RELAY: TRANSFERRING DENSE RESPONSE CONTRASTS TO DISTINGUISH UNSEEN PERTURBATIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

Predicting the effects of unseen perturbations requires capturing gene-level response tendencies while distinguishing the effects of different perturbations on the same readout gene. On PerturbQA, a gene prior (each readout gene's majority training label) attains Macro-F1 of 0.8814 for differential expression (DE), and every language-model and expression-based predictor we evaluate falls below it. Input ablations identify retrieved outcomes as the source of a retrieval LLM's within-gene signal; descriptions of biological mechanisms contribute little. Yet labels cover only about 1% of measured perturbation–gene pairs. We introduce RELAY, which transfers dense measured responses without a language model. Kernel ridge regression over a biological knowledge kernel transfers responses from related training perturbations. Task-specific readouts then combine each gene's tendency with the transferred contrast to predict DE and direction of change (DIR), retaining the gene prior's global accuracy while distinguishing perturbations that act on the same readout gene. Across four PerturbQA cell lines, RELAY Pareto-dominates every evaluated baseline on both global and within-gene metrics. It reaches DE and DIR Macro-F1 of 0.8941 and 0.8722 and improves within-gene balanced accuracy over the strongest baseline by 7.48 and 6.53 points. RELAY also extends to a held-out cell line and to unseen compounds, retaining its advantage over the competing methods.

1 INTRODUCTION

Predicting how an unseen genetic perturbation changes gene expression can help prioritize experiments. PerturbQA formulates this problem as two decisions: whether a readout gene is differentially expressed (DE) and, among DE-positive pairs, whether it is up- or downregulated (DIR) (Dixit et al., 2016; Adamson et al., 2016; Replogle et al., 2022; Wu et al., 2025). A useful predictor must capture the gene-level response tendencies that make a training-only *gene prior* strong: predicting each gene's majority training label already reaches DE Macro-F1 of 0.8814. It must also distinguish among perturbations affecting the same readout gene. A rule that gives every perturbation of a gene the same answer performs at chance on this within-gene criterion, even when its aggregate performance is strong (Yuan et al., 2026). We therefore evaluate two complementary objectives: global accuracy, measured by Macro-F1, and within-gene discrimination, measured by within-gene balanced accuracy (WG-BAcc; Figure 1b). The gene prior scores exactly chance on WG-BAcc.

None of the evaluated baselines meets both objectives on both tasks. A growing line of work applies large language models (LLMs) to these tasks and reports gains over expression-based models (Wu et al., 2025; Wang et al., 2026; Wei et al., 2026). Yet on DE, every LLM and expression-based predictor we evaluate falls below the gene prior in Macro-F1. AROMA comes closest but is near chance within genes, and SUMMER gains within-gene accuracy at a cost of 12.5 Macro-F1 points. SUMMER's within-gene signal depends on retrieved outcomes. Deleting mechanism descriptions changes WG-BAcc by at most 1.33 points, whereas deleting outcomes reduces DIR to chance. Such outcome labels cover only about 1% of measured perturbation–gene pairs. Expression-based predictors use every measurement but, in the runs we evaluate, predict held-out profiles no better than training gene means (Section 3.3).

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

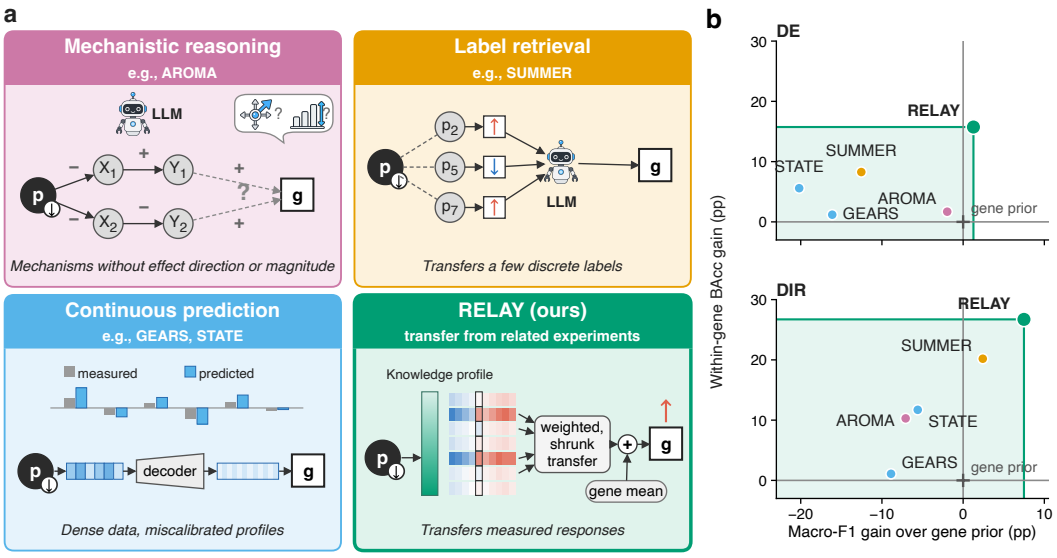


Figure 1: **Prediction strategies, global accuracy, and within-gene discrimination. a,** The evidence each strategy uses; RELAY transfers response contrasts through a knowledge kernel. **b,** Macro-F1 gain over the gene prior against WG-BAcc gain over chance (Section 2.2). In both tasks, every evaluated baseline lies in the region dominated by RELAY (shaded); CORE-Voting, which reports DE only and no WG-BAcc, is also below RELAY in DE Macro-F1 and pgAUROC (Table 2).

These observations suggest that gene-level response tendencies support global accuracy, while perturbation-specific measured responses support within-gene discrimination. **RELAY** combines these sources of information without a language model. Kernel ridge regression over a *knowledge profile* of curated annotations and interaction partners transfers the measured responses of related training perturbations as contrasts around gene means. Task-specific readouts then combine each gene's tendency with the transferred contrast. For direction, adding the gene mean improves global accuracy while preserving the within-gene ranking of contrasts. For significance, a learned readout relates the perturbation's response-wide pattern to how strongly the gene usually responds.

Our contributions are:

1. We diagnose why existing predictors fall short in global accuracy or within-gene discrimination. Input ablations trace the within-gene signal of retrieval LLMs to retrieved outcomes, and an exact error decomposition reveals gene-level error and excessive contrast amplitude in expression-based predictors.
2. We introduce RELAY, a lightweight and effective method that transfers measured responses of related training perturbations through a knowledge kernel. Its task-specific readouts combine each gene's tendency with the transferred contrast to support both objectives.
3. We show that RELAY Pareto-dominates every evaluated method across four cell lines, improving WG-BAcc over the strongest baseline by 7.48 (DE) and 6.53 (DIR) points. It retains its advantage over competing methods on an out-of-distribution cell line and on unseen compounds.

2 Task Formulation and Evaluation

2.1 Tasks and Available Information

For a fixed cell line, let $\mathcal{P}_{\text{tr}}$ and $\mathcal{P}_{\text{te}}$ denote disjoint training and test CRISPRi perturbations and $\mathcal{G}$ the measured readout genes. A query (p, g) asks how perturbation p affects readout gene g. DE asks whether g is significantly differentially expressed under p, and DIR, evaluated on ground-truth DE-positive pairs, asks whether it is down- or upregulated. PerturbQA provides selected query sets $\mathcal{Q}^T_{\text{tr}}$ and $\mathcal{Q}^T_{\text{te}}$ for each task $T \in \{\text{DE}, \text{DIR}\}$ (Wu et al., 2025), and pairs outside them are unlabeled.

Training experiments provide more than these labels. For each training perturbation and readout gene, the measured response $r(p, g)$ is the difference in mean $\log(1 + \mathrm{CP10K})$ expression between perturbed cells and batch-matched controls. The same screens also provide published effect estimates $F(p, g)$ with standard errors $\mathrm{SE}(p, g)$ (Nadig et al., 2025; Nadig, 2025). Knowledge graphs released with PerturbQA supply functional annotations and molecular relationships (Wu et al., 2025). Test responses and labels are used only for evaluation.

Beyond perturbations held out within a cell line, we consider two further settings. Prediction across cell lines uses training data from other cell lines and only unperturbed controls from the held-out cell line. Compound prediction poses the same DE and DIR tasks for drugs held out within compound-treated cell lines (Wei et al., 2026).

2.2 Global and Within-Gene Evaluation

PerturbQA scores DE and DIR by per-gene AUROC (pgAUROC) across perturbations of the same readout gene (Wu et al., 2025), whereas AROMA reports global class-wise F1 (Wang et al., 2026). Neither alone captures both objectives. A *gene prior* assigns each gene its majority training label, scoring highly globally without distinguishing perturbations. We compute Macro-F1 over all official queries for each cell line and task, and evaluate within-gene decisions on genes with both labels among their test queries. This mixed-label set, denoted $\mathcal{G}_{\mathrm{mix}}$, is shared by all models.

Within-gene ranking and decision quality are distinct. Continuous predictors produce fine-grained scores, whereas LLMs return coarse answers or vote counts, so pgAUROC also rewards score resolution. The vote-count audit in Appendix A.2 shows that adding votes raises pgAUROC without improving decisions. We therefore keep pgAUROC as a secondary ranking metric and score decisions by gene-level balanced accuracy, $\mathrm{BAcc}_g = (\mathrm{TPR}_g + \mathrm{TNR}_g)/2$ (Brodersen et al., 2010). The unweighted mean, pgBAcc, is noisy because more than half of the mixed-label DE genes have a single test query in one class. For these genes, one decision changes BAcc_g by 0.5. We therefore weight genes by their support in both classes, giving the *support-weighted within-gene balanced accuracy* (WG-BAcc),

$$w_g = \frac{n_g^+ n_g^-}{n_g^+ + n_g^-}, \qquad \text{WG-BAcc} = \frac{\sum_{g \in \mathcal{G}_{\mathrm{mix}}} w_g \, \mathrm{BAcc}_g}{\sum_{g \in \mathcal{G}_{\mathrm{mix}}} w_g}, \tag{1}$$

where n_g^+ and n_g^- count positive and negative test queries for gene g. The weight is proportional to the inverse of a Bernoulli variance bound derived in Appendix A.2. Macro-F1 and WG-BAcc are thus our primary metrics for global accuracy and within-gene discrimination, respectively.

2.3 Related Work

Methods designed for the discrete PerturbQA tasks rely on language models. *Mechanistic reasoning* generates explanations over biological relations before answering (Wang et al., 2026; Phillips et al., 2026; Wei et al., 2026), and *label retrieval* borrows outcomes of related training pairs (Wu et al., 2025; Yuan et al., 2026). Profile predictors instead estimate continuous expression profiles, which we convert into decisions when using them as baselines. *Continuous prediction* learns perturbation representations with deep models (Roohani et al., 2024; Adduri et al., 2026), whereas *similarity-based profile prediction* combines the profiles of related training perturbations, selected through knowledge-graph neighborhoods (Fawkes et al., 2026). RELAY also draws on related training perturbations. It centers response profiles on gene means, transfers the resulting contrasts through redundancy-aware kernel ridge regression, and converts them into DE and DIR decisions (Appendix A.3).

3 Diagnosing Existing Predictors

We examine three strategies through representative models: mechanistic reasoning (AROMA), label retrieval (SUMMER), and continuous prediction (GEARS and STATE; Section 2.3). We ask why each falls short in global accuracy or within-gene discrimination (Figure 1b).

3.1 From Biological Context to Response Evidence

Mechanistic reasoning traces how a perturbation alters biological processes and how these changes affect the readout gene. AROMA agrees with the gene prior on 96.77% of DE decisions. When it disagrees, it introduces more errors than it corrects, lowering accuracy by 0.70 points. It is fine-tuned on reasoning traces generated by a teacher model with access to the outcome, then rewarded for correct predictions (Wang et al., 2026). Neither stage controls for the gene prior, so recalling each gene's training tendency alone can earn the reward.

The curated relations used in the reasoning chain also lack the evidence needed for a decision. The coverage audit in Appendix B.1 shows that they describe biological context but not measured responses. GO BP (Ashburner et al., 2000) annotates both genes of 94.96% of DIR queries, but signed regulatory networks (Szklarczyk et al., 2025; Han et al., 2018; Türei et al., 2026) provide a short directed path from the perturbed to the readout gene ending in expression regulation for only 15.87%. These relations do not encode effect magnitudes, yet DE depends on whether the gene's overall expression change crosses a detection threshold. The AROMA explanations analyzed in Appendix B.1 hinge on unsupported claims, such as compensation strong enough to prevent DE. Multiplying edge signs along paths that end in a signed transcriptional relation yields conflicting directions for two thirds of the covered DIR queries, and a majority vote over paths performs close to chance. Without response-level evidence, a model trained this way tends to return the gene's prior outcome with a plausible explanation.

3.2 Outcome Retrieval and Its Coverage Limit

Label retrieval answers a query from the observed outcomes of related training pairs. SUMMER gives an LLM the descriptions and outcomes of 11–12 retrieved pairs per query (Wu et al., 2025) and is the strongest within-gene baseline, but it falls 12.5 DE Macro-F1 points below the gene prior. To locate its within-gene signal, we delete reference fields while fixing the query, the retrieved identities, and their order. Removing descriptions changes WG-BAcc by at most 1.33 points, whereas removing outcomes reduces DIR to the query-only level and removes more than half of the DE gain. The signal is also tied to which experiment each outcome comes from. In the exchange experiment of Appendix B.2, reassigning outcomes among references on Jurkat DIR lowers WG-BAcc by 4.83 points. Moving each identity with its outcome only reorders the references and has no clear effect (-0.55 points). Biological knowledge thus helps choose which pairs to consult, while their outcomes carry the within-gene signal.

Table 1: **SUMMER's within-gene signal comes from retrieved outcomes.** WG-BAcc averaged across four cell lines; changes are in points relative to the corresponding Full reference. DE and DIR decoding settings follow the main text (Appendix B.2); DE query-only answers non-DE for every query. Table B.4 gives intervals.

