COMPASS: Component-Wise Inference of Shared and Gene-Specific Perturbation Response

Huan Liang[1] and Rohit Singh[1,2,†]
[1]*Department of Biostatistics and Bioinformatics, Duke University, Durham, NC, USA*
[2]*Department of Cell Biology, Duke University, Durham, NC, USA*
[†]*E-mail: rohit.singh@duke.edu*

Predicting how a genetic perturbation reshapes a cell's transcriptome is a central goal of computational biology. Previous studies report that the mean response across training perturbations rivals specialized models on standard accuracy metrics, even though it cannot distinguish which perturbation occurred. Across 2,270 CRISPRi perturbations measured in each of six cell lines, we show that this apparent paradox reflects a conserved organization of perturbation responses. Perturbations span a continuum from responses strongly aligned with the mean to more targeted responses that depart from it. Crucially, a perturbation's position along this continuum is conserved across cell lines (Kendall's $W = 0.59$) and predictable from STRING protein-interaction embeddings ($R^2 = 0.35$). We formalize this structure with COMPASS, an interpretable linear model that decomposes each response into shared and gene-specific components and estimates them separately. The shared-response component is modeled as a cell-line-wide response scaled by a perturbation-specific coefficient. This coefficient is strongly conserved across cell lines. The residual gene-specific component—which is moderately conserved across cell lines—recovers pathway-level programs. COMPASS outperforms scGPT, CPA, GEARS, GenePert, and State in both response accuracy (de-biased Pearson delta 0.34 vs. ≤ 0.32) and perturbation discrimination (cosine PDS gain 0.23 vs. ≤ 0.08). These results recast perturbation prediction across cellular contexts as component-wise inference, with each component estimated from the evidence best suited to it.

Keywords: Perturb-seq; genetic perturbation prediction; mean baseline; gene embeddings; response decomposition; interpretable models.

1. Introduction

One of the long-standing puzzles in perturbation prediction is the strength of simple baselines. For gene knockdowns measured by Perturb-seq, multiple benchmarking efforts[1,2] report that the mean response across training perturbations in the same cell line consistently rivals or outperforms purpose-built models such as GEARS and CPA, single-cell foundation models such as scGPT, and architectures trained across diverse experimental contexts.[3–6] Systema further showed that much of the apparent performance of these models arises from a systematic expression shift shared across perturbations, rather than from recovery of the response specific

© 2026 The Authors. Open Access chapter published by World Scientific Publishing Company and distributed under the terms of the Creative Commons Attribution Non-Commercial (CC BY-NC) 4.0 License.

to the perturbed gene.[7]

While these observations can be read as a failure of model architectures, they also admit structural explanations. The first is about metrics: the mean baseline is *accurate*, recovering much of the measured perturbation effect, but it is ineffective at *discriminating* one perturbation from another. While early studies prioritized accuracy, discrimination is necessary for a complete assessment. The second is about biology: distinct knockdowns may converge on a common response direction because they perturb overlapping cellular programs. Such a shared response would explain the mean baseline's high accuracy and provide a scaffold for isolating perturbation-specific effects.

If such shared-response structure exists, individual perturbations should differ systematically in how strongly they engage the shared response. Knockdowns of core growth machinery are natural candidates for broad, stereotyped responses in which large numbers of genes shift together. Perturbations of specific regulatory machinery may instead produce responses that depart from the shared direction. To test this hypothesis, we characterize each perturbation p in cell line c using its response magnitude $m_{cp} \in \mathbb{R}$ and its alignment $a_{cp} \in \mathbb{R}$ with the cell line's mean perturbation direction. We explored how m and a are distributed and whether this organization is conserved across cell lines.

Across 2,270 perturbations shared by six cell lines from three independent CRISPRi studies,[8–10] we find that magnitude and alignment trace a similar continuum in every cell line. Strikingly, the same perturbations tend to occupy similar positions across studies, and their representations built from prior knowledge, particularly STRING network embeddings, predict those positions. This conservation motivates the decomposition $z_{cp} = \beta_p u_c + g_p + \varepsilon_{cp}$, where $u_c \in \mathbb{R}^G$ is the cell line's shared-response direction, $\beta_p \in \mathbb{R}$ measures how strongly perturbation p engages that direction, $g_p \in \mathbb{R}^G$ is the gene-specific residual response, and $\varepsilon_{cp} \in \mathbb{R}^G$ is the deviation specific to perturbation p in cell line c. Across cell lines, we compute $\bar{\beta}_p$ and g_p by averaging the per-cell-line coefficients and residuals, respectively. This converts the empirical observation of a perturbation continuum into two quantities with different conservation properties: $\bar{\beta}_p$ is strongly conserved and transferable across cell lines, while g_p is less well conserved but still retains meaningful cross-cell structure. Both quantities also carry substantial independent biological information.

The two components can therefore be estimated from different evidence and interpreted independently. This motivates COMPASS, an interpretable linear model that estimates these components under within-cell-line and cross-cell-line regimes and combines them to predict unseen perturbation responses. Across six cell lines, COMPASS substantially outperforms direct perturbation-response models and approaches the accuracy of the strongest fixed-representation predictor, TabICL, while matching or exceeding its cross-cell discrimination.

2. Data and Exploratory Analysis

2.1. *Data and evaluation contexts*

We analyzed six pooled CRISPRi Perturb-seq datasets from three independent studies: K562 and RPE1 from Replogle *et al.*, HepG2 and Jurkat from Nadig *et al.*, and HCT116 and HEK293T from X-Atlas/Orion.[8–10] After applying each study's guide-assignment and cell-

