), but $p = 0.017$ exceeds the Bonferroni threshold ($\alpha = 0.0167$) and the CI upper bound (-0.005) does not clear -0.05. Both criteria fail: neither p-value nor CI gate passes. The raw correlation ($\rho = -0.117$, $p \sim 10^{-22}$) is substantially larger, indicating that expression-level confounds inflate the apparent paralog-breadth effect as well.

4.4 Drug selectivity (underpowered)

Partial Spearman $\rho = 0.144$ (95% CI $[-0.088, 0.360]$, $p = 0.22$, $n = 75$), controlling for LINCS n_{contexts}. The primary metric is magnitude-corrected Gini of PRISM kill depth, residualizing raw Gini against mean kill magnitude to remove the mechanical confound between potency and selectivity. The correction reduced the apparent association from $\rho = 0.275$ (raw Gini) to $\rho = 0.144$ (corrected Gini), confirming the magnitude confound is real.

The CI spans $[-0.088, 0.360]$, consistent with anything from no effect to a moderate one. The pre-registration assumed $n \sim 1{,}000$; actual overlap between LINCS drugs and PRISM compounds with $|\text{mean LFC}| \geq 0.01$ is $n = 75$, an order of magnitude smaller. This experiment cannot distinguish a real $\rho = 0.15$ effect from zero.

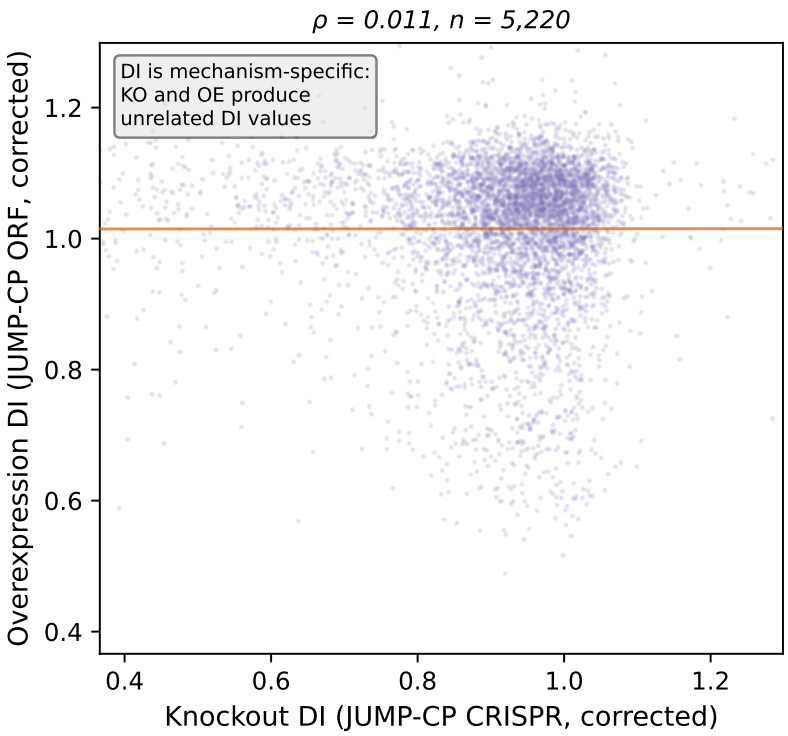

Figure 4: Knockout DI (JUMP-CP CRISPR) vs. overexpression DI (JUMP-CP ORF) for $n = 5{,}220$ genes profiled under both perturbation types. Spearman $\rho = 0.011$ ($p = 0.43$): no relationship. Context-dependence is a property of the perturbation mechanism interacting with cellular context, not a stable property of the target gene.

4.5 Summary of Batches 1–2A

In short: DI variation across genes and drugs is not explained by existing gene annotations, and drug selectivity remains unresolved at $n = 75$. Three reviewer-prompted experiments (Section 5) extend this characterization.

5 Reviewer-prompted extensions (Batch 3)

Batch 3 addresses three concerns raised in peer review of an earlier draft: (F) whether the atlas demonstrates utility beyond negative results, (G) whether the essentiality sign-flip survives on true cross-cell-line morphological data, and (H) whether convergent evidence from a different data modality supports the functional-constraint interpretation. All three experiments were pre-registered at commit SHA `4c34699` before execution, with Bonferroni-corrected $\alpha = 0.0167$ (three tests).