Input to SUMMER	DE		DIR	
	WG-BAcc	ΔWG-BAcc (pp)	WG-BAcc	ΔWG-BAcc (pp)
Full reference	.5849		.7019	
No descriptions	.5860	$+0.11$	.6886	-1.33
No outcomes	.5378	-4.71	.5083	-19.36
Query only	.5000	-8.49	.5061	-19.59

Label availability, however, limits this evidence. The median test gene has only 12–30 labeled training perturbations for DE and 4–7 for DIR, but measured responses for 1,086–1,596. Retrieval can also override a reliable prior. On the 42,153 DE queries whose gene has at least 20 identical training labels, the gene prior is 99.57% accurate but SUMMER only 88.85%. Retrieval thus improves within-gene discrimination at the expense of global accuracy on DE.

3.3 Continuous Response Predictors

Continuous response profiles can support both tasks. The sign of the measured response $r(p, g)$ reproduces every DIR label, and $|r|$ ranks DE queries within genes at pgAUROC 0.987. GEARS and

STATE both fall 5.6–20.2 Macro-F1 points below the gene prior across tasks, and the profile errors in Appendix B.3 show that their held-out profiles are no more accurate than training gene means. This aggregate error does not reveal which part of the profile fails. Gene means support global accuracy, whereas contrasts around them support within-gene discrimination. A predicted contrast can err in alignment or amplitude. We therefore write test responses as $m(g) + \delta(p, g)$ and predictions as $\widehat{m}(g) + \widehat{\delta}(p, g)$, with both contrasts centered over test perturbations. The exact identity

$$\mathrm{MSE} = \overline{(\widehat{m} - m)^2} + \sigma_\delta^2(1 - \rho^2) + \sigma_\delta^2(\eta - \rho)^2 \tag{2}$$

uses pooled contrast correlation ρ and amplitude ratio $\eta = \sigma_{\widehat{\delta}}/\sigma_\delta$. The terms are gene-level error, contrast error left after optimal linear rescaling, and amplitude error, which vanishes only at $\eta = \rho$.

GEARS and STATE reduce unexplained contrast error relative to gene means, but their gene-level and amplitude errors outweigh this gain. Their contrasts are weakly aligned ($\rho = 0.12$ and 0.26), with amplitudes too large for that alignment ($\eta = 0.36$ and 0.48); on GeneTAK, VCWorld likewise finds that STATE and scVI (Lopez et al., 2018) call two to five times too many DE genes (Wei et al., 2026). Restoring training gene means improves DIR Macro-F1 for both predictors without changing within-gene rankings.

Together, the diagnostics suggest transferring measured responses between related perturbations with controlled amplitude and anchoring decisions on gene-level tendencies. These choices address within-gene discrimination and global accuracy, respectively.

4 RELAY: Response Transfer and Task-Specific Decisions

RELAY meets these requirements in two stages (Figure 2; Algorithm 1). Kernel ridge regression over a knowledge kernel transfers measured response contrasts with controlled amplitude (Sections 4.1–4.2), and task-specific readouts anchor DE and DIR decisions on gene-level tendencies (Section 4.3).

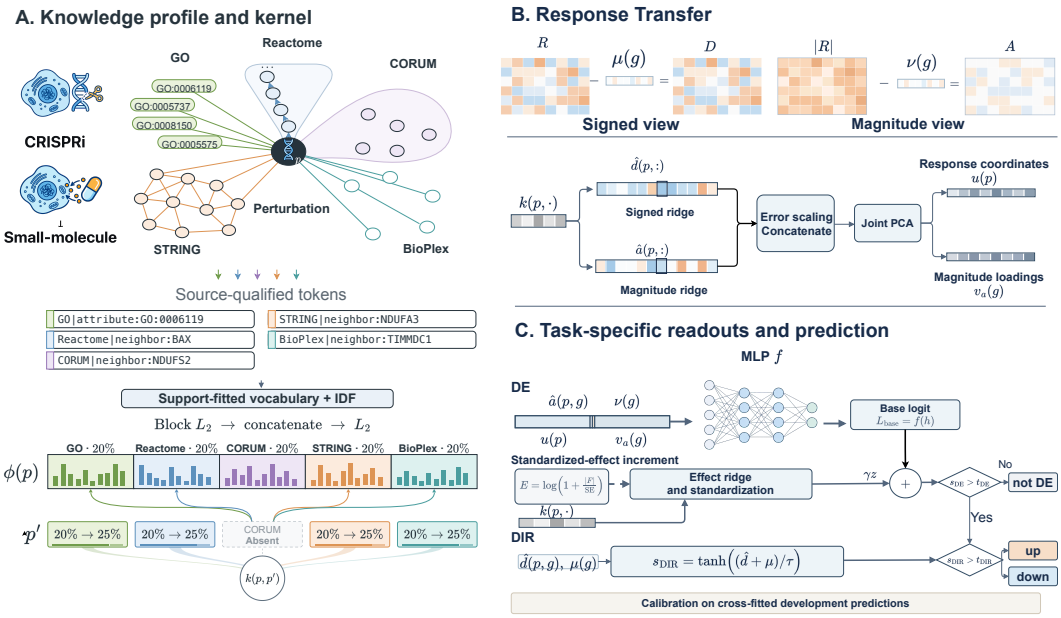


Figure 2: **RELAY's response-transfer pipeline. A,** Knowledge profiles define a knowledge kernel $k(p, p')$. **B,** Kernel ridge regressions use the query's similarities $k(p, \cdot)$ to transfer signed and magnitude contrasts, which a joint PCA summarizes for the DE readout. **C,** Task-specific readouts combine gene-level tendencies with the transferred contrasts. Source percentages are normalization weights, not learned weights; edges denote context records, not message passing.

4.1 Response Contrasts

Following the diagnosis in Section 3, RELAY retains the gene means and learns only the variation around them. All quantities below are fitted on the support $\mathcal{S} \subseteq \mathcal{P}_{\mathrm{tr}}$, the training perturbations used in a given fit. We arrange their responses in a matrix $R \in \mathbb{R}^{n \times |\mathcal{G}|}$ with $n = |\mathcal{S}|$, whose row for perturbation p contains its responses $r(p, g)$ across readout genes g. Each column mean of R gives a gene's average response; centering that column isolates the variation around it.

The two tasks need different views of this variation. DIR depends on the sign of a response, whereas DE depends on its size regardless of sign, and signed contrasts of related perturbations with opposite signs can cancel when transferred. RELAY therefore centers the columns of both R and its elementwise absolute value $|R|$, giving the signed and magnitude contrasts

$$d(p, g) = r(p, g) - \mu(g), \qquad a(p, g) = |r(p, g)| - \nu(g), \tag{3}$$

where $\mu(g) = |\mathcal{S}|^{-1} \sum_{p \in \mathcal{S}} r(p, g)$ and $\nu(g) = |\mathcal{S}|^{-1} \sum_{p \in \mathcal{S}} |r(p, g)|$ are the column means of R and $|R|$. The contrasts form matrices D and A of the same shape (Figure 2B), which Section 4.2 transfers to unseen perturbations.

4.2 Response Transfer over a Knowledge Kernel

Transferring contrasts to an unseen perturbation requires knowing which training perturbations are related to it. Because curated relations rarely determine a response, knowledge enters RELAY only through relatedness. A *knowledge profile* $\phi(p)$ encodes the annotations and interaction partners of the perturbed gene in GO (Ashburner et al., 2000; Aleksander et al., 2023), Reactome (Milacic et al., 2024), CORUM (Tsitsiridis et al., 2023), STRING (Szklarczyk et al., 2023), and BioPlex (Huttlin et al., 2021) as source-qualified tokens, such as `STRING|neighbor:NDUFA3`. Tokens are IDF-weighted (Spärck Jones, 1972) and normalized within each source and then jointly, with vocabularies and weights computed on the support only. The kernel $k(p, p') = \phi(p)^\top \phi(p')$ is thus a normalized sum of source-wise cosine similarities.

Given these similarities, RELAY must decide how much each training perturbation contributes, using the whole support while controlling the amplitude of the transferred contrasts. Kernel ridge regression meets both needs. Let $X \in \mathbb{R}^{n \times V}$ have the support profiles $\phi(p)$ as rows, where V is the support vocabulary size, and let $K = XX^\top$ and $k_p = k(p, \cdot) = (k(p, p'))_{p' \in \mathcal{S}} \in \mathbb{R}^n$. For each target $B \in \{D, A\}$, ridge regression (Hoerl & Kennard, 1970; Saunders et al., 1998) with a common penalty α predicts

$$\widehat{B}(p, :) = \phi(p)^\top W_B^* = k_p^\top (K + \alpha I_n)^{-1} B, \qquad W_B^* = \arg\min_W \ \|XW - B\|_F^2 + \alpha \|W\|_F^2. \tag{4}$$

The two views thus share transfer coefficients $w(p) = (K + \alpha I_n)^{-1} k_p$. Unlike the weights in a neighbor average, these coefficients account for redundancy among related perturbations, so they can be negative and need not sum to one. Larger penalties shrink predicted contrasts toward zero, limiting the excessive amplitudes observed in continuous predictors. We select α separately for each cell line. A perturbation without annotations ($k_p = 0$) receives zero contrasts, so its predicted response falls back to the gene means.

4.3 Task-Specific Readouts

Readouts combine each gene's tendency with the perturbation's predicted effect to make DE and DIR decisions. DIR depends only on the direction of the response. In contrast, DE prediction depends on the perturbation's overall effect, the gene's typical response, and uncertainty in the estimated effect.

Direction. The DIR score adds the gene mean back to the transferred signed contrast and rescales the result:

$$s_{\mathrm{DIR}}(p, g) = \tanh\left(\frac{\widehat{d}(p, g) + \mu(g)}{\tau(g)} \right), \tag{5}$$

where $\tau(g) > 0$ is a robust scale of the gene's signed prediction errors that makes scores comparable across genes. We predict upregulation when s_{DIR} exceeds a single calibrated threshold t_{DIR}. The gene mean shifts each gene toward its typical direction, while the score preserves the within-gene ranking of $\widehat{d}$.

Differential expression. The DE score combines a base logit computed from the predicted response profile with an increment from a transferred standardized effect. The base logit places the predicted effect on a gene in the context of the whole predicted profile. This context comes from a joint PCA of the predicted signed and magnitude contrasts, each first scaled by its prediction error and $\tanh$-transformed. The PCA represents each perturbation by response coordinates $u(p)$ in a shared component space and each readout gene by loadings $v_a(g)$ on the magnitude half of those components. The features combine the perturbation's global response pattern $u(p)$ with the gene's response characteristics $\nu(g)$ and $v_a(g)$. The base logit thus evaluates the pair-specific effect $\widehat{a}(p, g)$ in both contexts:

$$h(p, g) = \big[u(p),\ \widehat{a}(p, g),\ \nu(g),\ v_a(g)\big], \qquad L_{\text{base}}(p, g) = f\big(h(p, g)\big), \tag{6}$$

where f is a small MLP.

The increment accounts for effect uncertainty, which is not represented in the predicted response profile. A second kernel ridge regression, fitted on the training perturbations' effect statistics, transfers the standardized effect $E = \log(1 + |F|/\text{SE})$ using the same procedure as for the contrasts. This target expresses effect size relative to its standard error. The predictions, standardized per gene to $z(p, g)$, shift the base logit:

$$s_{\text{DE}} = \tanh\left(\frac{L_{\text{base}} + \gamma z}{2}\right), \tag{7}$$

where γ controls the contribution of the increment. We predict DE when s_{DE} exceeds a calibrated threshold t_{DE}.

Fitting and calibration. The readouts must interpret responses transferred to unseen perturbations, so they are fitted on out-of-sample development predictions. Cross-fitting generates these predictions by holding out groups of training perturbations and fitting regressions on the remaining groups. Prediction errors determine the scales τ. We use the cross-fitted predictions and training labels to train f and estimate the coefficient β in $\gamma = \lambda\beta$. We select the penalty α and decision thresholds by Macro-F1, and the factor λ by WG-BAcc, allowing at most a 0.01 Macro-F1 loss relative to the model without the increment. At test time, the regressions are refitted on all training perturbations; the query perturbation requires only its knowledge profile.

4.4 Held-Out Cell Lines and Unseen Compounds

For held-out cell lines and unseen compounds, RELAY retains the response-transfer framework and adapts its inputs. A held-out cell line contributes only unperturbed controls. We combine gene means and contrasts from source cell lines in response units, weighting each source by the similarity between its controls and the target controls. The readouts are refitted with each source cell line serving in turn as a pseudo-held-out cell line. We predict non-DE if $s_{\text{DE}} \leq t_{\text{DE}}$; otherwise, we predict upregulation if $s_{\text{DIR}} > t_{\text{DIR}}$ and downregulation if not. An unseen compound is represented by pooling the knowledge tokens of its protein targets into one knowledge profile. The standardized-effect increment is omitted when matched standard errors are unavailable.

5 Experiments

We evaluate whether RELAY improves both global accuracy and within-gene discrimination (Section 5.2), examine the contributions and trade-offs of its components (Section 5.3), and test whether its gains extend to a held-out cell line and unseen compounds (Section 5.4).

5.1 Experimental Setup

We evaluate on PerturbQA's 116,041 DE and 16,208 DIR test queries (Wu et al., 2025). Besides the gene prior, the baselines, detailed in Appendix D.1, follow the three diagnosed strategies: **AROMA** for mechanistic reasoning (Wang et al., 2026), **SUMMER** for label retrieval (Wu et al., 2025), and **GEARS** and **STATE**, trained on the official splits, for continuous prediction (Roohani et al., 2024; Adduri et al., 2026). A control, *RELAY with a GenePT kernel* (Chen & Zou, 2025), replaces only the knowledge profile.

5.2 Main Results

RELAY outperforms every baseline on both primary metrics in both tasks, and all improvements are significant (95% bootstrap CIs exclude zero). Within-gene discrimination therefore need not come at the expense of global accuracy.

Table 2: **Main comparison on PerturbQA.** Means weight cell lines equally; baselines are grouped by paradigm; bold marks the best value per column. Unanswered LLM queries receive the pooled training-rate fallback; AROMA gives a usable direction for 69.5% of DIR queries (Appendix D.1). CORE-Voting (Yuan et al., 2026) is listed with its reported DE values (Table D.1).