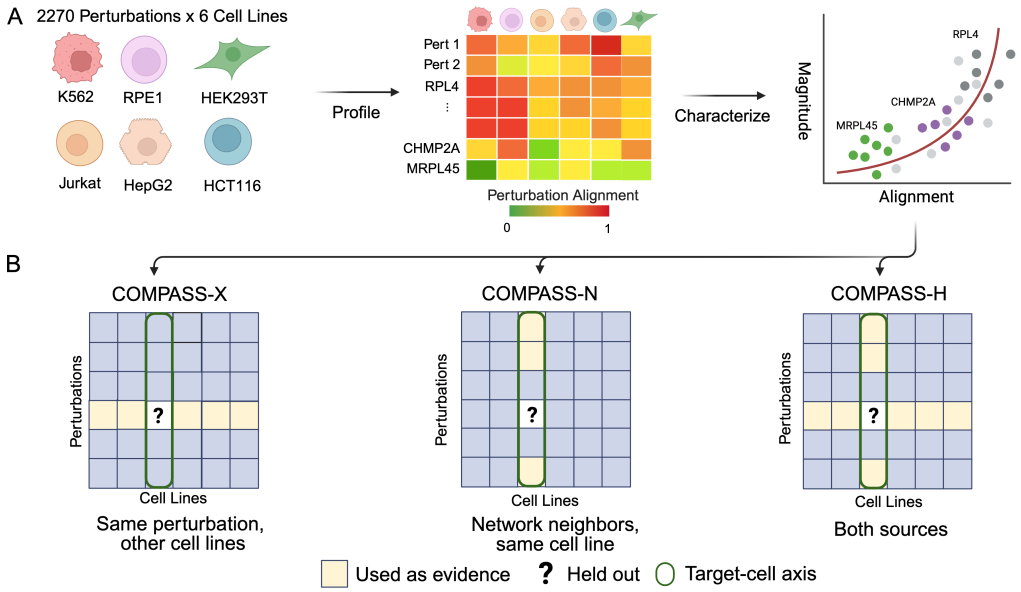


Fig. 1. **Study overview. (A)** We profiled 2,270 CRISPRi perturbations shared across six cell lines and characterized each perturbation by its response magnitude and its alignment with the cell-line average response. **(B)** COMPASS predicts unseen perturbations under three evidence regimes: COMPASS-X uses the same perturbation in other cell lines, COMPASS-N uses network-neighbor perturbations in the target cell line, and COMPASS-H combines both.

quality filters, we intersected the targets measured in all six lines to obtain 2,270 shared perturbations, predominantly essential genes from the Cancer Dependency Map project.[11,12] The shared perturbation set therefore supports matched cross-cell analysis but is enriched for genes with strong fitness effects, a limitation we return to in the Discussion. Counts were library-size normalized to 10,000 per cell and log-transformed as $\log(1 + x)$.

Predictions were scored on panels of 2,000 highly variable genes. We used two kinds of panels: cell-line-specific panels, selected independently for each line, and dataset-specific panels, shared across cell lines within a study group. Because the Replogle/Nadig (four cell lines) and Orion (two lines) dataset-specific panels overlap in only 282 readout genes, we kept our analysis dataset-specific; cross-cell quantities were estimated only within these groups. We evaluated predictions in two settings that correspond to different experimental situations. In the *within-cell-line* setting, a model observes a subset of perturbations in the query cell line and predicts held-out perturbations in the same line. This corresponds to extending a partially completed screen. In the *cross-cell-line* setting, the query perturbation has been measured in other cell lines but is held out from the target line. This corresponds to transferring knowledge from existing screens to a new cellular context. Extensive data processing details can be found in Supplementary Sections 6.1 and 6.2.

2.2. *Evaluation metrics: accuracy and discrimination*

A perturbation predictor must do two separable things: recover what the response looks like (*accuracy*) and distinguish one perturbation's response from another (*discrimination*). Measuring them requires distinct metrics.

For perturbation p, cell line c, and readout gene $g \in \mathcal{G}$ with $|\mathcal{G}| = G$, we define the pseudobulk perturbation effect

$$z_{cp,g} = \bar{x}_{cp,g} - \bar{x}_{c0,g}, \tag{1}$$

where $\bar{x}_{c0,g}$ is the mean among non-targeting controls. A predictive model returns a vector $\widehat{z}_{cp} \in \mathbb{R}^G$ for each held-out perturbation.

We measure accuracy with the **de-biased Pearson delta**, which quantifies agreement across readout genes between the predicted and measured perturbation-effect vectors. Following Nicol *et al.*, we form the predicted and observed effects using disjoint halves of the non-targeting controls, preventing their shared control reference from inflating the correlation.[13] We measure discrimination with the **cosine perturbation discrimination score (PDS) gain**: PDS tests whether each prediction is closer to its matched measured response than to those of other perturbations. We use cosine similarity, as ℓ_1/ℓ_2 is affected by magnitude of predicted vectors.[14] A score of 0.50 corresponds to chance, and we report the gain above chance.

We also computed DE overlap@N, which measures overlap between the top-ranked differential genes in the predicted and measured responses, with N set by the number of measured genes passing FDR < 0.05. Because DE overlap ranks methods nearly identically to the de-biased Pearson delta across all methods and cell lines (Supplementary Table A1), we report it in the supplement as a confirmatory accuracy measure. All formal metric definitions appear in Section 6.3.

2.3. *The mean baseline captures the shared response but cannot discriminate perturbations*

The training mean recovers much of the measured response but cannot distinguish which perturbation produced it. Across five random 80/20 splits per cell line, assigning every held-out perturbation the average training response matches or exceeds GEARS and scGPT on de-biased Pearson delta in every cell line, reaching 0.09–0.55 compared with 0.03–0.43 for GEARS (Supplementary Table A2). By construction, the same prediction is returned for every test perturbation, so its cosine PDS remains at chance (0.50). GEARS and scGPT also remain near chance, at gains of 0.01 and 0.00, despite producing perturbation-conditioned outputs. Their response-level agreement therefore comes primarily from structure shared across perturbations. All models and the training mean are described in Section 6.4.

GenePert outperforms the other baselines through a structure consistent with the decomposition we formalize in Section 3. The method fits a regularized regression from a gene embedding to the perturbation effect within a cell line and achieves the strongest response correlation among the within-cell baselines (0.32 on average, compared with 0.28 for the training mean). Its fitted intercept is the training mean, so each prediction is the shared response plus a shrunken gene-dependent correction. This gene-dependent correction raises its mean PDS gain to 0.08, although its discrimination remains modest relative to its response correlation.

Together, these results separate two prediction tasks: the mean's high accuracy shows that many perturbations share a dominant expression pattern, while its zero PDS gain shows that the perturbation-specific signal resides in the variation around the shared response. We next

characterize each response by both its overall strength and its agreement with the shared direction.