Scoping disclosure. During exploratory scoping for Experiment F, a raw (uncontrolled) one-way ANOVA of DI on MoA class was computed and showed $\eta^2 \approx 0.42$. The raw number was observed before the pre-registration was frozen. The confirmatory test registered here is the confound-controlled, permutation-tested version, whose result was unseen at freeze time. Experiment F is therefore exploratory-then-confirmed, and is disclosed as such.

5.1 Experiment F: DI is mechanism-structured

The Batch 1 and 2A results showed what DI is not explained by. A natural follow-up: does DI vary systematically with drug mechanism-of-action class? If core-machinery inhibitors (targeting fundamental cellular processes) show low DI while receptor-signaling drugs (engaging cell-type-specific pathways) show high DI, then DI captures biologically interpretable structure, and the atlas stratifies drug classes in a useful way.

Because DI depends mechanically on profiling depth (number of cell lines) and MoA classes may differ in profiling depth, the primary analysis uses residualized DI: raw DI residualized on $n_{\text{cell lines}}$ and mean signature norm via OLS. One-way ANOVA of residualized DI on MoA class (80 classes with ≥ 5 drugs, 1,031 drugs total) yields partial $\eta^2 = 0.201$. The raw (uncontrolled) η^2 is 0.417; controlling for profiling depth and signal strength halves the effect, confirming that the confound control is load-bearing.

A 1,000-permutation null (shuffling MoA labels, preserving class sizes) produced a null distribution with mean = 0.077, SD = 0.012, 99th percentile = 0.111. The observed $\eta^2 = 0.201$ exceeds the 99th percentile and the empirical permutation $p < 0.001$. Both pre-registered criteria are met.

The biological pattern is interpretable. The lowest-DI MoA classes are mTOR inhibitors, HDAC inhibitors, HSP inhibitors, and PI3K inhibitors—drugs targeting core cellular machinery [16] that operates uniformly across cell types. The highest-DI classes are serotonin receptor agonists, histamine receptor agonists, GABA receptor antagonists, and PPAR receptor agonists—drugs acting through receptor-mediated signaling pathways whose expression and coupling vary across cell lineages.

This result demonstrates that DI captures mechanism-level structure in drug pharmacology and that the atlas provides a resource for stratifying drug classes by their cross-context consistency.

5.2 Experiment G: cross-cell-line essentiality replication

The Batch 1 essentiality sign-flip ($\rho = -0.33$, $n = 7{,}653$) was computed on plate-replicate DI within a single cell line (U2OS). Plate replicates measure technical reproducibility, and reviewers asked whether the sign-flip reflects biology or plate-level noise.

The cpg0000 pilot dataset (Cell Painting Gallery) provides CRISPR profiles from two cell lines: A549 (10 plates) and U2OS (8 plates). With $K = 2$ cell lines, DI degenerates to a single pairwise cosine distance: $\text{DI} = 1 - \cos(\mathbf{s}_{\text{A549}}, \mathbf{s}_{\text{U2OS}})$. This is geometrically valid but extremely noisy per gene (no averaging over pairs). Cell line assignments were obtained from the companion metadata repository [3].

Of 128 genes with profiles in both cell lines, Spearman $\rho = -0.264$ (95% CI $[-0.418, -0.095]$, $p = 0.003$, $n = 128$). The correlation is negative, statistically significant at $\alpha = 0.0167$, and the CI excludes zero. The pre-registered verdict is **directional replication**: the essentiality sign-flip observed on plate replicates holds in direction on true cross-cell-line data, narrowing the interpretation from technical noise to biological context-dependence.

The point estimate ($\rho = -0.26$) is smaller than Batch 1 ($\rho = -0.33$), consistent with the expected attenuation from $K = 2$ noise. The CI is wide ($[-0.42, -0.10]$), reflecting the small sample size ($n = 128$ vs. Batch 1's $n = 7{,}653$) and the degeneracy of single-cosine DI. This experiment narrows the interpretation of the sign-flip; it does not close it. Properly powered cross-cell-line replication requires datasets with $K \geq 5$ cell lines per gene, such as future expansions of JUMP or single-cell CRISPR screens [19].