	DE			DIR		
Method	Macro-F1	WG-BAcc	pgAUROC	Macro-F1	WG-BAcc	pgAUROC
Gene prior	.8814	.5000	.5000	.7973	.5000	.5000
AROMA (reasoning)	.8618	.5169	.5152	.7265	.6029	.5526
SUMMER (retrieval)	.7560	.5827	.5891	.8215	.7019	.6637
GEARS (continuous)	.7202	.5119	.5143	.7085	.5109	.5462
STATE (continuous)	.6795	.5559	.5452	.7414	.6171	.6584
CORE-Voting (reported)	$\approx$ .8859	—	.7025	—	—	—
RELAY (ours)	**.8941**	**.6574**	**.7732**	**.8722**	**.7672**	**.8611**

RELAY achieves this discrimination by selectively revising only 5.09% of the gene prior's DE decisions. These changes raise DE WG-BAcc above chance by 15.74 points, with gains in every cell line (12.62–19.37). On genes with at least 20 identical training labels, it retains 99.39% accuracy compared with the prior's 99.57%.

5.3 Component Analysis

We examine three design choices: transferring measured responses from all training perturbations, transferring responses through a knowledge kernel, and applying task-specific readouts. We remove or replace components and build RELAY up one component at a time.

Table 3: **From label transfer to RELAY** (means weighted equally across cell lines). Starting from the base, each row adds one component; direct scores use no learned readout. †Separately tuned penalty (Appendix D.3.1).

	DE		DIR	
Configuration	Macro-F1	WG-BAcc	Macro-F1	WG-BAcc
Gene prior	.8814	.5000	.7973	.5000
Label transfer, direct scores†	.8987	.5598	.8510	.6788
Response transfer, direct scores (base)	.8133	.6341	.8649	.7709
+ learned DE readout	.8990	.6068	.8648	.7703
+ centering on gene means	.9016	.6152	.8722	.7672
+ magnitude view	.9045	.6228	.8722	.7672
+ standardized effect (**full RELAY**)	.8941	.6574	.8722	.7672
Full RELAY with a GenePT kernel	.8847	.6106	.8292	.7047

Table 3 examines response coverage, representation, and task-specific readouts. Response transfer expands coverage to all training perturbations, improving within-gene accuracy even after binarization. The knowledge kernel tracks response similarity more closely than GenePT. Appendix D.3 provides the full controls.

The readouts, in contrast, balance the two objectives through task-specific trade-offs. On DE, the learned readout raises Macro-F1 but lowers WG-BAcc, while the standardized-effect increment does the reverse. Aligning decisions with a gene's usual response favors global accuracy but reduces differences between perturbations; the increment restores some of that distinction. A DIR decision

depends only on the sign of the response, so adding the gene mean to the transferred contrast suffices and leaves within-gene rankings unchanged. A DE decision also depends on the perturbation's overall effect and the gene's typical response, which a single predicted magnitude cannot convey. Accordingly, removing the response-wide PCA context lowers both primary metrics (Appendix D.3).

5.4 Held-Out Cell Line and Unseen Compounds

For the out-of-distribution task, we transfer from K562, HepG2, and Jurkat to RPE1, which contributes only unperturbed controls. On AROMA's split and three-class task, we compare with AROMA and a source gene prior that assigns each gene its majority label across source cell lines. Appendix D.4.1 reports full binary results and a pair-level prior. RELAY reaches the highest Macro-F1, driven by gains in the up and down classes. On the binary tasks, it also retains within-gene discrimination when transferred to a cell line with no observed perturbation responses.

Table 4: **Out-of-distribution transfer to held-out RPE1.** Three-class F1 (ND, non-DE) and WG-BAcc of the binary tasks; AROMA reports only three-class values.

	Three-class				WG-BAcc	
Method	ND	Up	Down	Macro-F1	DE	DIR
Source gene prior	.9590	.4845	.6094	.6843	.5000	.5000
AROMA (reported)	.96	.57	.66	.73	—	—
RELAY (ours)	.9566	**.6243**	**.6980**	**.7597**	**.5933**	**.7333**

For compounds, we use GeneTAK's Tahoe-100M cell lines (Wei et al., 2026; Zhang et al., 2025), split as in Appendix D.4.2 into 284 training and 95 held-out drugs. Besides the gene prior, we compare with the LLM reasoning method VCWorld (Wei et al., 2026) and STATE trained on the same split.

Table 5: **Unseen compounds on GeneTAK.** Means across cell lines, weighted equally, over 95 held-out drugs; results by cell line are in Appendix D.4.2. Unanswered VCWorld queries receive the pooled training-rate fallback.

	DE			DIR		
Method	Macro-F1	WG-BAcc	pgAUROC	Macro-F1	WG-BAcc	pgAUROC
Gene prior	**.9196**	.5000	.5000	**.7894**	.5000	.5000
STATE	.8298	.5113	.5254	.6475	.5337	.5561
VCWorld	.6988	.5282	.5297	.5949	.5179	.5140
RELAY (ours)	.9179	**.5463**	**.6138**	.7828	**.5574**	**.6284**

RELAY nearly matches the gene prior's global accuracy while distinguishing compounds acting on the same readout gene better than every other method. Its within-gene gains over the prior have 95% CIs above zero in both tasks.

6 Conclusion

Predicting discrete perturbation effects requires both the global accuracy of a gene prior and the ability to distinguish perturbations acting on the same readout gene. None of the evaluated baselines meets both objectives on both tasks. Our diagnosis traces the within-gene signal to measured outcomes, but labels cover only a small fraction of measured perturbation–gene pairs. RELAY therefore transfers measured response contrasts through a knowledge kernel and anchors its decisions on gene-level tendencies. It Pareto-dominates every evaluated baseline on PerturbQA and retains its advantage over competing predictors on a held-out cell line and unseen compounds.

Two limitations remain. First, RELAY relates perturbations only through curated knowledge, so poorly annotated perturbations receive little or no transfer. Second, RELAY also uses response profiles and, on PerturbQA, effect statistics from the same screens as the labels, so its comparisons with LLM-based methods are system-level, as Table A.2 details.

Reproducibility Statement

Response construction and supervision are specified in Appendix A, and the evaluation protocol in Appendix D.1. Algorithm 1 and Table C.2 summarize fitting and calibration; Appendix D.4.1 specifies transfer across cell lines. We will release the code, full implementation notes, frozen models, and evaluation scripts upon publication.

AI Use Statement

AI was used to assist with code generation, language polishing of the manuscript, and searches for relevant literature and related work.

References

Britt Adamson, Thomas M. Norman, Marco Jost, et al. A multiplexed single-cell CRISPR screening platform enables systematic dissection of the unfolded protein response. *Cell*, 167(7):1867–1882.e21, 2016. doi: 10.1016/j.cell.2016.11.048.

Abhinav K. Adduri, Dhruv Gautam, Beatrice Bevilacqua, et al. Predicting cellular responses to perturbation across diverse contexts with State. *Cell*, 189(19):5914–5931.e20, 2026. doi: 10.1016/j.cell.2026.07.052.

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

Suzi A. Aleksander, James Balhoff, Seth Carbon, et al. The Gene Ontology knowledgebase in 2023. *Genetics*, 224(1):iyad031, 2023. doi: 10.1093/genetics/iyad031. The Gene Ontology Consortium.

Michael Ashburner, Catherine A. Ball, Judith A. Blake, et al. Gene Ontology: tool for the unification of biology. *Nature Genetics*, 25(1):25–29, 2000. doi: 10.1038/75556.

Kay Henning Brodersen, Cheng Soon Ong, Klaas Enno Stephan, and Joachim M. Buhmann. The balanced accuracy and its posterior distribution. In *Proceedings of the 20th International Conference on Pattern Recognition*, pp. 3121–3124, 2010. doi: 10.1109/ICPR.2010.764.

Payal Chandak, Kexin Huang, and Marinka Zitnik. Building a knowledge graph to enable precision medicine. *Scientific Data*, 10:67, 2023. doi: 10.1038/s41597-023-01960-3.

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

Atray Dixit, Oren Parnas, Biyu Li, et al. Perturb-seq: dissecting molecular circuits with scalable single-cell RNA profiling of pooled genetic screens. *Cell*, 167(7):1853–1866.e17, 2016. doi: 10.1016/j.cell.2016.11.038.

Bradley Efron and Robert J. Tibshirani. *An Introduction to the Bootstrap*. Chapman and Hall/CRC, 1994. doi: 10.1201/9780429246593.

Tom Fawcett. An introduction to ROC analysis. *Pattern Recognition Letters*, 27(8):861–874, 2006. doi: 10.1016/j.patrec.2005.10.010.

Jake Fawkes, Liam Hodgson, and Jason Hartford. Knowledge graphs and reasoning LLMs for finding simple yet effective transcriptomic perturbation predictors. *arXiv preprint arXiv:2606.08816*, 2026. URL https://arxiv.org/abs/2606.08816v1.

Gemma Team. Gemma 4 technical report. *arXiv preprint arXiv:2607.02770*, 2026.

Heonjong Han, Jae-Won Cho, Sangyoung Lee, Ayoung Yun, Hyojin Kim, Dasom Bae, Sunmo Yang, Chan Yeong Kim, et al. TRRUST v2: an expanded reference database of human and mouse transcriptional regulatory interactions. *Nucleic Acids Research*, 46(D1):D380–D386, 2018. doi: 10.1093/nar/gkx1013.

Arthur E. Hoerl and Robert W. Kennard. Ridge regression: biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67, 1970. doi: 10.1080/00401706.1970.10488634.

Edward L. Huttlin, Raphael J. Bruckner, Jose Navarrete-Perea, et al. Dual proteome-scale networks reveal cell-specific remodeling of the human interactome. *Cell*, 184(11):3022–3040.e28, 2021. doi: 10.1016/j.cell.2021.04.011.

Kasia Z. Kedzierska, Lorin Crawford, Ava P. Amini, and Alex X. Lu. Zero-shot evaluation reveals limitations of single-cell foundation models. *Genome Biology*, 26:101, 2025. doi: 10.1186/s13059-025-03574-x.

Hyomin Kim, Sang-Yeon Hwang, Jaechang Lim, Yinhua Piao, Yunhak Oh, Woo Youn Kim, Chanyoung Park, Sungsoo Ahn, and Junhyeok Jeon. Progressive multi-agent reasoning for biological perturbation prediction. *arXiv preprint arXiv:2602.07408*, 2026. URL https://arxiv.org/abs/2602.07408v2. Version 2.

Romain Lopez, Jeffrey Regier, Michael B. Cole, Michael I. Jordan, and Nir Yosef. Deep generative modeling for single-cell transcriptomics. *Nature Methods*, 15(12):1053–1058, 2018. doi: 10.1038/s41592-018-0229-2.

Marija Milacic, Deidre Beavers, Patrick Conley, et al. The Reactome pathway knowledgebase 2024. *Nucleic Acids Research*, 52(D1):D672–D678, 2024. doi: 10.1093/nar/gkad1025.

Sophia Müller-Dott, Eirini Tsirvouli, Miguel Vazquez, Ricardo O. Ramirez Flores, Pau Badia-i-Mompel, Robin Fallegger, Dénes Türei, Astrid Lægreid, and Julio Saez-Rodriguez. Expanding the coverage of regulons from high-confidence prior knowledge for accurate estimation of transcription factor activities. *Nucleic Acids Research*, 51(20):10934–10949, 2023. doi: 10.1093/nar/gkad841.

Ajay Nadig. Differential expression summary statistics for Perturb-Seq datasets analyzed in Nadig et al., 2025 (Nature Genetics). Figshare, 2025. URL https://doi.org/10.6084/m9.figshare.29498366.v1. Version 1.

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

Lawrence Phillips, Marc Boubnovski Martell, Aditya Misra, Josefa Lia Stoisser, Cesar A. Prada-Medina, Rory Donovan-Maiye, and Kaspar Märtens. SynthPert: Enhancing LLM biological reasoning via synthetic reasoning traces for cellular perturbation prediction. In *Machine Learning for Genomics Explorations Workshop at ICLR*, 2026. URL https://arxiv.org/abs/2509.25346v2. arXiv:2509.25346v2.

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

Yusuf Roohani, Kexin Huang, and Jure Leskovec. Predicting transcriptional outcomes of novel multigene perturbations with GEARS. *Nature Biotechnology*, 42:927–935, 2024. doi: 10.1038/s41587-023-01905-6. Published online 17 August 2023.

Craig Saunders, Alexander Gammerman, and Volodya Vovk. Ridge regression learning algorithm in dual variables. In *Proceedings of the Fifteenth International Conference on Machine Learning*, pp. 515–521, 1998.

Karen Spärck Jones. A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, 28(1):11–21, 1972. doi: 10.1108/eb026526.

Damian Szklarczyk, Rebecca Kirsch, Mikaela Koutrouli, 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(D1):D638–D646, 2023. doi: 10.1093/nar/gkac1000.

Damian Szklarczyk, Katerina Nastou, Mikaela Koutrouli, Rebecca Kirsch, Farrokh Mehryary, Radja Hachilif, Dewei Hu, Matteo E. Peluso, Qingyao Huang, Tao Fang, Nadezhda T. Doncheva, Sampo Pyysalo, Peer Bork, Lars J. Jensen, and Christian von Mering. The STRING database in 2025: protein networks with directionality of regulation. *Nucleic Acids Research*, 53(D1):D730–D737, 2025. doi: 10.1093/nar/gkae1113.

Sven Thiele, Luca Cerone, Julio Saez-Rodriguez, Anne Siegel, Carito Guziołowski, and Steffen Klamt. Extended notions of sign consistency to relate experimental data to signaling and regulatory network topologies. *BMC Bioinformatics*, 16:345, 2015. doi: 10.1186/s12859-015-0733-7.

George Tsitsiridis, Ralph Steinkamp, Madalina Giurgiu, et al. CORUM: the comprehensive resource of mammalian protein complexes–2022. *Nucleic Acids Research*, 51(D1):D539–D545, 2023. doi: 10.1093/nar/gkac1015.

Dénes Türei, Jonathan Schaul, Nicolás Palacio-Escat, Balázs Bohár, Yunfan Bai, Francesco Ceccarelli, et al. OmniPath: integrated knowledgebase for multi-omics analysis. *Nucleic Acids Research*, 54 (D1):D652–D660, 2026. doi: 10.1093/nar/gkaf1126.

Ramon Viñas Torné, Maciej Wiatrak, Zoe Piran, Shuyang Fan, Liangze Jiang, Sarah A. Teichmann, Mor Nitzan, and Maria Brbić. Systema: a framework for evaluating genetic perturbation response prediction beyond systematic variation. *Nature Biotechnology*, 44:1050–1059, 2026. doi: 10.1038/s41587-025-02777-8. Published online 25 August 2025.