2.4. *Magnitude and alignment characterize perturbation responses*

For each perturbation p in cell line c, we computed a signed Wilcoxon rank-sum statistic for every readout gene by comparing perturbed cells with non-targeting controls from the same experiment. The resulting profile $w_{cp} \in \mathbb{R}^G$ summarizes the direction and consistency of the response. With $\bar{w}_c = |\mathcal{P}|^{-1} \sum_{p \in \mathcal{P}} w_{cp}$ denoting the mean profile over the 2,270 shared perturbations, we define

$$m_{cp} = \|w_{cp}\|_2 , \qquad a_{cp} = \mathrm{corr}\,(w_{cp}, \bar{w}_c) . \tag{2}$$

Magnitude measures the overall response strength, while alignment measures how closely the response follows the cell line's shared perturbation direction.

Magnitude and alignment are positively correlated in all six cell lines (Fig. 2), with Spearman correlations ranging from $\rho = 0.41$ in HEK293T to 0.82 in HepG2. As elaborated in Section 2.6, high-magnitude, high-alignment perturbations elicit broad, similar responses, suggesting extensive disruption of the cellular system. At the other end are perturbations that elicit specific responses that are poorly aligned with the mean.

This correlation is partly geometric: high-magnitude responses contribute disproportionately to $\bar{w}_c$ and therefore tend to align with the direction they collectively define, whereas the alignment of weak responses is estimated less reliably. Recomputing the mean without each perturbation, or using the median response, leaves the correlation essentially unchanged (Supplementary Table A3). Because magnitude and alignment are naturally coupled, they define a continuum rather than independent axes, motivating a single coordinate along the dominant trajectory (Section 2.5). This structure is robust to the choice of readout panel: recomputing the analysis with 1,000- and 5,000-highly variable gene panels preserves the continuum, cross-cell concordance, and response regimes (Supplementary Fig. A1 and Table A4), and to raising the minimum number of cells per perturbation (Supplementary Table A6).

2.5. *A one-dimensional coordinate orders the magnitude–alignment continuum*

To reduce this continuum to a single quantity, we standardized m_{cp} and a_{cp} within each cell line and computed a one-dimensional Isomap coordinate, $\boldsymbol{s_{cp}}$, oriented so that larger values of s_{cp} correspond to the high-magnitude, high-alignment arm (Section 6.5). This ordering is conserved across cell lines and studies. Across the fifteen cell-line pairs, Pearson correlations average $r = 0.50$ for alignment, $r = 0.38$ for magnitude, and $r = 0.50$ for s_{cp}, and every pairwise comparison is positive; Kendall's coefficients of concordance across all six lines are $W_a = 0.587$, $W_m = 0.533$, and $W_s = 0.588$ for alignment, magnitude, and s_{cp}, respectively (Fig. 3). These values are robust to raising the minimum number of cells per perturbation (Supplementary Table A6). Perturbations with high (low) s_{cp} in one cell line therefore tend to rank highly (lowly) in the others. This conservation persists across study-specific processing pipelines and environmental factors: the strongest concordance in s_{cp} occurs for the cross-study RPE1–HepG2

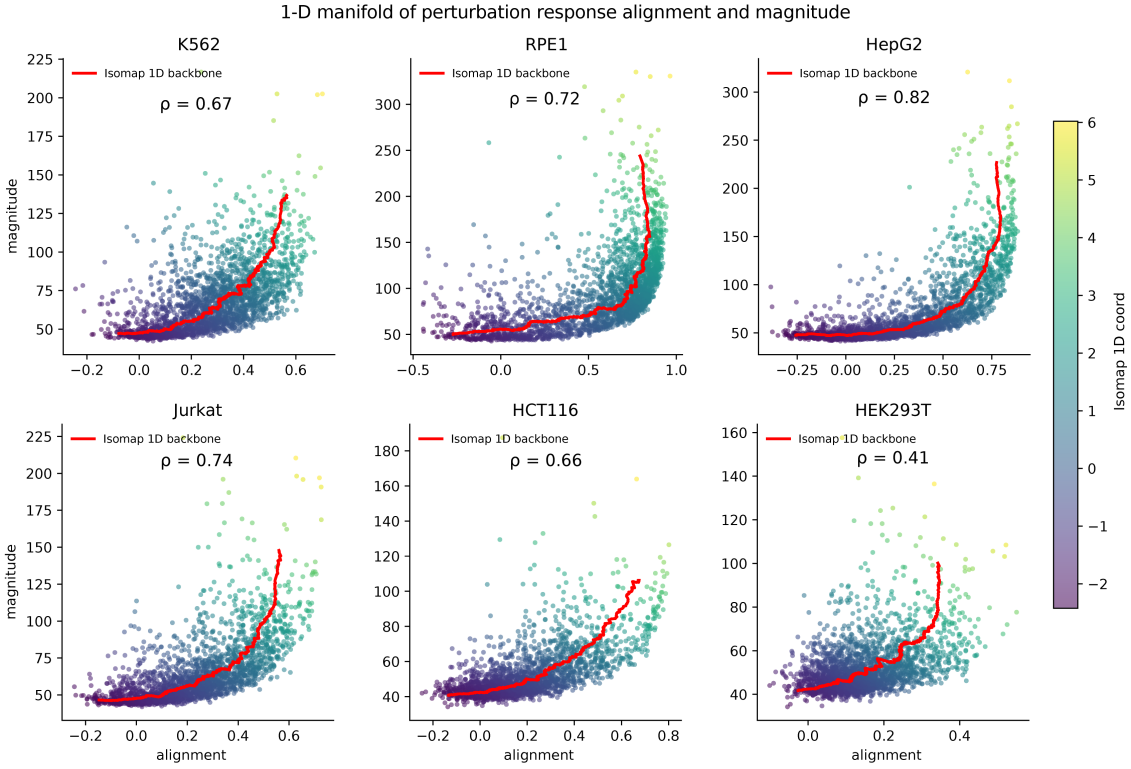


Fig. 2. **Magnitude and alignment trace a common response organization across six cell lines.** Response magnitude m_{cp} against alignment a_{cp} for the 2,270 shared perturbations. Cell-specific Spearman correlations between magnitude and alignment are marked on each plot. Points are coloured by the coordinate s_{cp} of Section 2.5.

pair ($r = 0.66$), and the next strongest, HepG2–HCT116 ($r = 0.62$), is also cross-study, while K562 and RPE1, the two Replogle lines, are among the least concordant ($r = 0.41$, eleventh of fifteen).