5.3 Experiment H: DepMap convergent evidence fails

The biological interpretation of the essentiality sign-flip—functional constraint suppresses context-dependent phenotypes—predicts that the same pattern should appear in dependency scores: broadly essential genes should show uniform dependency across cell lines (low variability), while selectively essential genes should show variable dependency (high variability). Testing this on DepMap Chronos scores provides convergent evidence from a fundamentally different measurement.

Partial Spearman $\rho = +0.811$ (95% CI $[0.806, 0.816]$, $p \approx 0$, $n = 17{,}916$), controlling for the number of non-missing Chronos values per gene. The pre-registered prediction was negative ($\rho < -0.05$); the observed correlation is strongly positive. This test fails.

We hypothesize the failure reflects a construct divergence rather than a biological contradiction. DepMap essentiality breadth (fraction of cell lines with Chronos < -0.5) and dependency-score SD are different geometric constructs from morphological DI. Morphological DI measures pairwise cosine distance between high-dimensional vectors; dependency-score SD measures the spread of scalar values. A gene essential in many cell lines has a wider range of Chronos scores (some near -1, others near -0.5), producing high SD. The positive correlation may reflect this scalar-range property: broadly essential genes show high dependency-score variability because the Chronos scale admits more spread as more cell lines cross the essentiality threshold. Consistent with this interpretation, dependency-score SD correlates with Chronos score range at $\rho = 0.83$: SD is largely a proxy for the mechanical range of scores a gene spans, which widens as more cell lines cross the essentiality threshold ($\rho = 0.79$ between score range and breadth).

The failure exposes a construct divergence: the functional-constraint hypothesis makes different predictions depending on whether context-dependence is measured as vector cosine distance (morphological DI) or scalar variability (dependency-score SD). The morphological sign-flip stands; the convergent pathway through DepMap SD does not replicate it. A convergent test using a vector-valued DepMap representation (such as gene-expression signatures of knockout effects across cell lines) would avoid this scalar artifact.

6 Pre-registration integrity

All hypotheses, decision criteria, statistical methods, and analysis code were frozen before execution. Batch 1 (concordance, essentiality, knockout vs. overexpression, PRISM) was frozen at commit SHA `0d07a01`; Batch 2A (expression variance, drug selectivity, paralog count, paralog breadth) at commit SHA `a17c125`; Batch 3 (MoA stratification, cross-cell-line CRISPR, DepMap convergent) at commit SHA `4c34699`. The pre-registration documents specify the statistical method, Bonferroni correction levels, decision criteria, power analyses, and interpretation guides for each outcome. All three documents are included in the repository.

Batch 3 was prompted by peer review of an earlier draft (desk-rejected from GigaScience for scope, GIGA-D-26-00229, 2026-07-08). The experiments are pre-registered before execution but are reviewer-prompted, post-hoc additions—timestamps are verifiable against the repository history. Experiment F includes a scoping disclosure: the raw $\eta^2 \approx 0.42$ was observed during exploratory scoping before the confirmatory (confound-controlled, permutation-tested) analysis was frozen. Experiment G's analysis logic is byte-identical between the frozen commit and the executed version; only the data download function was modified post-freeze to correct S3 paths (the original paths returned 404 errors). This deviation is documented in the repository.

6.1 The CI floor

At sample sizes above $n = 5{,}000$, trivially small effects reach statistical significance. The pre-registration introduced a uniform CI lower-bound gate ($|\text{CI bound}| > 0.05$ in the predicted direction) to prevent this. The gate blocked expression variance ($\rho = 0.060$), which would have passed on p-value alone. Paralog breadth ($\rho = -0.029$) fails on both criteria: its $p = 0.017$ marginally exceeds the Bonferroni threshold and its CI does not clear the floor. In both cases, the effects explain less than 0.4% of variance and lack biological plausibility as primary explanations for DI variation.

6.2 Deviations

Seven deviations from the pre-registration occurred during analysis. All are data-wrangling corrections or sample-size discrepancies; none alter the statistical method, decision criteria, or interpretation framework.