Zhenyu Wang, Geyan Ye, Wei Liu, and Man Tat Alexander Ng. AROMA: Augmented reasoning over a multimodal architecture for virtual cell genetic perturbation modeling. In *Findings of the Association for Computational Linguistics: ACL 2026*, 2026. URL https://arxiv.org/abs/2604.20263v1.

Zhijian Wei, Runze Ma, Zichen Wang, Zhongmin Li, Shuotong Song, and Shuangjia Zheng. VCWorld: A biological world model for virtual cell simulation. In *International Conference on Learning Representations*, 2026. URL https://arxiv.org/abs/2512.00306v2.

Frederik Wenkel, Wilson Tu, Cassandra Masschelein, et al. TxPert: using multiple knowledge graphs for prediction of transcriptomic perturbation effects. *Nature Biotechnology*, 2026. doi: 10.1038/s41587-026-03113-4.

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

Menghua Wu, Russell Littman, Jacob Levine, Lin Qiu, Tommaso Biancalani, David Richmond, and Jan-Christian Hütter. Contextualizing biological perturbation experiments through language. In *International Conference on Learning Representations*, 2025. URL https://arxiv.org/abs/2502.21290v1.

Xinyu Yuan, Xixian Liu, Jianan Zhao, Yashi Zhang, Hongyu Guo, and Jian Tang. Plausibility is not prediction: Contrastive evidence for LLM-based cellular perturbation reasoning. *arXiv preprint arXiv:2606.01042*, 2026. URL https://arxiv.org/abs/2606.01042v1. Version 1.

Jesse Zhang, Airol A. Ubas, Richard de Borja, et al. Tahoe-100M: A giga-scale single-cell perturbation atlas for context-dependent gene function and cellular modeling. *bioRxiv*, 2025. doi: 10.1101/2025.02.20.639398.

A BENCHMARK, SUPERVISION, AND RELATED WORK

This appendix provides supplementary material for Section 2.

A.1 BENCHMARK AND AVAILABLE SUPERVISION

PerturbQA supplies training and test identities and released DE/DIR labels for each cell line (Wu et al., 2025). We distinguish its *labeled support set*, containing selected perturbation–gene pairs, from the *response matrix*, which contains continuous measurements for training perturbations across the genes in the released single-cell expression matrix. These measurements come from the genome-wide K562 and the RPE1 screens of Replogle et al. (2022) and the HepG2 and Jurkat screens of Nadig et al. (2025). A pair absent from the label table is unlabeled, not negative. RELAY uses both sources of supervision: the response regressions fit the matrix, whereas the neural readout and calibration use the labeled pairs. Table A.1 gives the populations for each cell line, and Table A.2 compares supervision across methods.

PerturbQA data and labels. Expression is normalized to 10,000 counts per cell and $\log(1+x)$-transformed before averaging. Responses subtract batch-matched control means, weighted by each perturbation's cell fractions; magnitudes are taken after aggregation. Auxiliary effects and standard errors use the version-1 tables of Nadig (2025): GenomeWide for K562 and Essential for the other cell lines. Only finite effects with positive standard errors are retained; p-value tables are not used. All response and effect supervision is restricted to training perturbations.

PerturbQA uses normalized expression and differential-expression testing against non-targeting controls (Wu et al., 2025, App. A.2). K562 positive pairs must satisfy the DE criterion in both the genome-wide and the essential-gene screens; non-DE pairs are drawn from pairs not called DE in either. Other cell lines use adjusted significance cutoffs of $p_{\mathrm{adj}} < 0.01$ for positives and $p_{\mathrm{adj}} > 0.1$ for negative sampling. The benchmark retains up to 20 top-ranked DE genes and samples 100 non-DE genes per selected perturbation.

Test-pair exposure across cell lines. PerturbQA holds out perturbations within each cell line, but 82%–96% of each cell line's test perturbations appear in another cell line's training set. Some test queries therefore reappear in training data from other cell lines as the same ordered perturbation–gene pair: 4,717 of 116,041 DE queries (4.06%) and 3,098 of 16,208 DIR queries (19.11%). On these queries, copying the majority label of the same pair from other cell lines gives accuracy/Macro-F1 of 0.949/0.942 for DE and 0.983/0.982 for DIR. A model trained on all four cell lines (Wang et al., 2026) may therefore score well on this subset without knowing the response specific to the queried cell line. Our held-out-cell-line analysis also reports results on the non-overlapping pairs (Appendix D.4.1).

Table A.1: **Split sizes by cell line.** Entries are train / test counts; mixed-label entries count test queries / genes (Appendix A.1).

Cell line	Perturbations train / test	Response genes	DE pairs train / test	DE mixed queries / genes	DIR pairs train / test	DIR mixed queries / genes
K562	1,564 / 267	8,248	134,467 / 23,212	5,641 / 665	16,861 / 3,119	1,558 / 150
HepG2	1,086 / 278	9,624	101,140 / 25,749	5,163 / 707	14,257 / 3,603	1,712 / 166
Jurkat	1,227 / 313	8,882	113,684 / 29,138	6,321 / 750	15,937 / 4,121	2,599 / 181
RPE1	1,596 / 406	8,749	149,147 / 37,942	8,611 / 684	21,287 / 5,365	2,845 / 193
Total			498,438 / 116,041	25,736 / 2,806	68,342 / 16,208	8,714 / 690

GeneTAK data and labels. GeneTAK was curated by VCWorld from the Tahoe-100M single-cell atlas (Wei et al., 2026; Zhang et al., 2025) and covers five cell lines: C32, HepG2/C3A, HOP62, Hs 766T, and PANC-1. We use the released DE and DIR label files and the released 2,000 highly variable genes. VCWorld calls DE with a Wilcoxon test of each drug against vehicle controls (FDR at most 0.05 and $|\log \mathrm{FC}| \geq 0.25$), and DIR labels the direction of DE-positive pairs. All released doses of a drug form one perturbation. After drug names are normalized, 379 drugs appear in the label files of all five cell lines and both tasks.

Table A.2: **Supervision and external information used by the evaluated methods.** Entries describe our evaluated instances.

Method	Continuous experimental data	Discrete experimental outcomes	External knowledge / pretrained information
Gene prior	None	Training labels from the same cell line, pooled by gene and task	None
GEARS	Perturbed and control cells from the same cell line's official training split; a held-out 10% of training perturbations supplies validation	Validation labels calibrate binary thresholds	GO-derived perturbation graph and expression-derived coexpression graph; no generative LLM in GEARS (Roohani et al., 2024; Wu et al., 2025)
STATE	Training expression and non-targeting controls from the same cell line; a held-out 10% of training perturbations supplies validation. GeneTAK: training-drug expression and vehicle controls	Validation or development-drug labels calibrate binary thresholds	PerturbQA: cell embeddings from the pretrained SE-600M model and GenePT text embeddings from a pretrained language model. GeneTAK: Morgan fingerprints; retrained from scratch
SUMMER	None	Retrieved PerturbQA training outcomes from the same cell line and task	Biological descriptions and KG retrieval; Gemma4-26B-A4B-it produces three sampled answers
AROMA	None	None at inference; the authors' released checkpoint was trained on outcome-conditioned PerturbReason synthetic explanations	Text, graphs, protein-sequence features, and a pretrained generative LLM; published teacher is DeepSeek-R1 (Wang et al., 2026)
RELAY	Perturbed and control profiles from the same cell line's official training split; effect estimates and standard errors of training perturbations (PerturbQA only). GeneTAK: training-drug profiles and vehicle controls	DE/DIR training labels from the same cell line for readouts and development calibration	GO, Reactome, CORUM, STRING, and BioPlex from the PerturbQA knowledge graph that SUMMER also uses; no LLM or pretrained text embedding. GeneTAK compound targets come from the Tahoe-100M release and PrimeKG

A.2 Evaluation Metrics

Score resolution and pgAUROC. AUROC measures pairwise ranking and gives half credit to ties (Fawcett, 2006), so a gene's AUROC equals its BAcc when scores are binary, and the gap between pgAUROC and pgBAcc reflects score resolution. pgAUROC ranks each method's native scores, whose resolution differs: response predictors give effectively continuous scores, whereas SUMMER's mean of three categorical answers takes at most four values and our AROMA reproduction gives a single ternary answer. To test whether pgAUROC rewards resolution even when decisions do not improve, Figure A.1 raises SUMMER's resolution by adding votes. We collected seven independent answers for each of the 3,119 K562 DIR queries with the same prompts, model (Gemma4-26B-A4B-it), and sampling settings (temperature 1.0, top-p 0.95, top-k 64, at most 8,192 tokens, 2,048-token thinking budget).

A query's score is the mean of the valid votes in a subset, and a score of at least 0.5 maps to up; each statistic is averaged over all k-vote subsets. From $k = 1$ to $k = 7$ votes, pgAUROC rises

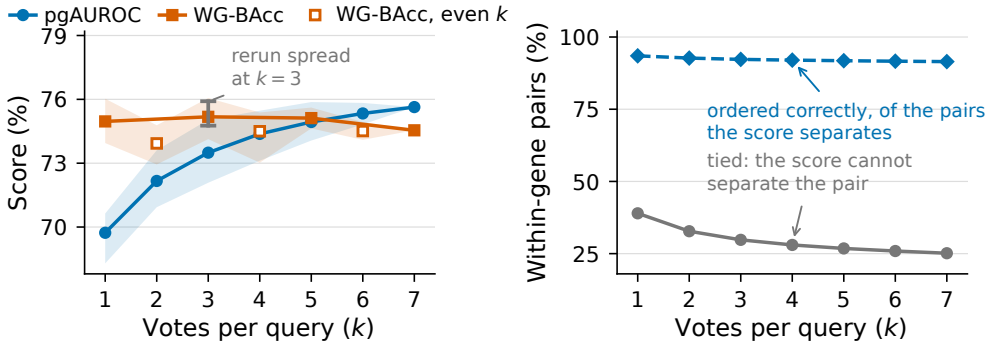


Figure A.1: **Ranking resolution and decision audit.** Left: pgAUROC and WG-BAcc; bands span vote subsets, and the gray bar compares two independent three-vote runs. Open even-k points break ties toward the majority class and are not directly comparable to odd-k decisions. Right: share of tied pairs and ordering accuracy among untied pairs.

from 69.73% to 75.64%, whereas at odd k WG-BAcc stays between 74.54% and 75.18% and full-population accuracy between 86.85% and 87.14%. Because a k-vote mean takes at most $k+1$ values, the share of tied within-gene positive–negative pairs falls from 38.96% to 25.14% as k grows, while ordering accuracy among untied pairs falls slightly (93.48% to 91.48%). More votes resolve ties, which receive half credit under AUROC; decision quality remains stable. Two independent three-vote runs differ by 1.15 WG-BAcc points, more than the range across odd vote counts, so the changes in WG-BAcc with k are within sampling noise. Even vote counts admit exact 0.5 scores, which the rule assigns to up, the majority class (2,530 of 3,119 queries), so the plotted WG-BAcc line connects only odd counts.

Support weighting of WG-BAcc. The gene weight w_g of WG-BAcc in Equation (1) is the inverse of a variance bound. If observations within each class are independent Bernoulli trials, the variance of a gene's estimated BAcc is one quarter of the sum of the sensitivity and specificity variances; with n^+ positive and n^- negative queries, it is bounded by

$$\mathrm{Var}(\widehat{\mathrm{BAcc}}) \le \frac{1}{16}\left(\frac{1}{n^+} + \frac{1}{n^-}\right). \tag{8}$$

The weight w_g in Equation (1) is proportional to the inverse of this bound and equals half the harmonic mean $H_g = 2n^+n^-/(n^+ + n^-)$ of the two counts, so weighting by H_g gives the same WG-BAcc. Because the harmonic mean is dominated by the smaller class, which limits how precisely BAcc_g can be estimated, support weighting favors genes with many queries in both classes.

The 1,601 of 2,806 mixed-label DE genes with $H_g \le 2$, most of which have a single query in one class, account for 76% of the aggregate variance bound under equal weighting but only 34% under support weighting.

pgBAcc and WG-BAcc thus share decisions and differ only in gene weights, so reporting all three metrics separates ranking quality, decision quality, and the influence of low-support genes.

A.3 Additional Related Work

This subsection expands the related work of Section 2.3, following its grouping of methods.

Under the PerturbQA protocol, mechanistic reasoning models generate explanations over biological relations before answering. AROMA and SynthPert train on synthetic reasoning traces, whereas VCWorld and PBio-Agent ground their reasoning in curated knowledge (Wang et al., 2026; Phillips et al., 2026; Wei et al., 2026; Kim et al., 2026). Label retrieval instead borrows observed outcomes of related training pairs: SUMMER passes them to an LLM alongside retrieved descriptions (Wu et al., 2025), and CORE aggregates contrastive evidence to reduce gene-frequency confounding (Yuan et al.,

2026). RELAY uses knowledge to relate perturbations and transfers dense response measurements between them.

GEARS combines coexpression and GO graphs (Roohani et al., 2024), and STATE models cellular contexts (Adduri et al., 2026). TxPert learns perturbation embeddings from multiple knowledge graphs, adds them to an encoded basal state, and decodes the combined representation into expression profiles (Wenkel et al., 2026). This latent transfer differs from directly combining measured training responses. These models output continuous expression profiles. Applying them to the DE and DIR tasks in PerturbQA and GeneTAK requires converting those profiles into discrete decisions (Wu et al., 2025; Wei et al., 2026); Appendix D.1 specifies our conversion and calibration procedures. RELAY instead separates response transfer from task-specific readouts fitted to the discrete labels.

Graph-neighbor averaging reuses measured training responses to predict continuous expression profiles, averaging the training deltas of a perturbation's knowledge-graph neighbors and optionally refining the neighbors with an LLM (Fawkes et al., 2026). RELAY shares this direct use of measured responses but transfers contrasts around gene means by kernel ridge regression, whose coefficients account for redundancy among related perturbations, and reads the transferred responses out as DE and DIR decisions.