2.6. *The two ends of the continuum are biologically distinct*

The conserved ordering separates perturbations with distinct biological roles. We averaged the cell-specific coordinates to $\bar{s}(p) = |\mathcal{C}|^{-1} \sum_c s_{cp}$, ranked the 2,270 perturbations, and tested the top and bottom 30% for Gene Ontology overrepresentation (Supplementary Table A7). The high-magnitude, high-alignment end is dominated by core growth machinery, including cytoplasmic translation, rRNA processing, spliceosomal mRNA processing, RNA polymerase II preinitiation, and DNA replication. Representative genes include RPL4, RPS13, SF3B3, PRPF31, MCM2, MCM3, MCM6, GINS1, and GINS4. The low-$\bar{s}$ tail is enriched for regulatory and organelle-specific functions, including mitochondrial translation, chromatin remodeling, transcriptional coactivation, protein folding, and regulation of DNA-templated transcription, with genes such as MRPL45, MRPS7, KDM6A, BRD4, NCOA4, and PFDN1. The same terms remain significant when the tails are defined at 20%, 25%, or 30%; the full sensitivity analysis appears in Supplementary Table A8. These enrichments support the biological distinction underlying the continuum: core growth machinery is concentrated among strong, shared-

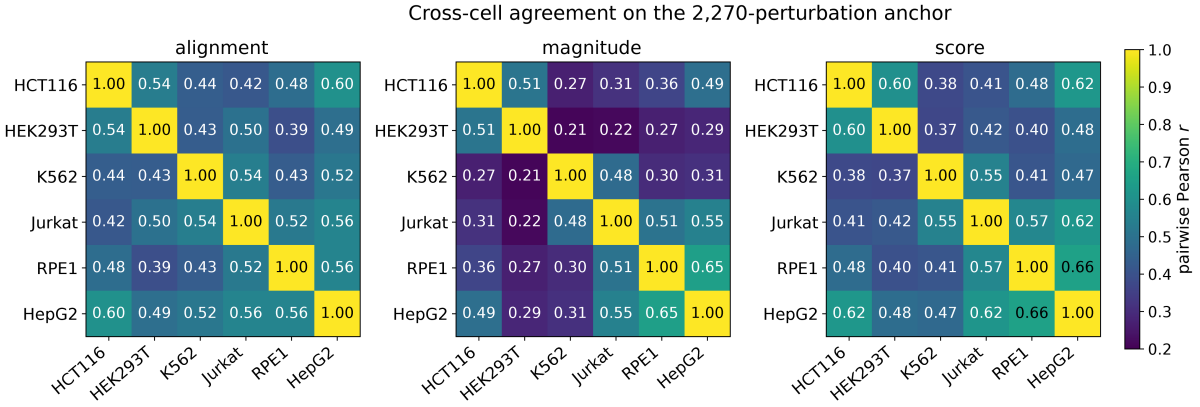


Fig. 3. **The response organization is conserved across cell lines and studies.** Pairwise cross-cell correlations of alignment, magnitude, and the coordinate s_{cp} across the 2,270 shared perturbations; all fifteen pairs are positive for each measurement.

response-dominated perturbations, whereas regulatory and organelle-specific functions are concentrated among less aligned responses.

2.7. *Gene representations predict conserved response position*

Because the same perturbations occupy similar positions across cell lines, we asked whether cell-agnostic prior knowledge about the targeted gene could predict that position. We trained a gradient-boosted regressor to predict the cross-cell means $\bar{m}(p)$, $\bar{a}(p)$, and $\bar{s}(p)$ from gene representations derived from protein-interaction networks, knowledge graphs, transcriptomic data, Gene Ontology annotations, protein sequences, and language models, evaluating every representation under identical perturbation-level five-fold splits, prediction targets, model class, and hyperparameters (Section 6.6).

STRING predicts a perturbation's position more accurately than the other representations (Supplementary Fig. A2). Consistent with its strong cross-cell agreement, alignment is the most predictable of the three quantities, with cross-validated $R^2 = 0.39$. Magnitude and $\bar{s}(p)$ are also well-predicted, albeit less strongly, at $R^2 = 0.24$ and 0.35, respectively. Graph-neural, knowledge-graph, Gene Ontology, and language-model representations also recover meaningful signal, while somewhat surprisingly, transcriptomic and protein-sequence representations perform less well.

3. Methods

3.1. *COMPASS: a shared-plus-specific decomposition of the perturbation response*

The response structure characterized above motivates a direct decomposition in expression space. For the pseudobulk effect vectors z_{cp} defined in Eq. (1), let $\bar{z}_c = |\mathcal{P}|^{-1} \sum_{p \in \mathcal{P}} z_{cp} \in \mathbb{R}^G$ denote the mean perturbation response in cell line c, and let $u_c = \bar{z}_c / \|\bar{z}_c\|_2$ be the unit vector along that mean. We decompose each response as

$$z_{cp} = \beta_{cp}\, u_c + r_{cp}, \qquad \beta_{cp} = \langle z_{cp}, u_c \rangle, \qquad r_{cp} = g_p + \varepsilon_{cp}, \tag{3}$$

Held-out cell line	K562	RPE1	HepG2	Jurkat	Mean
shared-response coefficient $\widehat{\beta}_{Cp} \to \beta_{Cp}$	0.61	0.70	0.71	0.62	**0.66**
gene-specific component $\widehat{g}_p \to r_{Cp}$	0.22	0.22	0.21	0.21	**0.22**

Table 1. **The shared-response coefficient transfers more strongly than the gene-specific response.** Each of the four Replogle/Nadig cell lines is held out in turn, and both components are estimated from the other three. The coefficient is a scalar per perturbation, compared by Pearson correlation across the 2,270 perturbations; the gene-specific response is a vector, compared by Pearson correlation across genes within each perturbation and averaged. A perturbation-shuffled null gives ≈ 0 for both.