DepMap version. All DepMap analyses use the pre-registered 24Q4 release (figshare article 27993248): Chronos dependency scores (file 51064667; 1,178 cell lines, 17,916 genes) for the essentiality test, expression (file 51065489; 1,673 cell lines, 19,193 genes) for expression variance and paralog tests. Initial runs inadvertently used 22Q2 Chronos (essentiality) and 24Q2 expression (expression variance, paralogs). Re-running on the correct 24Q4 release changed no verdicts: the essentiality test moved from $\rho = -0.330$ to -0.332, expression variance from 0.064 to 0.060, and paralog count/breadth were identical to three decimal places.

Same-modality concordance matching method. The pre-registration specified InChIKey matching for drug concordance. An initial run erroneously used case-insensitive drug-name matching ($n = 114$, $\rho = -0.003$), which introduced approximately 38 spurious pairs from salt forms, stereoisomers, and name collisions. Reverting to InChIKey matching via PubChem CID cross-reference ($n = 76$, $\rho = 0.254$) moved toward the pre-registered method, resolving the spurious null. The InChIKey result is the pre-registered analysis; the name-matched result is documented in Appendix C.

Batches 1–2A deviations. (i) Cross-modal concordance yielded $n = 378$ instead of the estimated $n = 330$, because `pert_iname` matching recovered more valid InChIKey pairs than anticipated. (ii) `pert_iname` was used instead of `pert_id` for LINCS metadata lookup; `pert_id` matching was a data-wrangling error yielding only 8 compounds. (iii) Context definitions differ across datasets (cell lines vs. plate replicates); see Section 2 and Section 7. (iv) Drug selectivity yielded $n = 75$ instead of the estimated $n \sim 1{,}000$ due to the small LINCS-PRISM overlap after viability filtering. (v) Nine genes were dropped from the paralog tests due to missing expression values, leaving $n = 6{,}894$ (paralog count) and $n = 6{,}882$ (paralog breadth).

Batch 3 deviations. (i) Experiment F: The scoping disclosure (raw η^2 observed before freeze) is described above. The confirmatory analysis (confound-controlled, permutation-tested) was unseen at freeze time. (ii) Experiment G: The pre-registration estimated $n \approx 160$ shared genes; the actual overlap with DepMap is $n = 128$ (deterministic: genes present in both cell lines with non-zero signature norms). The data download function was modified post-freeze to use the correct S3 paths; the analysis logic (DI computation, correlation, CI) is byte-identical to the frozen version. (iii) After obtaining the Experiment G result, bootstrap and timepoint-stratified sensitivity analyses were run

exploratorily. These were not pre-registered and are excluded from the manuscript (documented in `DEVIATIONS.md`).

7 Discussion

7.1 What DI measures—and what it does not

Eleven pre-registered experiments produce a consistent characterization of direction instability. DI is concordant across measurement modalities, confirming that it reflects a biological property of perturbation-context interactions. It is mechanism-specific: knockout and overexpression of the same gene produce unrelated DI profiles, so DI characterizes how a perturbation mechanism interacts with cellular context, not a stable property of the target gene. Functional constraint suppresses DI: essential genes produce stereotyped disruption, yielding low DI—a sign-flip that directionally replicates on true cross-cell-line CRISPR data ($\rho = -0.26$, cpg0000 pilot). Three plausible proxy explanations—expression variance, paralog count, paralog breadth—each explain less than 0.4% of DI variance after confound control.

DI is also mechanism-structured at the pharmacological level: drug MoA classes explain 20% of residualized DI variance, with an interpretable pattern distinguishing core-machinery inhibitors (low DI) from receptor-signaling drugs (high DI). This result demonstrates that the atlas provides a resource for stratifying drug classes by cross-context consistency, a property complementary to traditional selectivity metrics.

The DepMap convergent test exposes a boundary of the functional-constraint interpretation. Dependency-score SD and essentiality breadth correlate positively ($\rho = +0.81$), opposite to the morphological sign-flip, because the scalar SD estimand mechanically inflates with the Chronos score range. The functional-constraint hypothesis makes different predictions depending on whether context-dependence is measured as vector cosine distance or scalar variability. The morphological sign-flip is about directional consistency of high-dimensional profiles; DepMap SD indexes scalar spread. Future convergent tests should use vector-valued representations of dependency effects.