Simple baselines often match complex models on average responses (Ahlmann-Eltze et al., 2025; Wong et al., 2025), and zero-shot evaluations show that single-cell foundation models can underperform them (Kedzierska et al., 2025). Systema separates perturbation-specific effects from systematic variation shared across perturbations (Viñas Torné et al., 2026), while PerturbQA evaluates ranking within each readout gene (Wu et al., 2025). We retain that ranking view and complement it with global and within-gene decision metrics. Our error decomposition (Section 3.3) further separates contrast alignment from gene-level and amplitude error.

TRRUST, STRING, and OmniPath supply signed relations (Han et al., 2018; Szklarczyk et al., 2025; Türei et al., 2026; Müller-Dott et al., 2023), and sign-consistency methods test whether observed changes agree with network constraints (Thiele et al., 2015). Our signed-path audit (Appendix B.1) instead uses these relations to predict held-out directions and finds that composed signs conflict for most covered queries.

Choice of baselines. Our expression-based baselines are those that LLM-based methods on these benchmarks compare against. PerturbQA, AROMA, and SynthPert report GEARS, and AROMA and VCWorld report STATE (Wu et al., 2025; Wang et al., 2026; Phillips et al., 2026; Wei et al., 2026), so we retrain both on the official splits. We do not reproduce similarity-based profile predictors or TxPert, because their reported systems depend on resources that are not publicly available. The full TxPert and graph-neighbor models use two graphs derived from perturbation screens that the TxPert release describes as proprietary (Wenkel et al., 2026; Fawkes et al., 2026). We instead test their shared design, averaging the responses of neighboring perturbations, in the controlled comparisons of Appendix D.3.2.

B Diagnostic Experiments

This appendix collects the diagnostic analyses behind Section 3, with one subsection each for mechanistic reasoning, label retrieval, and continuous prediction.

B.1 Mechanistic Reasoning

Table B.1: **Knowledge resources in the coverage audit.** Context resources describe shared function or complex membership; signed resources specify the direction of a regulatory effect. The audit uses GO (Ashburner et al., 2000; Aleksander et al., 2023), CORUM (Tsitsiridis et al., 2023), STRING (Szklarczyk et al., 2025), TRRUST (Han et al., 2018), OmniPath (Türei et al., 2026), and CollecTRI (Müller-Dott et al., 2023).

Resource	Role	Relations used	Version
GO	Context	Biological-process annotations	2026-05-21; ontology 2026-07-26
CORUM	Context	Protein-complex membership	5.3
STRING	Signed	Regulatory links and expression events	12.5
TRRUST	Signed	TF–target relations with a known sign	v2
OmniPath signaling	Signed	Curated causal signaling, acting on protein activity	Queried 2026-09-14
CollecTRI	Signed	TF–target records, acting on expression	Queried 2026-09-14
HGNC	Identifiers	Approved gene symbols	2026-09-07

The coverage audit distinguishes functional or complex context from signed regulatory relations (Table B.1). We analyze the OmniPath signaling and CollecTRI layers separately. In these layers, a record is signed when exactly one of its consensus stimulation and inhibition flags is set. We exclude records with a protein-complex endpoint because they do not specify a relation between individual genes. Gene identifiers are mapped to approved HGNC symbols, and queries with unmapped genes remain in all denominators.

For DIR, GO BP annotations cover both genes in 94.96% of queries and CORUM membership covers both in 42.86%. Coverage by signed relations at both genes ranges from 1.35% (TRRUST) to 11.79% (STRING, confidence at least 0.4). We enumerate simple paths of at most three edges using the signed relations above, retaining STRING relations at confidence at least 0.4 and excluding the query genes as intermediate nodes (Table B.2). For the expression-regulatory criterion, the final edge must be a STRING expression event, a signed TRRUST relation, or a CollecTRI relation. Short directed paths cover 17.99% of DE-positive queries, which form the DIR set, and 9.38% of non-DE queries.

Table B.2: **Coverage of query-specific directional evidence.** Percentages are averaged with equal weight across cell lines. Counts are pooled across cell lines, with 116,041 DE and 16,208 DIR queries in total.

Evidence condition	DE (%)	DIR (%)	Pairs (DE / DIR)
Local directed relations at both endpoints	26.27	35.91	30,721 / 5,895
Directed connection, at most three edges	10.59	17.99	12,315 / 2,957
Local relations, including gene expression regulation	18.57	31.45	21,738 / 5,183
Short connection ending in expression regulation	7.73	15.87	9,002 / 2,604

The sign-propagation experiment in Section 3.1 uses the signed relations above, without query-specific effect magnitudes or DE probabilities. Each path must end in a transcriptional relation with a known sign from TRRUST or CollecTRI. This stricter final-edge requirement reduces coverage from 2,604 to 1,955 DIR queries. For a path $\mathcal{P}$, the predicted response direction is $-\prod_{e\in\mathcal{P}} \zeta_e$, where $\zeta_e \in \{-1, +1\}$ is the sign of edge e and the initial minus sign represents reduced activity of the perturbed gene under CRISPRi knockdown. The rule propagates signs across activity and expression levels.

The consistency rule assigns a direction when all possible terminal signs agree. The majority rule assigns one vote to each distinct node sequence, excluding sequences whose edge records allow both

terminal signs. It selects the direction with more votes and falls back to the training-set majority DIR label in a tie.

Table B.3: **Coverage and accuracy of path-sign aggregation on DIR.** Coverage and BAcc are computed on queries resolved by path signs before fallback. Mean BAcc weights cell lines equally; pooled BAcc combines these queries across cell lines.

Measure	Consistency	Majority
Queries resolved by path signs	660	1,618
Coverage of all 16,208 DIR queries (%)	4.07	9.98
Mean BAcc across cell lines (%)	48.89	53.69
Pooled BAcc (%)	45.12	49.90

Path-based predictions have low coverage and BAcc close to chance on DIR (Table B.3).

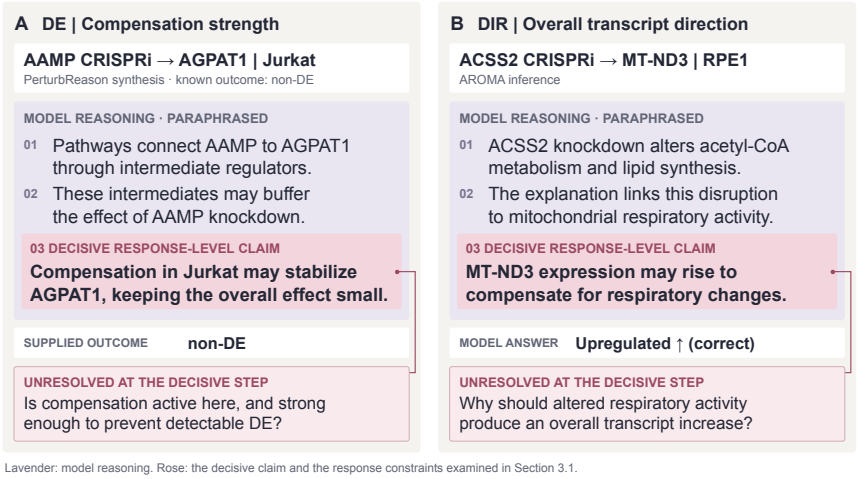


Figure B.1: **Decisive response-level claims in mechanistic reasoning.** Steps are paraphrased from the AROMA examples (Wang et al., 2026). Rose highlights mark the claims used to reach the answer, and callouts identify the missing evidence. **(A)** The AAMP–AGPAT1 explanation is generated with a known non-DE outcome and assumes effective compensation in Jurkat. **(B)** The ACSS2–MT-ND3 inference assumes that altered respiratory activity leads to an overall increase in MT-ND3 expression in RPE1.

PerturbReason and the final SynthPert synthesis strategy give the teacher the observed outcome before it generates the explanation (Wang et al., 2026; Phillips et al., 2026). At inference, the student receives the query inputs and must predict the outcome.

For AAMP→AGPAT1 in Jurkat (Figure B.1A), AROMA v1 (Wang et al., 2026, App. F.3, with sentence-level source tracing in App. G) generates a PerturbReason explanation after the non-DE outcome is supplied. The instruction permits plausible buffering, compensation, or feedback to fill gaps in the evidence. The output invokes "compensatory mechanisms that stabilize AGPAT1's expression." The source descriptions make compensation plausible but do not establish that it is strong enough to keep the AGPAT1 response below the benchmark's detection criterion.

In the ACSS2→MT-ND3 inference example in RPE1 (Figure B.1B), AROMA's explanation uses metabolic and respiratory-chain relations supported by the source descriptions (Wang et al., 2026, AROMA v1, App. E.2). The model then assumes that MT-ND3 "may increase to compensate" for the respiratory changes. The prediction matches the measured increase, but the decisive step depends on this assumption.

B.2 Label Retrieval

This subsection supports Section 3.2. All interventions use SUMMER's public pipeline with Gemma4-26B-A4B-it (Gemma Team, 2026). On DIR, the Full row reuses the main-comparison run and each deletion condition reruns the same sampling configuration: three sampled answers with native thinking enabled (2,048-token budget), temperature 1.0, top-p 0.95, top-k 64, and an 8,192-token limit. The DE runs use one answer without thinking, temperature 0.6, top-p 0.95, top-k 20, and a 4,096-token limit. The deletion runs retain the written reasoning questions in the prompt, and each intervention is compared with its own Full row.

Every condition with references retains the query and reference gene names, the cell line, and the question sequence. *Full* contains descriptions and observed outcomes; *No descriptions* removes the description prose; *No outcomes* removes the observed-outcome lines; *Query only* removes the reference instances and descriptions and states that no related observations were retrieved.

Table B.4: **Reference interventions (deletion runs, means across four cell lines).** ΔWG-BAcc is relative to Full with 95% paired perturbation-cluster intervals, in percentage points. Macro-F1 uses full populations.

Task	Input	WG-BAcc	ΔWG-BAcc [95% CI]	Macro-F1
DE	Full reference	0.5849	—	0.7895
	No descriptions	0.5860	$+0.11$ $[-0.65,\ 1.06]$	0.7865
	No outcomes	0.5378	-4.71 $[-5.71,\ -3.87]$	0.4590
	Query only	0.5000	-8.49 $[-9.49,\ -7.76]$	0.4626
DIR	Full reference	0.7019	—	0.8215
	No descriptions	0.6886	-1.33 $[-2.41,\ -0.27]$	0.8215
	No outcomes	0.5083	-19.36 $[-22.08,\ -17.82]$	0.4218
	Query only	0.5061	-19.59 $[-22.45,\ -17.95]$	0.4442

Because the retrieval index is fixed, the no-descriptions condition measures the contribution of the description prose given the retrieved references, not its role in choosing them.

Identity–outcome exchange. In the Jurkat DIR exchange experiment, 1,332 of the 4,121 queries have references of opposite labels that can be exchanged while matching neighbor status and repetition counts, and each is run under four conditions: unchanged, outcomes exchanged, identities exchanged, and both exchanged. Exchanging only outcomes or identities breaks the pairing while preserving label counts; exchanging both restores each tuple and changes only its order. The other 2,789 queries have identical inputs in every condition and share one prediction, and evaluation uses all 4,121 queries. The unchanged, outcome-exchange, identity-exchange, and joint-exchange WG-BAcc values are 0.6703, 0.6221, 0.6348, and 0.6649, with paired changes of -4.83 (95% CI, -6.82 to -3.03), -3.56, and -0.55 points (95% CI, -1.56 to 0.61).

B.3 Continuous Response Predictors

This subsection supports Section 3.3. Unless stated otherwise, analyses use model predictions on the 1,264 test perturbations over the genes shared by RELAY's response matrix and STATE's decoder (8,198–9,622 per cell line). Intervals come from 2,000 perturbation-cluster bootstrap replicates (seed 17), each averaging cell lines with equal weight.

In addition to the DIR signs and within-gene DE ranking reported in Section 3.3, a single global threshold on measured $|r(p, g)|$ gives DE Macro-F1 0.946 and WG-BAcc 0.924. Because the threshold is chosen using the test labels, this is an oracle diagnostic.

Table B.5 indicates that the perturbation-specific signal recovered by GEARS and STATE is outweighed by prediction errors, while RELAY converts response transfer into a net gain over gene means.

For the decomposition in Equation (2), let $\widehat{m}$ and m average the predicted and measured responses over test perturbations, and write $\widehat{\delta} = \widehat{r} - \widehat{m}$ and $\delta = r - m$. With overlines averaging over

Table B.5: **Signed response-profile prediction on test perturbations.** Means weight cell lines equally and cover 1,264 test perturbations. Errors are averaged over perturbations and genes against measured batch-matched $\log(1 + \mathrm{CP10K})$ responses; MSE is $\times 10^3$ and MAE $\times 10^2$. Pearson r is computed per perturbation across genes and per gene across perturbations, excluding constant vectors; dashes mark undefined values. Gene means predict each gene's mean training response. GEARS is retrained on the official split (Appendix D.1); STATE is the main-comparison reproduction.

Predictor	MSE ↓	MAE ↓	Per-perturbation r ↑	Per-gene r ↑
RELAY	**4.849**	**4.565**	**.5165**	**.3141**
Gene means	6.154	4.962	.4190	—
STATE	6.504	5.041	.3335	.1811
GEARS	6.851	5.422	.3019	.0535
Zero response	7.905	5.352	—	—

test perturbations and genes, $\sigma_\delta^2 = \overline{\delta^2}$, $\sigma_{\widehat{\delta}}^2 = \overline{\widehat{\delta}^2}$, $\rho = \overline{\widehat{\delta}\,\delta}/(\sigma_{\widehat{\delta}}\sigma_\delta)$, and $\eta = \sigma_{\widehat{\delta}}/\sigma_\delta$. Because $\sum_p(\widehat{\delta} - \delta) = 0$ for every gene, the cross term between gene and contrast errors vanishes.