Here $\beta_{cp} \in \mathbb{R}$ is the shared-response coefficient, computed by projecting z_{cp} onto u_c, and $\beta_{cp}u_c \in \mathbb{R}^G$ is the corresponding shared-response component. The vector $r_{cp} \in \mathbb{R}^G$ is the gene-specific response remaining after this component is removed; it is orthogonal to u_c by construction. Across cell lines with a common readout space, g_p denotes the conserved mean of this gene-specific response and ε_{cp} its cell-line-specific deviation. A single cell line cannot identify g_p and ε_{cp} separately, so within-cell estimators target r_{cp} directly. Here and below, *gene-specific* means specific to the perturbed gene p, rather than to an individual readout gene.

The shared-response coefficient β_{cp} is the expression-space counterpart of the empirical coordinate s_{cp}: both order perturbations by how strongly their responses follow the cell line's shared response. β_{cp} is slightly better conserved across cell lines than s_{cp} itself: per-cell-line estimated values correlate at mean pairwise Pearson 0.55 among the four Replogle/Nadig lines and 0.56 between the two Orion lines (Supplementary Table A9). The two orderings also agree closely: the Pearson correlation between the (Replogle/Nadig) cross-cell mean $\bar{\beta}_p = |\mathcal{C}|^{-1} \sum_c \beta_{cp}$ and $\bar{s}(p)$ is 0.88. Ranking by $\bar{\beta}_p$ yields similar gene function enrichments (Supplementary Table A10). Because $\bar{\beta}_p$ can be estimated from source cell lines and summarizes a perturbation's shared-response engagement in a single portable scalar, it can be used as an input in cross-cell prediction settings.

To measure how well these components transfer across cellular contexts, we performed leave-one-cell-line-out assessments on the four Replogle/Nadig lines, writing C for the held-out line. The estimate $\widehat{\beta}_{Cp}$, formed by averaging over the remaining three lines, predicts the held-out shared-response coefficient β_{Cp} with mean Pearson correlation 0.66, while the corresponding estimate $\widehat{g}_p$ matches the held-out gene-specific response r_{Cp} at mean Pearson 0.22 (Table 1). The shared-response coefficient thus transfers considerably more reliably than the gene-specific residual. With only one source line available, as for the Orion pair, transfer remains predictive but is clearly weaker than from three, 0.56 and 0.09 for the two components (Supplementary Table A11).

Together, these analyses identify two components with complementary conservation properties: the shared-response coefficient $\bar{\beta}_p$ is strongly conserved across cell lines and transfers reliably, while the gene-specific response, though more cell-line-specific, still retains meaningful cross-cell-line structure.

3.2. *Estimating COMPASS under different information regimes*

We now consider predicting z_{Cp} for an unmeasured perturbation p in a target cell line C. COMPASS assumes that the target shared-response direction u_C can be estimated from a calibration set of perturbations measured in C; the low-data analysis below shows that this estimate stabilizes within a few tens of perturbations (also see Discussion). The prediction task then reduces to estimating β_{Cp} and r_{Cp}. COMPASS provides three estimators for these quantities: a cross-cell estimator that carries both components over from other cell lines, a within-cell estimator that infers them from STRING neighbors of p measured in C, and a hybrid that combines both sources with component-specific weights (Supplementary Algorithms 1 to 3). In each case, the final prediction is assembled as

$$\widehat{z}_{Cp} = \widehat{\beta}_{Cp}\, u_C + \widehat{r}_{Cp}. \tag{4}$$

COMPASS-X: the perturbation is measured in other cell lines. When p is available in a set of source cell lines $\mathcal{S}_p = \{d \neq C : p \in \mathcal{P}_d\}$, we estimate its shared-response projection and conserved gene-specific component by

$$\widehat{\beta}^{\mathrm{X}}_{Cp} = \frac{1}{|\mathcal{S}_p|} \sum_{d \in \mathcal{S}_p} \beta_{dp}, \qquad \widehat{g}^{\mathrm{X}}_{p} = \frac{1}{|\mathcal{S}_p|} \sum_{d \in \mathcal{S}_p} r_{dp}. \tag{5}$$

Since the target-specific deviation ε_{Cp} is unobserved, we approximate $\widehat{r}^{\mathrm{X}}_{Cp} = \widehat{g}^{\mathrm{X}}_{p}$. We emphasize that this differs from simply averaging the raw source response vectors z_{dp} and transferring them directly. COMPASS-X first decomposes each source-line response into its shared-response projection and gene-specific component, then transfers these separately, applying the averaged coefficient $\widehat{\beta}^{\mathrm{X}}_{Cp}$ to the target-cell direction u_C rather than retaining the source-cell directions. Critically, this decomposed transfer outperforms direct source averaging, as we show below.

COMPASS-N: related perturbations are measured in the target cell. When the target line contains a sufficiently large perturbation set, or when p has not been screened elsewhere, we estimate both components from STRING neighbors measured in C. With $\mathcal{N}_k(p)$ denoting the k nearest training perturbations,

$$\widehat{\beta}^{\mathrm{N}}_{Cp} = \frac{1}{k_b} \sum_{q \in \mathcal{N}_{k_b}(p)} \beta_{Cq}, \qquad \widehat{r}^{\mathrm{N}}_{Cp} = \frac{1}{k_r} \sum_{q \in \mathcal{N}_{k_r}(p)} r_{Cq}. \tag{6}$$

The neighborhood sizes k_b and k_r can be selected separately on validation data (Section 6.7); we recommend default values of $k_b = 5$ and $k_r = 10$.

COMPASS-H: both sources of information are available. When both sources of information are available, we combine the β and r estimates from COMPASS-X and COMPASS-N. Because the shared-response coefficient and gene-specific response have different transfer properties, COMPASS-H assigns them separate mixing weights:

$$\begin{aligned} \widehat{\beta}^{\mathrm{H}}_{Cp} &= \lambda_b \widehat{\beta}^{\mathrm{X}}_{Cp} + (1-\lambda_b)\widehat{\beta}^{\mathrm{N}}_{Cp} \\ \widehat{r}^{\mathrm{H}}_{Cp} &= \lambda_r \widehat{r}^{\mathrm{X}}_{Cp} + (1-\lambda_r)\widehat{r}^{\mathrm{N}}_{Cp} \end{aligned} \tag{7}$$

We set $\lambda_b = 1$ because the cross-cell estimate comes from the same perturbation in other cell lines and better preserves its magnitude along the shared response direction; a sensitivity

analysis confirms that this choice gives the strongest discrimination (Supplementary Fig. A3). We select λ_r jointly with k_r on the validation split, allowing the gene-specific estimate to draw on target-cell neighbors when the conserved component alone is insufficient. As defaults, we recommend $\lambda_b = 1, \lambda_r = 0.5$.