This combination of positive, negative, and failed results shows that DI captures how perturbation mechanisms interact with cell-type-specific biology. It is orthogonal to essentiality, expression variance, and paralog structure. Other untested features such as protein interaction network topology or chromatin accessibility might correlate with DI; the present results constrain the space of explanations without exhausting it. For atlas users, the practical implication is that DI provides information beyond what is available from essentiality screens, expression atlases, or paralog databases.

7.2 The resource

The atlas provides magnitude-corrected DI values with bootstrap confidence intervals for 8,949 LINCS drugs, 379 Tahoe drugs, 7,946 CRISPR knockouts, and 12,590 ORF overexpressions. The companion JUMP-CP compound DI values (25,254 compounds) are available from a prior study [20]. Together, these cover the perturbation-biology datasets most widely used for drug mechanism prediction, gene function annotation, and virtual screening benchmarks.

The atlas is designed for downstream integration. Each row provides the columns needed for a secondary analysis: corrected DI, raw DI, CI bounds, mean signature norm, magnitude CV, and number of contexts. Pre-registration documents, analysis scripts, and result JSON files are released alongside the data.

7.3 Limitations

Statistical power. The PRISM viability analysis ($\rho = 0.359$, $n = 75$) is exploratory: uncorrected p-value, no covariate adjustment, small sample. It provides corroboration only. The same-modality concordance test (LINCS-Tahoe, $n = 76$) was underpowered by design: the LINCS-Tahoe compound overlap is structurally small. The point estimate ($\rho = 0.254$) is consistent with genuine concordance, but the wide CI ($[0.027, 0.456]$) precludes a firm conclusion. The drug selectivity test is underpowered at $n = 75$ and cannot distinguish a moderate effect from zero. Both underpowered tests are data-limited: the small n reflects the structural size of the dataset intersection, not a design choice. We report them for completeness and for the atlas resource record, not as adjudicated hypotheses. Experiment G computes DI from $K = 2$ cell lines, the minimum for a pairwise comparison. With $K = 2$, DI is a single cosine distance rather than an average over pairs, making it extremely noisy per gene. No magnitude correction was applied ($n = 128$ is too small for meaningful OLS), and no bootstrap CIs are possible. The A549 plates vary in timepoint (96h vs. 144h) and antibiotic treatment, introducing uncontrolled confounds. The experiment's value is as a directional check, not a definitive test.

Construct and interpretation. Context definitions vary across datasets. LINCS and Tahoe use cell lines as contexts; JUMP-CP CRISPR and ORF use plate replicates. Plate replicates sample technical variability within one cell line, while cell lines sample biological variability across cell types. DI computed over plate replicates measures reproducibility within a fixed biological context, whereas DI over cell lines measures context-dependence across biological contexts. Cross-dataset comparisons match on context type, but the CRISPR/ORF DI values used in the essentiality, mechanism-specificity, and proxy tests index technical variability rather than biological context-dependence. Extending the atlas to single-cell CRISPR screens with cell-line diversity [19] would resolve this asymmetry. The essentiality sign-flip interpretation (functional constraint suppresses DI) is post-hoc. The pre-registration predicted the opposite sign. The mechanistic explanation is plausible, connects to established results in functional genomics [7, 15], and is independently corroborated by a separate study that found universal-essential knockdowns diverge rather than converge across cell types in Perturb-seq data ($d = 0.60$, $p_{\text{perm}} < 10^{-4}$; 21); it was, however, generated after observing the data. Prospective testing of a falsifiable prediction—among non-essential genes, tissue-restricted expression should positively predict DI (Section 4)—would convert this post-hoc rationalization into a validated mechanism. Experiment F is exploratory-then-confirmed (scoping disclosure, Section 5); the confound-controlled $\eta^2 = 0.20$ is the primary result.

Experiment H exposes a construct divergence: scalar dependency-score SD and vector cosine DI measure different geometric properties of context-dependence. The positive $\rho = +0.81$ does not falsify the morphological sign-flip; it shows that the functional-constraint hypothesis generates different predictions under different operationalizations.