The remaining contrast error, for $\sigma_\delta, \sigma_{\widehat{\delta}} > 0$, expands as

$$\overline{(\widehat{\delta} - \delta)^2} = \sigma_\delta^2(1 + \eta^2 - 2\rho\eta) = \sigma_\delta^2\big[(1 - \rho^2) + (\eta - \rho)^2\big]. \tag{9}$$

Together with the gene-level error, this gives the three terms reported in Table B.6: the gene term, the unexplained-contrast term $\sigma_\delta^2(1 - \rho^2)$, and the amplitude-error term $\sigma_\delta^2(\eta - \rho)^2$. At fixed gene component, the difference from zero-contrast error is $\sigma_\delta^2\eta(\eta - 2\rho)$, so predicted contrasts lower the error only when $\eta < 2\rho$. If predicted contrasts vanish, correlation is undefined (shown as a dash in the table); setting $\eta = \rho = 0$ solely for evaluating the identity allocates σ_δ^2 to unexplained contrast and zero to amplitude error. Adding a per-gene constant or rescaling $\widehat{\delta}$ by a positive factor leaves signed within-gene rankings unchanged.

Table B.6: **Decomposition of held-out profile error.** Means weight cell lines equally on test perturbations; errors $\times 10^3$. Slope: regression of the predictor's gene component on the training gene means across genes. Brackets: 95% perturbation-bootstrap intervals. Slope is ideal at 1 and η at ρ.

Predictor	MSE ↓	Gene-level ↓	Amplitude ↓	Unexplained ↓	Slope	ρ ↑	η
Zero response	7.905	1.784	0	6.120	0	—	0
Gene means	6.154	0.033	0	6.120	1	—	0
GEARS	6.851	0.570	0.341	5.940	.829 [.811,.848]	.119 [.106,.133]	.362
STATE	6.504	0.430	0.434	5.640	.569 [.533,.605]	.257 [.231,.281]	.479
RELAY	4.849	0.023	0.053	4.772	1.053 [1.027,1.078]	.470 [.450,.489]	.512

At the cell-line level, the slope of STATE's gene component on the training gene means averages 0.57 and ranges from 0.21 to 0.81, whereas GEARS's deviates from 1 in both directions (0.64–0.74 in three cell lines, 1.27 in RPE1). Contrast amplitudes exceed the MSE-minimizing scale $\eta/\rho = 1$: the ratio averages 1.72 across cell lines for STATE and is 1.6–7.9 for GEARS in HepG2, Jurkat, and RPE1, whereas GEARS's K562 contrasts are uncorrelated with the measured ones ($\rho = 0.001$). RELAY, a ridge regression with gene means as its default, has η/ρ of 1.09. Relative to RELAY, unexplained contrast accounts for 52% of STATE's excess error and 58% of GEARS's.

To isolate the gene-level contribution, we replace each predictor's gene component with the training gene means ($\mu + \widehat{\delta}$). The resulting MSE is 6.107 for STATE and 6.314 for GEARS, compared with 6.154 for the training gene means (all $\times 10^{-3}$). Using diagnostic-specific response-based decision rules, the restoration raises DIR Macro-F1 from 0.733 to 0.754 for STATE and from 0.675 to 0.709 for GEARS. Within-gene decisions do not improve consistently: DE WG-BAcc moves from 0.553 to 0.552 for STATE and from 0.510 to 0.509 for GEARS, and DIR WG-BAcc from 0.609 to 0.594 and from 0.510 to 0.517, with signed DIR pgAUROC unchanged by construction.

A further oracle diagnostic combines this restoration with rescaling the contrasts to their MSE-optimal amplitude, using measured test responses. The MSE falls to 5.67 for STATE and 5.97 for GEARS,

compared with 4.81 for RELAY under the same rescaling (all $\times 10^{-3}$). The remaining gap therefore lies in contrast alignment.

C Notation and RELAY Implementation

C.1 Notation

Table C.1 collects the main symbols used throughout the paper. Hats denote estimates; quantities are specific to each cell line unless stated otherwise.

Table C.1: **Notation for tasks, evaluation, RELAY, and diagnostics.**

Symbol	Meaning
p, g, c, T	Perturbation, readout gene, cell line, and task $T \in \{\mathrm{DE}, \mathrm{DIR}\}$
$\mathcal{P}_{\mathrm{tr}}, \mathcal{P}_{\mathrm{te}}, \mathcal{S}$	Training, test, and current support perturbations; $n = \lvert\mathcal{S}\rvert$
$\mathcal{G}, \mathcal{Q}_{\mathrm{tr}}^T, \mathcal{Q}_{\mathrm{te}}^T$	Readout genes and labeled training/test query sets for task T
$\mathcal{G}_{\mathrm{mix}}, n_g^+, n_g^-, w_g$	Mixed-label test genes, class counts, and the gene weight in WG-BAcc
$\pi(g), \bar{\pi}$	Gene-specific and pooled training positive-label rates
r, R	Signed response and its support matrix $R \in \mathbb{R}^{n \times \lvert\mathcal{G}\rvert}$
μ, ν	Per-gene support means of r and $\lvert r\rvert$
d, a, D, A	Signed/magnitude contrasts and their matrices, shaped as R
$\phi(p), X, V$	Knowledge profile, support design matrix $X \in \mathbb{R}^{n \times V}$, and vocabulary size
k, K, k_p	Scalar kernel, support Gram matrix, and query kernel column
$\alpha, B, W_B, w(p)$	Ridge penalty, target $B \in \{D, A\}$, primal coefficients, and query transfer coefficients (distinct from w_g)
u, v_a	Response coordinates and magnitude gene loadings
h, f, L_{base}	DE features, neural readout, and its base logit
$\tau(g), s_T, t_T$	Signed-error scale, task score, and decision threshold for the cell line
F, SE, E, z	Training effect, standard error, standardized-effect target, and its normalized prediction
β, λ, γ	Fitted increment coefficient, selected strength, and product $\gamma = \lambda\beta$
m, δ	Test-centered gene mean and response contrast; diagnostic only, unlike the support-based μ, d
$\rho, \eta, \sigma_\delta$	Pooled contrast correlation, predicted/observed amplitude ratio, and observed contrast standard deviation

C.2 Fitting and Prediction

Algorithm 1 runs separately for each cell line. Out-of-fold (OOF) predictions exclude the predicted perturbation from the fit. Every regression fit rebuilds its vocabulary, IDF weights, and response means on its support only.

Algorithm 1 RELAY fitting and prediction.

Require: Training responses and DE/DIR labels; training effects (F, SE); knowledge records; four perturbation groups; query pairs

Ensure: DE/DIR scores and three-class decisions

1: **for all** candidate penalties α and outer query groups among the first three groups **do**
2: Set the support to the other two development groups.
3: Inner-cross-fit regressions for D, A, E on this support.
4: Fit error scales, PCA, input transforms, and the neural base using support OOF predictions.
5: Refit support regressions and score the outer queries with support-fitted transforms.
6: **end for**
7: Pool outer predictions; select α, fit β, and calibrate $\lambda, t_{\mathrm{DE}}, t_{\mathrm{DIR}}$.
8: Evaluate the fixed configuration on the fourth group, without retuning.
9: Fit the final readout using four-fold OOF predictions on all training perturbations.

10: Refit upstream regressions on all training perturbations; keep the calibrated auxiliary normalization, γ, and thresholds fixed.
11: **for all** query pairs (p, g) **do**
12: Predict contrasts and z; construct DE features and obtain L_{base}.
13: Compute DE and analytic DIR scores by Equations (7) and (5).
14: Return non-DE if $s_{\text{DE}} \leq t_{\text{DE}}$; otherwise up if $s_{\text{DIR}} > t_{\text{DIR}}$, and down otherwise.
15: **end for**

C.3 Key Configuration

The neural base uses gene- and class-balanced binary cross-entropy with equal total DE/DIR weight; its DIR head is auxiliary and is not used at inference. The increment coefficient β uses support-weighted binary cross-entropy, with zero weight for single-class genes. We exclude missing effect statistics.

Table C.2: **Key settings for RELAY on PerturbQA.** Selection and calibration are performed on development predictions only.

Component	Setting
Ridge	No intercept; selected $\alpha = 1$ (K562, HepG2, Jurkat), 0.1 (RPE1)
PCA and features	32 joint components; 66D task features
Neural base	$66 \to 64 \to 32$ ReLU trunk; separate DE/DIR logit heads
Optimizer	AdamW; learning rate 10^{-3}; weight decay 10^{-4}
Training	Batch 8,192; 40 epochs; seeds 17, 23, 31; average seed scores
Perturbation groups	Proportions $0.4/0.3/0.2/0.1$; shuffle seed 17
Increment strength	$\lambda \in \{0, 0.05, \ldots, 1\}$ maximizes WG-BAcc within a 0.01 Macro-F1 loss from $\lambda = 0$
Model selection	α and task thresholds selected by development Macro-F1

D Evaluation Details and Ablations

This appendix follows Section 5: the evaluation protocol and baselines (Appendix D.1), the full main results (Appendix D.2), the component ablations (Appendix D.3), and the evaluations on a held-out cell line and unseen compounds (Appendix D.4).

D.1 Evaluation Protocol and Baselines

The main experiment scores every test query with the metrics of Appendix A.2. All methods with query-level predictions are scored on the same rows; CORE-Voting uses published aggregates (Table D.1).

Standalone DIR is evaluated on ground-truth DE-positive pairs; three-class prediction instead uses the predicted DE gate of Algorithm 1.

The gene prior provides a reference based on training-label frequencies. For each gene, cell line, and task, $\pi(g)$ is the fraction of training pairs from that cell line and task with a positive label; every test gene has such training pairs, so $\pi(g)$ is always defined. The gene prior predicts the positive label when $\pi(g) > 1/2$. Pooling all training pairs of the cell line and task instead gives the *pooled training rate* $\bar{\pi}$, whose majority label resolves exact ties ($\pi(g) = 1/2$, 235 of 116,041 DE test queries) and serves as the fallback for unanswered queries. For comparison, predicting non-DE for every query gives mean DE Macro-F1 of 0.4626, with cell lines weighted equally.

The expression predictors require calibrated thresholds for DE and DIR decisions. For GEARS, subtracting the control profile from predicted $\log(1 + \text{CP10K})$ profiles gives the responses used for both the binary-task scores and the profiles in Table B.5. GEARS and STATE use absolute log fold change as the DE score and signed log fold change as the DIR score. For each cell line and task, the binary threshold maximizes Macro-F1 on the 10% of training perturbations held out from model fitting. GEARS also uses this subset for epoch selection. STATE scores are represented as empirical

Table D.1: **Reproduction strategies for the main PerturbQA baselines.**

Method	Strategy	Implementation and settings
GEARS	Retrain the authors' package	`cell-gears` 0.1.2 with default hyperparameters (20 epochs, hidden size 64, batch size 32, learning rate 10^{-3}). Official training perturbations for each cell line, without highly-variable-gene pre-filtering. K562 uses the genome-wide screen because the essential-gene screen covers only about half of its training perturbations.
STATE	Train the transition model and decoder from scratch	PerturbQA-aligned split with validation seed 42 and a negative-binomial gene decoder. Inputs are GenePT model-3 text embeddings for perturbations and cell embeddings from the frozen, pretrained SE-600M State Embedding model.
AROMA	Use the public checkpoint without additional training	The authors' checkpoint (Wang et al., 2026) was trained on all four cell lines (Appendix A.1). A single parsed ternary answer is projected to DE or DIR; a non-DE answer to a DIR query provides no direction.
SUMMER	Run the public code with a different backbone	Gemma4-26B-A4B-it (Gemma Team, 2026) with the three-answer aggregate. The labeled retrieval index uses the queried cell line's official training split for the queried task.
CORE-Voting	Use published results	The official implementation and ReasonKG resource were unavailable to us. We use the reported DE results (Yuan et al., 2026) and reconstruct Macro-F1 from accuracy and positive-class F1.

quantiles of the validation-score distribution, and pairs outside STATE's output gene space or lacking a perturbation representation are unanswered.

Using the resulting predictions, we compute paired intervals for the main comparison from 2,000 bootstrap replicates (Efron & Tibshirani, 1994), with draws shared across methods. For pgAUROC, pgBAcc, and WG-BAcc, each replicate resamples mixed-label genes within each cell line and retains their recorded metric values and class-support weights; for accuracy and Macro-F1, it resamples query perturbations and carries all their query rows together. Each replicate averages the values across the four cell lines before computing the paired RELAY-minus-baseline difference, and the 2.5th and 97.5th percentiles give the interval. The SUMMER reference interventions (Appendix B.2) instead resample query perturbations within each cell line, carry all query rows of a perturbation together, and recompute eligibility and effective support in each replicate. All intervals are conditional on fixed predictions.

D.2 Main Results by Cell Line

Table D.2: **Main-experiment results by cell line.** All entries include the common fallback; DIR uses the official DE-positive query population. Mean rows weight the four cell lines equally and are computed from unrounded values.

		DE				DIR			
Cell line	Method	Macro-F1	WG-BAcc	pgBAcc	pgAUROC	Macro-F1	WG-BAcc	pgBAcc	pgAUROC
K562	Gene prior	.8458	.5000	.5000	.5000	.7299	.5000	.5000	.5000
	GEARS	.6295	.5001	.4990	.5008	.6323	.5293	.5359	.5160
	STATE	.6942	.6053	.5904	.5859	.7549	.7082	.6739	.7365
	AROMA	.8307	.5173	.5169	.5165	.6665	.5948	.5628	.5530
	SUMMER	.7394	.6004	.5886	.6086	.7952	.7591	.7021	.7342
	RELAY	.8890	.6937	.6855	.8086	.8842	.8498	.8062	.9375
HepG2	Gene prior	.8831	.5000	.5000	.5000	.8245	.5000	.5000	.5000
	GEARS	.7397	.5392	.5308	.5260	.7485	.5188	.5145	.5807
	STATE	.6880	.5753	.5584	.5605	.7667	.6683	.6316	.7076
	AROMA	.8652	.5186	.5155	.5159	.7545	.6180	.5747	.5641
	SUMMER	.7679	.5893	.5872	.5996	.8408	.6853	.6207	.6532
	RELAY	.8911	.6615	.6474	.7858	.8886	.7671	.7029	.8518
Jurkat	Gene prior	.8915	.5000	.5000	.5000	.7753	.5000	.5000	.5000
	GEARS	.7161	.5026	.5025	.5056	.6528	.5020	.4898	.5195
	STATE	.6254	.4908	.4864	.4872	.6572	.5125	.4782	.5078
	AROMA	.8809	.5160	.5147	.5155	.7265	.6305	.5572	.5568
	SUMMER	.7616	.5625	.5522	.5661	.8031	.7114	.6381	.6547
	RELAY	.8948	.6262	.6215	.7478	.8396	.7592	.6924	.8480
RPE1	Gene prior	.9050	.5000	.5000	.5000	.8595	.5000	.5000	.5000
	GEARS	.7956	.5059	.5042	.5249	.8002	.4934	.4899	.5685
	STATE	.7104	.5520	.5445	.5471	.7869	.5795	.5736	.6819
	AROMA	.8702	.5157	.5129	.5129	.7585	.5682	.5417	.5367
	SUMMER	.7553	.5784	.5681	.5821	.8470	.6519	.5939	.6126
	RELAY	.9014	.6482	.6354	.7505	.8765	.6926	.6620	.8071
Mean	Gene prior	.8814	.5000	.5000	.5000	.7973	.5000	.5000	.5000
	GEARS	.7202	.5119	.5091	.5143	.7085	.5109	.5075	.5462
	STATE	.6795	.5559	.5449	.5452	.7414	.6171	.5893	.6584
	AROMA	.8618	.5169	.5150	.5152	.7265	.6029	.5591	.5526
	SUMMER	.7560	.5827	.5740	.5891	.8215	.7019	.6387	.6637
	RELAY	.8941	.6574	.6475	.7732	.8722	.7672	.7159	.8611

Table D.3: **Paired differences, RELAY minus each baseline (percentage points).** Means across cell lines, weighted equally, on the PerturbQA test set, with 95% bootstrap intervals from the protocol above. CORE-Voting lacks query-level predictions and is omitted.