		Accuracy: De-biased Pearson delta							Discrimination: Cosine PDS gain						
		K562	RPE1	HepG2	Jurkat	HCT116	HEK293T	Mean	K562	RPE1	HepG2	Jurkat	HCT116	HEK293T	Mean
Cross-cell	Source average	0.28	0.37	0.35	0.26	0.10	0.08	0.24	**0.25**	**0.32**	**0.24**	0.23	0.18	0.16	**0.23**
	STATE	0.14	0.44	0.31	0.17	0.10	0.08	0.21	0.18	0.04	0.04	0.11	0.07	0.06	0.08
	TabICL	**0.41**	**0.63**	**0.46**	**0.41**	**0.28**	**0.20**	**0.40**	0.24	0.18	0.20	**0.27**	0.15	0.16	0.20
	COMPASS-X	0.33	0.43	0.40	0.32	0.12	0.12	0.29	**0.25**	0.30	0.22	0.24	0.18	**0.19**	**0.23**
	COMPASS-H	0.36	0.51	0.43	0.35	0.22	0.16	0.34	0.24	0.28	0.21	0.24	**0.23**	**0.19**	**0.23**
Within-cell	Training mean	0.26	0.55	0.36	0.24	0.19	0.09	0.28	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	CPA	0.05	0.06	0.06	0.07	0.01	0.02	0.05	0.01	0.00	0.01	0.02	0.01	0.00	0.01
	GEARS	0.13	0.43	0.25	0.13	0.06	0.03	0.17	0.00	0.01	0.03	0.03	−0.01	0.02	0.01
	scGPT	0.18	0.55	0.34	0.19	0.16	0.06	0.25	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	GenePert	0.32	0.57	0.38	0.28	0.23	0.13	0.32	0.11	0.05	0.06	0.09	0.08	0.07	0.08
	TabICL	**0.39**	**0.61**	**0.43**	**0.35**	**0.28**	**0.19**	**0.37**	**0.21**	**0.14**	**0.15**	**0.19**	**0.16**	**0.15**	**0.16**
	COMPASS-N	0.34	0.57	0.39	0.30	0.25	0.16	0.34	0.19	0.13	0.13	0.18	0.15	0.12	0.15

Table 2. **Cross-context prediction across the six cell lines (random splits).** De-biased Pearson delta and cosine PDS gain, averaged over five 80/20 partitions per cell line. TabICL is run with all six prior-knowledge modalities as its authors do. **Bold** marks the best value in each column and metric, underline the second best. Variability across partitions and paired significance tests against the COMPASS variants are given in Supplementary Table A12.

4. Experiments

4.1. *COMPASS is competitive with the strongest baseline using a linear, interpretable model*

To benchmark the predictive performance of our approach, we compare the COMPASS suite against three families of existing models (Table 2): *structural baselines* (training mean, source average), *perturbation-conditioned deep learning models* (CPA, GEARS, scGPT, STATE), and *embedding-based regressors* (GenePert[15] and TabICL[16,17]). Our adaptation of TabICL follows Palla *et al.*'s work. Implementation details for each are given in Section 6.4. Compared to these, the COMPASS suite exploits the response structure characterized in Section 3.1: each prediction is a shared-response component plus a gene-specific residual, with both terms estimated from within-cell (COMPASS-N), cross-cell (COMPASS-X) or both (COMPASS-H). All models are scored on the same test perturbations, and any tuned hyperparameters are selected on validation perturbations (Section 6.9).

The COMPASS suite, specifically the hybrid model COMPASS-H, matches or outperforms every non-TabICL model on both accuracy (de-biased Pearson delta) and discrimination (cosine PDS gain) (Table 2). COMPASS-X uses the same inputs as the source average but decomposes them first, and is more accurate at comparable discrimination. This suggests that transferring information from other cell lines via our linear decomposition is more effective than using it directly. Adding within-cell information, i.e. COMPASS-H, further improves performance, while the source averaging baseline is unable to accommodate such additional information.

COMPASS substantially outperforms the direct perturbation-response deep learning models evaluated here. In within-cell contexts, COMPASS-N leads the deep learning approaches scGPT,

GEARS, and CPA. As shown in Section 2.3, several of these models also underperform simple averaging baselines on response accuracy.

GenePert and TabICL are regression-based frameworks that use one and six sets of gene embeddings, respectively, to predict response. GenePert works in a within-cell setting and uses regression; TabICL works in a hybrid setting and uses pre-trained tabular foundation models. We argue they are both best thought of as powerful variants of the training mean or source averaging baseline, as they both learn the overall response and add gene-specific tilts. Interestingly, both methods are stronger on accuracy than discrimination, analogous to the training mean.

TabICL is overall more accurate than COMPASS variants but slightly less discriminative; GenePert is outperformed by COMPASS on both metrics. TabICL leads COMPASS-H on de-biased Pearson delta by 0.06 cross-cell (0.40 vs. 0.34) and COMPASS-N by 0.03 within-cell (0.37 vs. 0.34). On cosine PDS gain, however, COMPASS-X and COMPASS-H match or exceed TabICL cross-cell (0.23 vs. 0.20), and COMPASS-N is within 0.01 within-cell (0.15 vs. 0.16). COMPASS is therefore competitive with a much heavier model, using STRING embeddings only as a distance metric to select neighbors rather than as regression inputs, and its decomposition additionally produces components that are interpretable in isolation (Section 4.2). For both COMPASS and TabICL, accuracy and discrimination decrease as test perturbations become more isolated from training perturbations, based on STRING embedding space (Supplementary Fig. A4).

COMPASS variants reflect the tension between accuracy and discrimination COMPASS-X is the stronger discriminator (PDS gain 0.23 vs. 0.15) because its perturbation-specific information comes from measurements of the same perturbation in other cell lines. COMPASS-N is the stronger accuracy model (de-biased Pearson 0.34 vs. 0.29) because averaging over STRING neighbors in the target line recovers the local shape of the response, but this neighbor-averaging cannot effectively separate a perturbation from its neighbors. COMPASS-N is therefore the appropriate estimator when the target-cell perturbation set is large but the target perturbation has not been measured elsewhere, while COMPASS-X is the appropriate estimator when the same perturbation has been measured in a related cell line but the target perturbation set is sparse (Supplementary Fig. A5). COMPASS-H estimates each component from the source that recovers it better, matching the stronger of the two on both axes at once.