Data scope. LINCS L1000 has well-documented batch effects across production sites and plates [17]. If a drug was profiled across batches, batch-driven direction differences would inflate DI independently of biology. The companion study's split-half stability analysis provides indirect reassurance, but a direct test of whether DI correlates with batch structure has not been performed. All DI values are computed at the dose used in each screening campaign. Dose-response effects could change a drug's DI: a compound selective at low dose might become promiscuous at high dose, altering its context-dependence. The atlas does not capture this dose-dependent structure.

7.4 Future directions

Four extensions follow naturally from the current results. First, Batch 2B will test whether DI correlates with Shesha geometric stability [13, 14], a concurrent and independently developed measure of perturbation coherence computed on single-cell data. Second, Gene Ontology enrichment analysis on the directional divergence gene sets from the mechanism-specificity test (1,331 knockout-stable / overexpression-unstable genes; 289 with the opposite pattern) may reveal functional categories associated with mechanism-specific context-dependence. Third, expanding the Tahoe drug-InChIKey annotation would provide a properly powered same-modality concordance test. Fourth, a properly powered cross-cell-line CRISPR replication ($K \geq 5$ cell lines per gene) would resolve the essentiality sign-flip interpretation definitively; future expansions of JUMP or single-cell CRISPR screens [19] may provide such data.

More broadly, the failure of tested gene-level annotations to explain DI, combined with the success of pharmacological MoA class in structuring it, suggests that DI indexes an aspect of perturbation biology—the interaction between mechanism and cellular context—that gene-level annotations do not capture. Whether this interaction can be predicted from protein interaction networks, pathway topology, or chromatin accessibility remains an open question.

7.5 Conclusions

Eleven pre-registered experiments across three batches characterize direction instability as a metric that is concordant across measurement modalities, mechanism-specific, shaped by functional constraint, and structured by pharmacological mechanism-of-action class. DI is orthogonal to expression variance, paralog count, and paralog breadth. The essentiality sign-flip—functional constraint suppresses DI—directionally replicates on cross-cell-line CRISPR data. A convergent test on DepMap dependency-score variability fails, exposing a construct divergence between scalar SD and vector cosine DI that bounds the generality of the functional-constraint interpretation. The atlas provides magnitude-corrected DI values with bootstrap confidence intervals for 8,949 drugs, 379 single-cell-profiled drugs, 7,946 CRISPR knockouts, and 12,590 ORF overexpressions, ready for integration with drug-mechanism prediction, gene-function annotation, and screening quality control.

Materials and methods

Data and code availability

All analysis scripts are written in Python 3 with standard scientific libraries (NumPy, SciPy, pandas, scikit-learn) and are available at `https://github.com/elliottower/direction-instability-atlas` under the MIT License.

All atlas data files (per-perturbation DI values with bootstrap CIs for each of the five datasets), experiment result files, pre-registration documents, and analysis code are deposited on Zenodo (DOI: 10.5281/zenodo.21223667) under a CC-BY-4.0 license. Pre-registration commits can be verified against the repository history at the SHA values reported in the text. LINCS L1000 compound DI values (8,949 drugs) and JUMP-CP compound DI values (25,254 compounds) were previously computed and are available from the companion study (DOI: 10.5281/zenodo.21213699).

Source datasets are available from their original repositories: LINCS L1000 (GEO: GSE92742), Tahoe-100M (HuggingFace: tahoebio/Tahoe-100M), JUMP Cell Painting (Cell Painting Gallery, cpg0016), DepMap 24Q4 (figshare: 27993248), PRISM (figshare: Repurposing Public 24Q2), Ensembl paralogs (BioMart, GRCh38).

Method details

DI computation, magnitude correction, and bootstrap confidence intervals are described in Section 2. Pre-registration protocol and deviations are documented in Section 6. All analysis scripts reproduce the results from source data; scripts and frozen pre-registration documents are included in the Zenodo deposit.

Quantification and statistical analysis

All hypothesis tests use partial Spearman correlation computed by Pearson-residualizing both variables against the covariate matrix (with intercept), then computing Spearman's ρ on the residuals (Section 2). Confidence intervals use the Fisher z-transform with degrees of freedom adjusted for the number of covariates. Bonferroni correction is applied within each batch ($\alpha = 0.0125$ for Batch 1's four primary tests; $\alpha = 0.0167$ for Batch 2A and Batch 3, three tests each). A uniform CI lower-bound gate (± 0.05) prevents trivially small effects from passing at large n. Permutation nulls (1,000 shuffles) are used for the knockout vs. overexpression test and the MoA stratification test. Experiment F uses one-way ANOVA with partial η^2 on residualized DI. All statistical analyses were performed in Python 3.11 using SciPy 1.12, NumPy 1.26, pandas 2.2, and scikit-learn 1.4.