Task	Baseline	ΔMacro-F1	ΔWG-BAcc	ΔpgAUROC
DE	Gene prior	1.27 [0.82, 1.73]	15.74 [15.01, 16.49]	27.32 [26.33, 28.31]
	GEARS	17.38 [16.73, 18.10]	14.55 [13.64, 15.48]	25.88 [24.35, 27.35]
	STATE	21.46 [20.63, 22.36]	10.16 [9.29, 11.01]	22.80 [21.35, 24.27]
	AROMA	3.23 [2.73, 3.73]	14.05 [13.25, 14.87]	25.80 [24.76, 26.86]
	SUMMER	13.80 [13.32, 14.26]	7.48 [6.58, 8.39]	18.41 [17.18, 19.64]
DIR	Gene prior	7.49 [6.16, 8.88]	26.72 [25.33, 27.85]	36.11 [34.38, 37.77]
	GEARS	16.37 [14.89, 17.78]	25.63 [23.92, 27.11]	31.49 [28.62, 34.42]
	STATE	13.08 [11.48, 14.74]	15.00 [13.44, 16.46]	20.27 [17.81, 22.82]
	AROMA	14.57 [13.06, 16.17]	16.43 [14.76, 18.14]	30.85 [28.85, 32.97]
	SUMMER	5.07 [4.01, 6.11]	6.53 [5.28, 7.85]	19.74 [17.37, 22.05]

D.3 Component Ablations

This subsection follows Section 5.3 and examines the contributions and trade-offs of RELAY's components in four steps: the transferred quantity (Appendix D.3.1), the knowledge kernel and response-transfer operator (Appendix D.3.2), the response decomposition and task readouts (Appendix D.3.3), and the resulting trade-off with the gene prior (Appendix D.3.4).

All ablations use the four PerturbQA cell lines. Encoding diagnostics evaluate response error on development folds; the transfer-quantity experiment selects hyperparameters on development labels and uses direct scores, while readout ablations include matched retraining. Downstream performance is evaluated on the test set. Differences are reported in percentage points unless stated otherwise.

Table D.4 extends Table 3 with pgBAcc, pgAUROC, and three-class Macro-F1, averaging cell lines equally. Block (a) compares direct scores without a learned readout. Ridge penalties are selected on development predictions; L_g marks variants restricted to a gene's labeled training perturbations, and the remaining variants use all training perturbations (Appendix D.3.1). Block (b) adds response and readout components sequentially. Its initial row scores DE by $|\widehat{r}|$ and DIR by $\widehat{r}$; the final row is full RELAY. The magnitude view and standardized-effect increment affect only DE scores. Block (c) replaces only the knowledge profile with GenePT. Dashes indicate metrics not recorded.

Table D.4: **Response-transfer, buildup, and representation variants.**

	DE				DIR			Three-class
Configuration	Macro-F1	pgBAcc	WG-BAcc	pgAUROC	Macro-F1	WG-BAcc	pgAUROC	Macro-F1
(a) What is transferred								
Gene prior	.8814	.5000	.5000	.5000	.7973	.5000	.5000	—
Labels (L_g)	.8987	—	.5598	.6553	.8510	.6788	.7622	—
Contrasts (L_g)	.8505	—	.5511	.6282	.8508	.6603	.7443	—
Binarized responses	.8148	—	.6354	.6837	.8454	.7749	.8221	—
Binarized, RELAY penalty	.8344	—	.6362	.7005	.8585	.7757	.8359	—
Contrasts $\lvert\widehat{d}+\mu\rvert$	.8173	—	.6323	.6687	.8659	.7669	.8611	—
Magnitude $\widehat{a}+\nu$	.8544	—	.6389	.7086	—	—	—	—
(b) Buildup								
Direct response readout	.8133	.6188	.6341	.6670	.8649	.7709	.8590	.7427
+ learned DE readout	.8990	.5988	.6068	.7370	.8648	.7703	.8590	.8077
+ centering on gene means	.9016	.6083	.6152	.7465	.8722	.7672	.8611	.8148
+ magnitude view	.9045	.6156	.6228	.7672	.8722	.7672	.8611	.8185
+ standardized effect	.8941	.6475	.6574	.7732	.8722	.7672	.8611	.8115
(c) Representation								
GenePT in RELAY	.8847	—	.6106	.7094	.8292	.7047	.7735	.7794

D.3.1 What Is Transferred

Every variant uses the knowledge kernel of Section 4.2, the fold plan and threshold rule of Appendix C, direct task scores without a learned readout, and development-calibrated thresholds; ridge penalties are selected on pooled development predictions by WG-BAcc. For a gene g, L_g denotes the training perturbations with a label for the task; because DIR labels exist only on DE-positive pairs, a quarter of test genes have at most one or two labeled training perturbations for DIR (medians in Section 3.2). *Labels* transfer y over L_g by kernel ridge regression around the label mean. *Contrasts* transfer $d = r - \mu$, either over L_g or over all training perturbations (RELAY's regression), and score DE by $|\widehat{d} + \mu|$; the magnitude view instead scores DE by $\widehat{a} + \nu$. *Binarized responses* transfer $\mathrm{sign}(r)$ for DIR and $\mathbb{I}[|r| > t]$ for DE over all training perturbations, with t chosen within each support fold to maximize balanced accuracy against the training DE labels; on training pairs, $\mathrm{sign}(r)$ matches every DIR label and the DE indicator ($t = 0.10$–0.14) reaches 95.5–97.2% balanced accuracy.

Transferring contrasts from all training perturbations outperforms transferring labels in WG-BAcc by 8.81 DIR points and 7.25 DE points. At equal support, labels outperform raw contrasts by 1.85 DIR points; over all training perturbations, binarized responses differ from contrasts in WG-BAcc

by $+0.80$ DIR and $+0.31$ DE points but lose 3.91 DIR pgAUROC and 2.05 DIR Macro-F1 points, because 61–76% of training pairs have $|r| < 0.05$ and their binarized signs are weakly determined. Coverage of measured responses therefore carries the within-gene decision gain, and continuous values add ranking resolution and global accuracy.

D.3.2 Knowledge Kernel and Response Transfer

Contrast regression predicts an unseen perturbation's contrasts as a combination of training contrasts weighted through the kernel $k(p, p')$ (Equation (4)). An encoding is therefore useful to the extent that perturbations it deems similar also have similar response contrasts. Genome-scale Perturb-seq supports this premise for shared molecular context: genes whose knockdowns produce similar transcriptional phenotypes frequently belong to the same complex or pathway (Replogle et al., 2022). The knowledge profile exposes such relations directly, since two perturbations are similar only when they share a specific annotation, pathway, complex, or interaction partner, and inverse document frequencies discount widely shared entities.

Ridge predictions depend on ϕ only through inner products (Equation (4)), which decompose by source. Let $\kappa_j(p, p')$ denote the IDF-weighted cosine similarity within knowledge source j, and let $\mathbb{I}_j(p) \in \{0, 1\}$ indicate whether p retains any token of j. Then

$$k(p,p') = \begin{cases} \dfrac{\sum_j \mathbb{I}_j(p)\mathbb{I}_j(p')\kappa_j(p,p')}{\sqrt{h(p)h(p')}}, & h(p)h(p') > 0, \\ 0, & h(p)h(p') = 0, \end{cases} \qquad h(p) = \sum_j \mathbb{I}_j(p). \tag{10}$$

Each nonzero profile remains unit length after a source is missing, but a pairwise similarity is normalized by both profiles' source counts, not just their intersection.

Table D.5 supports source complementarity: STRING alone recovers most of the advantage over GenePT, and the combined encoding retains a further gain. Source-removal effects are broadly consistent across cell lines and response views, with only small exceptions.

Table D.5: **Contrast prediction error by encoding.** Development folds only; mean squared error over all genes and query perturbations ($\times 10^3$), averaged equally over folds and then over cell lines. All encodings share the support partitions, the frozen penalties ($\alpha = 1$; 0.1 for RPE1), and the no-intercept solver. Δ is the relative change from the full knowledge profile.

	Signed contrast		Magnitude contrast	
Encoding	MSE ↓	Δ	MSE ↓	Δ
Knowledge profile (RELAY)	**4.886**	—	**2.924**	—
Without CORUM	4.894	+0.2%	2.927	+0.1%
Without Reactome	4.912	+0.5%	2.933	+0.3%
Without GO	4.938	+1.1%	2.960	+1.2%
Without BioPlex	4.948	+1.3%	2.964	+1.4%
Without STRING	4.999	+2.3%	2.996	+2.4%
STRING only	5.092	+4.2%	3.030	+3.6%
Reactome only	5.304	+8.6%	3.171	+8.4%
BioPlex only	5.443	+11.4%	3.224	+10.3%
CORUM only	5.474	+12.0%	3.265	+11.6%
GO only	5.488	+12.3%	3.280	+12.2%
GenePT substitution	5.600	+14.6%	3.289	+12.5%

Per-source kernel entries correlate at only 0.41–0.65 across sources. Learning per-source weights lowers contrast error by under 0.7%, and the learned weights vary across folds.

To separate the encoding from the rest of the pipeline, we replace only ϕ with L_2-normalized GenePT model-3 text embeddings (Chen & Zou, 2025) and keep response supervision, readouts, and the calibration procedure fixed. The full nested cross-fitting is rerun with the frozen penalties, and DIR thresholds are refitted on pooled development anchor predictions with the Macro-F1 rule of the main model. The substitution worsens contrast recovery in every cell line and response view, and its downstream losses link the representation choice to both DE and DIR decisions (Table D.4c).

Table D.6 tests the premise stated at the start of this subsection. For each representation, we correlate kernel similarity with response similarity over all pairs of training perturbations in a cell line. Response similarity is the Pearson correlation of their signed contrast profiles across genes (0.006 for random pairs). For each representation, we also evaluate anchored ridge regression at the penalty from $\{0.01, 0.03, 0.1, 0.3, 1, 3, 10\}$ that minimizes error on the same development folds. The GO/GO-Jaccard comparison shows that response alignment alone is insufficient to determine contrast-prediction error.

Sharpening GenePT similarities by taking the fourth power partially closes the prediction gap, and the pooled-vocabulary control supports preserving source identity in the knowledge profile.

Table D.6: **Encoding geometry and contrast prediction.** Alignment: Pearson correlation, over pairs of training perturbations, between kernel similarity and response-contrast similarity (mean across four cell lines). Background: mean off-diagonal similarity. Kernel spread: median effective support of the raw similarity vector k_p, $(\sum_{p'} k)^2 / \sum_{p'} k^2$, as a share of the training support, not the ridge regression weights. Error and ρ: anchored ridge regression on the development folds at the best penalty on the grid; Δ is relative to the knowledge profile at its best penalty (4.789×10^{-3}). GenePT uses the model-3 text embeddings of the substitution experiment; the power is elementwise.

Representation	Alignment ↑	Background	Kernel spread (%)	Error Δ ↓	ρ ↑
Knowledge profile (RELAY)	**.172**	.024	37.9	—	**.430**
Knowledge profile, pooled vocabulary	.153	.027	37.3	+1.1%	.420
STRING only	.141	.032	35.4	+3.6%	.400
GO only	.132	.036	47.4	+5.1%	.388
GO Jaccard (GEARS similarity)	.078	.137	86.3	+4.5%	.387
GenePT	.053	.427	97.5	+10.4%	.324
GenePT, fourth power	.102	.041	69.2	+6.4%	.370

The knowledge kernel achieves the lowest development error among the tested representations; both response alignment and the distribution of similarities appear relevant to transfer quality.

Two comparisons examine how responses are transferred (Table D.7). The first keeps the knowledge kernel, folds, labels, and standardized-effect prediction fixed, replaces only the kernel ridge regressions for D and A with neighbor averages, and retrains every downstream component. Equal-weight averaging uses the top-k training perturbations with positive similarity, and weighted averaging weights them by $k(p, p')$. For each cell line, k is selected on pooled development predictions by the mean of DE and DIR Macro-F1, giving $k = 5, 5, 5, 10$ (equal) and $10, 20, 20, 20$ (weighted) for K562, HepG2, Jurkat, and RPE1. A query without a positive neighbor receives zero contrasts. The ridge reference uses the DE model from the matched retraining in Table D.8 and the analytic DIR anchor. Ridge regression's advantage over weighted neighbor averaging lies mainly in global accuracy, with similar point estimates for within-gene decision quality.

The second comparison is a public-resource analogue of graph-neighbor averaging (Fawkes et al., 2026). It takes the top 20 in-neighbors of each query and averages the measured responses of those in the same cell line's training set. It thresholds this average directly using development predictions and falls back to the mean training response when no neighbor is available. RELAY's full pipeline outperforms this direct public-graph transfer baseline.

Table D.7: **Response operators and neighbor transfer** (cell lines weighted equally; point estimates). The first block replaces RELAY's ridge regressions with neighbor averages over the same knowledge kernel and retrains all downstream components. The graph-neighbor row averages the responses of public STRING and GO neighbors and uses no learned readout.