For a new cell line, COMPASS is more data-efficient than TabICL When investigating a new cell line or tissue, a practical question is how many perturbations are needed before data from other cell lines can be effectively used. To assess this, we included all available cross-cell perturbations and increased the number of within-cell training perturbations from 10 to 130. Across ten seeds in each of the six cell lines (Fig. 4), COMPASS-X converges quickly: its de-biased Pearson delta rises from 0.253 to 0.288 across the range and its PDS gain stays flat near 0.24. This is because the target cell only enters through the shared response u_C, which converges within a few tens of perturbations. In contrast, TabICL starts well below COMPASS-X, and only overtakes it on accuracy at approximately 80 profiled perturbations; it does not outperform on discrimination even with maximal data in this scenario.

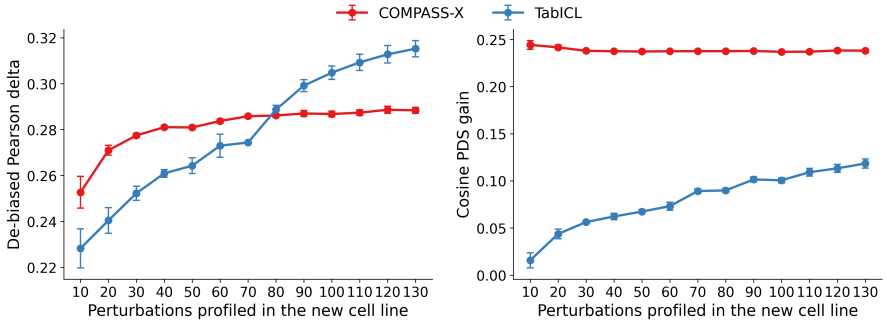


Fig. 4. **Cross-context transfer versus in-context learning as the profiling budget grows.** De-biased Pearson delta (left) and cosine PDS gain (right) against the number of perturbations profiled in a new cell line, with the remainder held out for test.

4.2. *The gene-specific response g_p encodes pathway-level biology*

The gene-specific component contains coherent biological programs rather than unstructured variation left after removing the shared response (Section 6.10). Averaging r_{cp} across the four Replogle/Nadig cell lines gives a 2,270 × 2,000 matrix of conserved components g_p. Clustering their normalized directions groups perturbations targeting common biological machinery, including mitochondrial translation, ribosome biogenesis, rRNA processing, mitosis, DNA replication, RNA splicing, and proteasome function (Supplementary Table A13). All fifteen clusters in the highest-coefficient tertile are enriched for perturbations targeting a coherent pathway or complex (Supplementary Table A15). The organization persists among perturbations that engage the shared response weakly, where clusters recover mitochondrial translation, N-glycosylation, and mitochondrial organization programs (Supplementary Table A14).

Individual gene-specific components also recover expected downstream responses. The TFAM component is dominated by reduced expression of mitochondrially encoded oxidative-phosphorylation transcripts. Proteasome-subunit perturbations, including PSMB5, repress cytoplasmic translation while inducing vacuolar and protein-stress programs; the PSMB5 component alone is enriched for unfolded-protein ($q = 8 \times 10^{-5}$) and heat-shock ($q = 3 \times 10^{-5}$) responses among its fifteen most induced genes. None of these analyses uses pathway labels during fitting. The gene-specific component therefore captures both relationships among perturbed genes and the downstream programs that distinguish their responses.

5. Discussion

The key conceptual advancement of our work is to efficiently disentangle gene-specific effects from shared responses in single-cell perturbation data. This separation both clarifies which pathways a perturbation engages and improves prediction of unmeasured perturbations. Across 2,270 CRISPRi knockdowns, response magnitude and alignment define a continuum from shared-response-dominated perturbations to those eliciting more gene-specific responses. The ordering of perturbations in the continuum is conserved across cell lines and studies. COMPASS formalizes this organization as $z_{cp} = \beta_{cp} u_c + r_{cp}$, with $r_{cp} = g_p + \varepsilon_{cp}$ across cell lines sharing a readout space. Averaging the shared-response coefficient for the same perturbation across source cell lines yields the portable perturbation-level quantity $\bar{\beta}_p$, which predicts how strongly that

perturbation will engage the shared response in a new cellular context. The shared-response coefficient transfers substantially more strongly than the gene-specific response g_p, which retains coherent pathway-level structure. Concurrent work by Molina and Zhang[18] similarly decomposes shared and perturbation-specific response structure, but our work additionally shows that the perturbation-specific coefficient β_{cp} on the shared response is itself portable across cell lines, enabling COMPASS to reconstruct the shared response as $\bar{\beta}_p u_c$, contributing to its efficiency and interpretability (Section 6.12).

The decomposition also explains why accuracy and discrimination diverge in recent perturbation-prediction benchmarks.[1,2,7] Accuracy metrics reward recovery of the response shared across many perturbations, whereas discrimination rewards capturing what makes each perturbation's response distinct. The COMPASS variants expose a corresponding tension between sources of evidence. In their corresponding regimes, COMPASS-X and COMPASS-N outperform scGPT, CPA, GEARS, GenePert, and STATE on both metrics. COMPASS-H combines the strengths of COMPASS-X and COMPASS-N and approaches TabICL's higher accuracy, exceeds its cross-cell discrimination, and is also more data and parameter-efficient than TabICL.

Our work suggests several directions for future work. As others have noted, perturbation predictors should be evaluated along two distinct axes, response accuracy and perturbation discrimination. Our results further suggest stratifying both metrics along the magnitude–alignment continuum, and model development should similarly treat the two components separately: cross-cell measurements and cell-agnostic gene knowledge may be particularly effective for estimating $\bar{\beta}_p$, whereas target-cell data may be needed to refine r_{cp}. The continuum also suggests a strategy for experimental design. A relatively small set of high β_p perturbations can establish the shared-response direction in a new cell type, allowing more of the profiling budget to be directed toward low-alignment perturbations whose responses carry distinctive and less predictable biology.