Acknowledgements

The author used Claude Code (Anthropic, `https://claude.ai/code`) to assist with drafting and editing manuscript text and with writing and debugging analysis code. Perplexity (`https://perplexity.ai`) was used as an independent reviewer of the near-final draft. The author reviewed and verified all outputs, takes full responsibility for the accuracy and integrity of the work, and confirms that all hypotheses, pre-registered predictions, interpretations, and conclusions are the author's own. No AI tool is listed as an author, and no AI-generated content was used as a primary data source.

Author contributions

E.T.: Conceptualization, Methodology, Software, Formal analysis, Investigation, Data curation, Writing – original draft, Writing – review & editing, Visualization.

Declaration of interests

The author declares no competing interests. This work was conducted independently with no external funding.

References

[1] Badonyi, M., Marsh, J.A. (2025). Prevalence of loss-of-function, gain-of-function and dominant-negative mechanisms across genetic disease phenotypes. *Nature Communications*, 16:8392.

[2] Bray, M.-A., et al. (2016). Cell Painting, a high-content image-based assay for morphological profiling using multiplexed fluorescent dyes. *Nature Protocols*, 11:1757–1774.

[3] Chandrasekaran, S.N., et al. (2024). Three million images and morphological profiles of cells treated with matched chemical and genetic perturbations. *Nature Methods*, 21:1114–1121.

[4] Chandrasekaran, S.N., et al. (2025). Morphological map of under- and overexpression of genes in human cells. *Nature Methods*.

[5] Corsello, S.M., et al. (2020). Discovering the anticancer potential of non-oncology drugs by systematic viability profiling. *Nature Cancer*, 1:235–248.

[6] Dempster, J.M., et al. (2021). Chronos: a cell population dynamics model of CRISPR experiments that improves inference of gene fitness effects. *Genome Biology*, 22:343.

[7] Funk, L., et al. (2022). The phenotypic landscape of essential human genes. *Cell*, 185(24):4634–4653.

[8] Haghighi, M., et al. (2022). High-dimensional gene expression and morphology profiles of cells across 28,000 genetic and chemical perturbations. *Nature Methods*, 19:1550–1557.

[9] Hart, T., et al. (2015). High-resolution CRISPR screens reveal fitness genes and genotype-specific cancer liabilities. *Cell*, 163(6):1515–1526.

[10] Lamb, J., et al. (2006). The Connectivity Map: using gene-expression signatures to connect small molecules, genes, and disease. *Science*, 313(5795):1929–1935.

[11] McFarland, J.M., et al. (2020). Multiplexed single-cell transcriptional response profiling to define cancer vulnerabilities and therapeutic mechanism of action. *Nature Communications*, 11:4296.

[12] Meyers, R.M., et al. (2017). Computational correction of copy number effect improves specificity of CRISPR–Cas9 essentiality screens in cancer cells. *Nature Genetics*, 49:1779–1784.

[13] Raju, P.C. (2026). From syntax to semantics: Geometric stability as the missing axis of perturbation biology. *arXiv*, 2603.00678.

[14] Raju, P.C. (2026). Geometric coherence of single-cell CRISPR perturbations reveals regulatory architecture and predicts cellular stress. *arXiv*, 2604.16642.

[15] Ramezani, M., et al. (2025). A genome-wide atlas of human cell morphology. *Nature Methods*, 22:621–633.

[16] Santos, R., et al. (2017). A comprehensive map of molecular drug targets. *Nature Reviews Drug Discovery*, 16:19–34.

[17] Subramanian, A., et al. (2017). A next generation connectivity map: L1000 platform and the first 1,000,000 profiles. *Cell*, 171(6):1437–1452.

[18] Tsherniak, A., et al. (2017). Defining a cancer dependency map. *Cell*, 170(3):564–576.

[19] Tahoe Therapeutics (2025). Tahoe-100M: A giga-scale single-cell perturbation atlas for context-dependent gene function and cellular modeling. *bioRxiv*, 2025.02.20.639398.