	DE			DIR			Three-class
Response operator	Macro-F1	WG-BAcc	pgAUROC	Macro-F1	WG-BAcc	pgAUROC	Macro-F1
Kernel ridge regression (RELAY)	**.8953**	.6593	**.7729**	**.8722**	**.7672**	.8611	**.8131**
Weighted neighbor average	.8914	**.6599**	.7656	.8637	.7623	**.8667**	.8043
Equal-weight neighbor average	.8899	.6508	.7574	.8477	.7495	.8534	.7952
Public graph-neighbor average	.8215	.6264	.6622	.8500	.7500	.8392	.7439

D.3.3 Response Decomposition and Task Readouts

Following Section 5.3, the buildup (Table D.4b) and matched DE retraining (Table D.8) examine how response decomposition and readout inputs shape global and within-gene decisions. Matched removals are retrained under the same supervision and calibration protocol and compared with the same-batch full model.

Table D.8: **DE component ablation under matched retraining** (cell lines weighted equally). The first column gives the same-batch full model; the other columns give paired differences from it in percentage points. Brackets are 95% perturbation-cluster bootstrap intervals (2,000 replicates, seed 17, resampling shared across cell lines).

Metric	Full (same batch)	− PCA context	Joint simplification	Mean training effect
DE Macro-F1	.8953	-3.74 [$-4.11, -3.38$]	$+0.19$ [$-0.11, +0.49$]	-5.89 [$-6.40, -5.38$]
DE pgAUROC	.7729	-5.67 [$-6.47, -4.34$]	-2.39 [$-3.60, -1.34$]	-27.29 [$-29.24, -26.43$]
DE WG-BAcc	.6593	-2.96 [$-3.65, -2.21$]	-3.88 [$-4.68, -3.21$]	-15.94 [$-17.16, -14.85$]

The learned DE readout favors global classification in the buildup, with a cost to within-gene discrimination. The PCA control probes the information used by that readout: it retains the MLP and local response features while zeroing both $u(p)$ and $v_a(g)$. The DE performance differences have intervals below zero in every cell line, supporting the value of interpreting a local effect within perturbation-wide and gene-wide response patterns (Section 4.3).

Mean–contrast decomposition specifies the default response under weak transfer: ridge shrinkage sends $\widehat{d}$ toward zero, leaving μ as the prediction (Section 4.1). The mean-effect control removes the contrasts altogether and gives every perturbation of a gene the same prediction, eliminating within-gene discrimination by construction.

For DE, the magnitude view adds a complementary response target. Modeling $|r|$ directly preserves magnitude information that can be lost when opposite-signed effects are combined during transfer. The corresponding DE gains are consistent with this role and cost one additional ridge solve.

The standardized-effect increment then adds information about effect size relative to measurement uncertainty (Section 4.3). Its gain in within-gene discrimination comes with a global-accuracy cost, consistent with its support-weighted fitting objective (Appendix C.3). Selection under the one-point development Macro-F1 budget constrains this trade-off during calibration, and test-set DE Macro-F1 remains above the gene prior.

Jointly removing centering, the magnitude view, and the standardized-effect increment lowers DE within-gene discrimination while leaving Macro-F1 nearly unchanged.

DIR makes the distinction between rankings and decisions explicit (Table D.9). Within a gene, adding μ shifts every predicted response by the same amount, so all ranking information comes from $\widehat{d}$. A global decision threshold turns this shift into a gene-specific preference for one direction. The term ablation shows the resulting trade-off: retaining μ improves global Macro-F1, while the contrast

alone favors within-gene decisions, even with identical rankings. The per-gene scale τ puts genes on comparable score scales, and the monotone $\tanh$ map bounds the score; both preserve within-gene rankings.

Table D.9: **DIR anchor terms** (cell lines weighted equally). μ is constant within a gene, so its within-gene metrics equal 0.5 by construction. Dashes mark metrics not recorded for these variants.

Anchor terms	Macro-F1	pgBAcc	WG-BAcc	pgAUROC
μ only (gene mean)	.7645	.5000	.5000	.5000
$\widehat{d}$ only (query contrast)	.7650	—	**.7929**	.8611
$\widehat{d} + \mu$ without τ	.8706	—	.7680	.8611
$\tanh((\widehat{d} + \mu)/\tau)$ (anchor)	**.8722**	.7159	.7672	.8611

D.3.4 Trade-offs Between Global and Within-Gene Performance

Three components move global and within-gene metrics in opposite directions: the learned DE readout ($+8.57$ Macro-F1, -2.73 WG-BAcc), the gene mean in the DIR anchor ($+10.72$ Macro-F1, -2.57 WG-BAcc), and the standardized-effect increment (-1.04 Macro-F1, $+3.46$ WG-BAcc). In these comparisons, the learned DE readout and the gene mean in the DIR anchor improve Macro-F1 while reducing within-gene decision accuracy; the standardized-effect increment moves the DE metrics in the opposite directions. The final configuration exceeds the gene prior in equal-cell mean Macro-F1 on both tasks. The increment is selected subject to a one-point development Macro-F1 loss budget.

Departures from the gene prior. Of the gene prior's DE decisions, RELAY changes 5.09%, correcting 2.730% that the prior gets wrong and overturning 2.357% that it gets right, a net accuracy gain of 0.373 points. When each query is weighted so that the corrected minus the overturned share equals the WG-BAcc difference, corrections outweigh overturned decisions by 22.60% to 6.86%, the 15.74-point gain of Table D.3. The net gain is positive in every cell line, ranging from 12.62 to 19.37 points, and every 95% perturbation-bootstrap interval lies above zero.

D.4 Held-Out Cell Line and Unseen Compounds

D.4.1 Held-Out RPE1

RPE1 contributes only unperturbed controls. A source-fitted 32D SVD of batch-averaged controls gives cell-line coordinates b_c; source weights satisfy $\omega_c \propto 1 + \cos(b_{\mathrm{RPE1}}, b_c)$ and sum to one. Source means, contrasts, and standardized-effect predictions are combined in response units before normalization and PCA, with weights renormalized over the sources covering each gene. Genes absent from all sources receive zero contrast, mean, and anchor; no query is dropped. The source gene union contains 11,296 genes, and response-coordinate and standardized-effect coverage on RPE1 are both 93.41%.

To fit the response transforms and readouts, each source cell line serves in turn as a pseudo-held-out cell line and receives predictions from the other sources. Perturbation folds follow Appendix C.2, excluding each held-out perturbation from every contributing source regression. All source ridge penalties are one; increment strength and thresholds are calibrated on pooled source-development OOF predictions with the rules of Table C.2. Final source regressions use all training groups, and DIR uses the analytic anchor on the combined response. Target perturbation responses and labels are used only for evaluation.

For evaluation, a query is classified as overlapping if its ordered perturbation–gene pair occurs in at least one source cell line's official DE training split. The non-overlap subset contains 36,431 queries, including 4,286 DIR queries (Table D.10).

The source gene prior uses pooled source training labels for each gene, with the majority label of the pooled source training rate as fallback. The exact-pair prior uses the matching source pair's majority label when available and falls back to the source gene prior otherwise.

Table D.10: **Binary tasks after transfer to held-out RPE1.** RELAY's DE and DIR results under the main metric definitions; non-overlap excludes the 1,511 of 37,942 queries whose ordered pair appears in training data from source cell lines. Table 4 gives the three-class results, with AROMA values as published (Wang et al., 2026, Table 2).

Task / population	Macro-F1	Accuracy	pgBAcc	WG-BAcc	pgAUROC
DE / all	.8459	.9254	.6018	.5933	.7091
DIR / all	.8507	.8557	.7114	.7333	.8125
DE / non-overlap	.8191	.9251	.5908	.5807	.6852
DIR / non-overlap	.8264	.8325	.6702	.6576	.7686

On the 1,511 overlapping pairs, the exact-pair prior reaches Macro-F1 0.9402 versus 0.9006 for RELAY; on the remaining pairs it reduces to the source gene prior. On the three-class task of Table 4, its Macro-F1 is .7246 overall and .6552 on non-overlapping queries. RELAY reaches .7176 on the non-overlapping queries.

D.4.2 Unseen Compounds

Using the GeneTAK data and labels described in Appendix A.1, we randomly permute the sorted drug identities with seed 17 and assign 284 drugs to training and the remaining 95 to a held-out set shared by all cell lines (Table D.11). This follows PerturbQA's 75/25 perturbation-disjoint ratio. The training drugs are permuted again with the same seed and assigned to four development groups of 71 drugs each, following Appendix C.2.

Table D.11: **GeneTAK populations under the shared 284/95 drug split.** Mixed-label genes have both classes among a cell line's held-out queries and define the within-gene metrics.

	DE pairs			DIR pairs		
Cell line	Train	Held-out	Mixed genes	Train	Held-out	Mixed genes
C32	102,446	36,680	850	54,487	20,627	478
HepG2/C3A	86,005	29,954	608	37,682	13,813	379
HOP62	107,824	38,122	915	59,711	22,178	507
Hs 766T	87,272	30,815	706	38,872	14,724	473
PANC-1	86,025	30,631	934	37,882	14,490	439
Total	469,572	166,202	4,013	228,634	85,832	2,276

For each training drug, the response averages the released log-normalized expression over its cells and doses and subtracts plate-matched DMSO vehicle controls. A compound's targets are the union of those listed in the Tahoe-100M release (Zhang et al., 2025) and PrimeKG drug–protein edges (Chandak et al., 2023). Pooling the knowledge tokens of these targets (Section 4.2) gives one knowledge profile per compound. Thirteen of the 95 held-out drugs and 31 of the 284 training drugs retain no token; their predicted contrasts are zero, so they receive the gene means.

For these compound responses, RELAY uses ridge penalty 1, the 66-dimensional DE readout, and the analytic DIR anchor, without the standardized-effect increment (Section 4.4). Decision thresholds maximize Macro-F1 on pooled development predictions for the 213 drugs in the first three groups, and the 71 drugs of the fourth group form the held-out development split.

We use the authors' public repository to evaluate VCWorld with Gemma4-26B-A4B-it on the same 95 held-out drugs in all five cell lines. The release provides no query-level predictions for per-gene metrics on this split. Both DE and DIR use the released prompt templates, drug-similarity and gene-neighbor files, and drug and gene descriptions. We use the released retrieval procedure and prompts, replacing randomly sampled example outcomes with their true training labels. Unanswered queries receive the pooled training-rate fallback used for SUMMER and AROMA (Appendix D.1). Each variant is a single pass without repeats.

The gene prior uses the same training drugs and the rule in Appendix D.1. The 1,078 DIR queries whose gene has no DIR training label receive the majority label of the pooled training rate $\bar{\pi}$.

For STATE, we retrain from scratch for each cell line on the 284 training drugs with the released Tahoe ST-HVG configuration. Morgan fingerprints provide representations for both training and held-out drugs. Doses are pooled as in the labels, the checkpoint is selected by validation loss, and thresholds maximize Macro-F1 on development-drug predictions. The released Tahoe checkpoint is excluded because Tahoe-100M contains GeneTAK and every held-out drug appears in its drug vocabulary. GEARS represents perturbations as genes in a gene-interaction graph and has no compound input, so it is not evaluated on GeneTAK.

For the compound comparison in Table D.12, intervals use 2,000 paired bootstrap replicates (seed 17) that resample held-out drugs with weights shared across the five cell lines, recompute eligibility and effective support in each replicate, and average cell lines equally.

For unweighted pgBAcc, RELAY improves on the gene prior by 5.71 DE points (95% CI, 1.33–9.07) and 6.37 DIR points (95% CI, 2.74–9.66).

Table D.12: **GeneTAK results by cell line on the 95 held-out drugs.** VCWorld is evaluated using the authors' public repository (Appendix D.4.2). Mean rows weight cell lines equally.

		DE				DIR			
Cell line	Method	Macro-F1	WG-BAcc	pgBAcc	pgAUROC	Macro-F1	WG-BAcc	pgBAcc	pgAUROC
C32	Gene prior	.9240	.5000	.5000	.5000	.8062	.5000	.5000	.5000
	STATE	.8295	.5027	.5065	.5160	.6443	.5340	.5341	.5530
	VCWorld	.7094	.5236	.5253	.5253	.6008	.5136	.5154	.5157
	RELAY	.9239	.5481	.5557	.6210	.7990	.5640	.5667	.6401
HepG2/C3A	Gene prior	.9386	.5000	.5000	.5000	.7855	.5000	.5000	.5000
	STATE	.8556	.5115	.5187	.5208	.6605	.5265	.5066	.5375
	VCWorld	.7135	.5324	.5379	.5379	.6068	.5324	.5239	.5235
	RELAY	.9365	.5395	.5568	.6005	.7586	.5445	.5497	.6062
HOP62	Gene prior	.9104	.5000	.5000	.5000	.7907	.5000	.5000	.5000
	STATE	.8270	.5079	.5047	.5137	.6755	.5391	.5421	.5671
	VCWorld	.7015	.5260	.5170	.5170	.5901	.5075	.5045	.5044
	RELAY	.9048	.5376	.5390	.5848	.7979	.5654	.5754	.6399
Hs 766T	Gene prior	.9192	.5000	.5000	.5000	.7597	.5000	.5000	.5000
	STATE	.8228	.5229	.5263	.5558	.6228	.5422	.5415	.5765
	VCWorld	.6874	.5340	.5426	.5426	.5879	.5261	.5257	.5255
	RELAY	.9203	.5452	.5596	.6116	.7629	.5664	.5845	.6604
PANC-1	Gene prior	.9058	.5000	.5000	.5000	.8051	.5000	.5000	.5000
	STATE	.8142	.5113	.5070	.5207	.6346	.5267	.5179	.5466
	VCWorld	.6824	.5250	.5258	.5258	.5891	.5100	.5012	.5009
	RELAY	.9038	.5613	.5745	.6513	.7955	.5469	.5419	.5954
Mean	Gene prior	.9196	.5000	.5000	.5000	.7894	.5000	.5000	.5000
	STATE	.8298	.5113	.5127	.5254	.6475	.5337	.5284	.5561
	VCWorld	.6988	.5282	.5297	.5297	.5949	.5179	.5142	.5140
	RELAY	.9179	.5463	.5571	.6138	.7828	.5574	.5637	.6284