There are limitations to this work, including the modest number of cells per perturbation which results in uncertain estimation of the components (Supplementary Table A5). The current perturbation set is enriched for essential genes, so testing functionally diverse perturbations will establish whether the same shared-plus-specific organization persists beyond stress-dominated responses. Extending the analysis to primary cells and additional cellular contexts will similarly test how well $\bar{\beta}_p$ and g_p transfer beyond immortalized and cancer-derived cell lines and to out-of-distribution pathways. Finally, COMPASS currently requires perturbations in the target cell line to estimate u_c; learning this direction from unperturbed cells, for example through single-cell foundation model embeddings, could enable prediction in entirely new cellular contexts without a perturbation calibration set.

Code, data and supplemental materials availability Code, processed response matrices, split definitions, and scripts for reproducing the figures and tables is released at `https://github.com/rohitsinghlab/compass`. Source Perturb-seq datasets remain available from their respective studies. Supplemental sections, figures and tables can be found at `https://singhlab.net/compass/`.

References

1. C. Ahlmann-Eltze, W. Huber and S. Anders, Deep-learning-based gene perturbation effect prediction does not yet outperform simple linear baselines, *Nature Methods* **22**, 1657 (2025).
2. Z. Wei, Y. Wang, Q. Liu *et al.*, Benchmarking algorithms for generalizable single-cell perturbation response prediction, *Nature Methods* (2025).
3. M. Lotfollahi, A. Susmelj, C. De Donno, Y. Ji, I. L. Ibarra, F. A. Wolf, N. Yakubova and F. J. Theis, Predicting cellular responses to complex perturbations in high-throughput screens, *Molecular Systems Biology* **19**, p. e11517 (2023).
4. Y. Roohani, K. Huang and J. Leskovec, Predicting transcriptional outcomes of novel multigene perturbations with GEARS, *Nature Biotechnology* **42**, 927 (2024).
5. H. Cui, C. Wang, H. Maan, K. Pang, F. Luo and B. Wang, scGPT: toward building a foundation model for single-cell multi-omics using generative AI, *Nature Methods* **21**, 1470 (2024).
6. A. K. Adduri, D. Gautam *et al.*, Predicting cellular responses to perturbation across diverse contexts with State, *bioRxiv* (2025).
7. R. Viñas Torné *et al.*, Systema: a framework for evaluating genetic perturbation response prediction beyond systematic variation, *Nature Biotechnology* (2025).
8. J. M. Replogle, R. A. Saunders, A. N. Pogson, J. A. Hussmann, A. Lenail, A. Guna, L. Mascibroda, E. J. Wagner, K. Adelman, G. Lithwick-Yanai *et al.*, Mapping information-rich genotype-phenotype landscapes with genome-scale Perturb-seq, *Cell* **185**, 2559 (2022).
9. A. Nadig, J. M. Replogle, A. N. Pogson, M. Murthy, S. A. McCarroll, J. S. Weissman, E. B. Robinson and L. J. O'Connor, Transcriptome-wide analysis of differential expression in perturbation atlases, *Nature Genetics* **57**, 1228 (2025).
10. A. C. Huang, T.-H. S. Hsieh, J. Zhu, J. Michuda, A. Teng, S. Kim, E. M. Rumsey, S. K. Lam, I. Anigbogu, P. Wright *et al.*, X-atlas/Orion: Genome-wide Perturb-seq datasets via a scalable Fix-Cryopreserve platform for training dose-dependent biological foundation models, *bioRxiv* , 2025 (2025).
11. A. Tsherniak, F. Vazquez, P. G. Montgomery, B. A. Weir, G. Kryukov, G. S. Cowley, S. Gill, W. F. Harrington, S. Pantel, J. M. Krill-Burger *et al.*, Defining a cancer dependency map, *Cell* **170**, 564 (2017).
12. J. M. Dempster, J. Rossen, M. Kazachkova, J. Pan, G. Kugener, D. E. Root and A. Tsherniak, Extracting biological insights from the Project Achilles genome-scale CRISPR screens in cancer cell lines, *bioRxiv* , p. 720243 (2019).
13. P. B. Nicol *et al.*, Spurious correlation inflates performance in single-cell perturbation prediction, *bioRxiv* (2026).
14. Q. Liu, Q. Zhang, J. Du, S. Zhao and J. Wang, Effects of distance metrics and scaling on the perturbation discrimination score, *arXiv* (2025).
15. Y. Chen and J. Zou, GenePert: Leveraging GenePT embeddings for gene perturbation prediction, *bioRxiv* (2024).
16. J. Qu, D. HolzmÃžller, G. Varoquaux and M. L. Morvan, Tabiclv2: A better, faster, scalable, and open tabular foundation model, *arXiv preprint arXiv:2602.11139* (2026).
17. G. Palla, A. Hillsley, Y.-J. Kim and L. A. Royer, Tabular foundation models are competitive cellular perturbation predictors across biological scales, *bioRxiv* , 2026 (2026).
18. A. Molina and X. Zhang, Perturbation response decomposition enables biologically aligned generalization to unseen perturbations and cellular contexts, *bioRxiv* , 2026 (2026).
19. E. Cole, G.-J. Huizing, S. Addagudi, N. Ho, E. Hasanaj, M. Kuijs, T. Johnstone, M. Carilli, A. Davi, C. Ellington *et al.*, Foundation models improve perturbation response prediction, *BioRxiv* , 2026 (2026).
20. D. Szklarczyk, R. Kirsch, M. Koutrouli, K. Nastou, F. Mehryary, R. Hachilif, A. L. Gable, T. Fang,

N. T. Doncheva, S. Pyysalo *et al.*, The STRING database in 2023: protein-protein association networks and functional enrichment analyses for any sequenced genome of interest, *Nucleic Acids Research* **51**, D638 (2023).
21. A. Subramanian, P. Tamayo, V. K. Mootha, S. Mukherjee, B. L. Ebert, M. A. Gillette, A. Paulovich, S. L. Pomeroy, T. R. Golub, E. S. Lander *et al.*, Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles, *Proceedings of the national academy of sciences* **102**, 15545 (2005).
22. M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig *et al.*, Gene ontology: tool for the unification of biology, *Nature Genetics* **25**, 25 (2000).