[20] Tower, E. (2026). Direction instability tracks drug mechanism transport while geometric alternatives measure magnitude. Zenodo. `doi:10.5281/zenodo.21213699`.

[21] Tower, E. (2026). Construct validity for geometric perturbation scores: when does a number license a mechanistic claim? Zenodo. `doi:10.5281/zenodo.21284088`.

A Pre-registration protocol summary

All hypotheses, decision criteria, statistical methods, and analysis code were frozen before execution. The three pre-registration documents are included in the Zenodo deposit.

Table 3: Pre-registration batches.

Batch	Commit SHA	Experiments
Batch 1	`0d07a01`	A2 (LINCS–JUMP cross-modal concordance), A1 (LINCS–Tahoe same-modality concordance), C (essentiality), D (KO vs. OE), PRISM (exploratory)
Batch 2A	`a17c125`	E2 (expression variance), E3 (drug selectivity), E5a (paralog count), E5b (paralog breadth)
Batch 3	`4c34699`	F (MoA stratification), G (cross-cell-line CRISPR essentiality), H (DepMap convergent evidence)

Two Batch 1 experiments were dropped before analysis: A3 (Tahoe–JUMP concordance; zero compound overlap between clinical drugs and screening library) and B (MOA stratification in Tahoe; structurally underpowered at $n = 7$ vs. $n = 17$). Batch 2A labels E2/E3/E5 are non-consecutive because the pre-registration proposed three specific explanatory hypotheses, not a sequential series. Batch 3 was prompted by peer review and frozen at a separate SHA with its own $\alpha = 0.0167$ (three tests).

What did not change across batches. The DI computation (Equation 1), the magnitude correction (Equation 2), the partial Spearman method (Pearson-residualize then rank), the CI floor gate (± 0.05), and the Bonferroni correction framework were constant across all three batches. Batch 2A and Batch 3 each added hypotheses with their own $\alpha = 0.0167$ (Bonferroni for three tests each).

B Knockout vs. overexpression permutation null

The mandatory permutation null for the knockout vs. overexpression DI correlation used 1,000 label shuffles preserving plate structure within each dataset. The null distribution had mean = 0.0003, SD = 0.014, and 95th percentile = 0.023. The observed $\rho = 0.011$ falls at permutation $p = 0.215$, well within the null. At $n = 5{,}220$, the study had power to detect effects as small as $\rho = 0.04$ (SE ≈ 0.014), so the null result is well-powered.

C Same-modality concordance matching method comparison

The pre-registration specified InChIKey matching for drug concordance. An initial run erroneously used case-insensitive drug-name matching, which inflated the sample size and introduced spurious pairs.

Reverting to InChIKey matching resolved the spurious null. The name-matched result ($\rho = -0.003$) reflects contamination by chemically distinct compounds sharing common drug names, not absence of concordance.

Table 4: Same-modality concordance results by matching method. InChIKey matching is the pre-registered analysis.

Method	Match type	n	Partial ρ	Note
Drug-name (erroneous)	Case-insensitive	114	-0.003	$\sim$38 spurious pairs from salt forms, stereoisomers
InChIKey (pre-registered)	Connectivity prefix	76	0.254	Via PubChem CID cross-reference

D Magnitude correction details

The OLS magnitude correction (Equation 2) is applied independently per dataset. For each dataset, raw DI is regressed on mean signature L_2 norm. The residuals plus the intercept yield $\mathrm{DI}_{\mathrm{corrected}}$.

Table 5: Magnitude correction parameters per dataset. Negative slopes indicate that weaker perturbations (lower norm) have higher raw DI, as expected from the noise-inflation mechanism.

Dataset	n	OLS slope	R^2
Tahoe-100M	379	$+0.00013$	0.002
JUMP-CP CRISPR	7,946	-0.0097	0.081
JUMP-CP ORF	12,590	-0.0126	0.182

LINCS L1000 and JUMP-CP compound DI values were corrected in the companion study [20] and are distributed with pre-corrected values. The near-zero Tahoe slope reflects Tahoe-100M's high signal-to-noise ratio (deeply sequenced single-cell data); the correction has negligible effect on this dataset.