PIE: GENERALIZING PERTURBATION EFFECTS ACROSS UNSEEN PERTURBATIONS, CONTEXTS AND DATASETS

Anonymous authors
Paper under double-blind review

ABSTRACT

Predicting cellular response to perturbations is key to understanding biological mechanisms and selecting therapeutic targets. However, generalizing across cellular contexts, perturbations, and experimental datasets remains challenging due to the difficulty of measuring true perturbation effects and incomplete representations of the system being perturbed. We introduce PIE (Perturbation is Everything), which addresses these challenges by reformulating the learning task around population-level perturbation effects and incorporating auxiliary inputs that better characterize the biological system. PIE directly predicts differentially expressed genes and changes in gene expression for each (context, perturbation) pair, using external biological knowledge, baseline gene expression, and observed perturbation responses. On the Replogle-Nadig dataset, PIE achieves state-of-the-art performance on most metrics across all tested generalization settings, including the most challenging setting where both the context and perturbation are unseen during training. PIE achieves 1.2–3.2 times the AUPRC of the strongest baseline in each setting for predicting differentially expressed genes: across unseen contexts, unseen perturbations, and jointly unseen contexts and perturbations. PIE also outperforms existing baselines on zero-shot transfer across experimental datasets. Overall, PIE provides a framework for predicting cellular responses across datasets, contexts, and perturbations.

1 INTRODUCTION

Understanding how cells respond to genetic and chemical perturbations is important for uncovering the mechanisms that drive biological systems and disease. Every perturbation experiment is defined by the cellular context in which the perturbation is performed, the applied perturbation itself, and the broader experimental context or dataset. Computational models can help guide experiments by predicting responses beyond the conditions already measured. However, generalizing across variations in these experimental conditions requires distinguishing the biological effects of the perturbation from dataset-specific factors. This remains challenging due to the difficulty of measuring true perturbation effects and incomplete representations of the system being perturbed. Most existing perturbation prediction models rely mainly on transcriptomic data or a single auxiliary data source to distinguish between contexts and perturbations. Moreover, they attempt to predict perturbation effects at the level of individual cells, which is challenging because current single-cell perturbation technologies cannot directly measure perturbation effects at this level. This is also reflected in how most evaluation metrics focus on pseudobulk measurements (Adduri et al., 2026b), such as the set of differentially expressed genes.

We introduce a new model, PIE (Perturbation is Everything), that addresses these challenges by simplifying the learning task and incorporating auxiliary inputs that better characterize the biological system. Unlike existing perturbation prediction models that attempt to predict the absolute gene expression state of a cell following perturbation, PIE focuses on learning population-level effects for each condition, or (context, perturbation) pair. This enhances the quality of the training signal and improves generalization to new conditions. To further limit dataset-specific noise, we focus on predicting coarse outcomes such as the set of differentially expressed genes, the direction of differential expression, and the binned magnitude of mean expression change. Together, these choices greatly reduce the space of possible outcomes that the model must predict compared to conventional

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.

transcriptomic models, thereby enhancing transferability to new conditions. Beyond enhancing the perturbation signal, PIE leverages three kinds of auxiliary inputs to build more complete representations of the system being perturbed: knowledge sources, which embed each context and perturbation using external databases such as Cellosaurus, NCBI Gene, and PubChem; the mean control expression of the query context; and summary statistics of responses in other observed biological conditions.

Unlike existing baselines, PIE can generalize across multiple axes of experimental covariates, such as dataset, context, and perturbation, including combinations of these axes. Within each axis, PIE significantly outperforms existing baselines that are specialized for that axis on several metrics. We see the largest increase on AUPRC in classifying differentially expressed genes, where PIE outperforms several evaluated baselines by more than 3-fold on both within and cross dataset generalization tasks, across multiple experimental settings. PIE's modular architecture allows ablation studies to test the relationship between input sources, generalization settings, and metrics.

2 Related Works

Predicting transcriptional responses to interventions encompasses several distinct generalization problems. Methods for context generalization transfer the effect of a known perturbation to a new background, either by transporting its unperturbed cells (Bunne et al., 2023; Klein et al., 2025; Adduri et al., 2026b; Dong et al., 2026) or by learning how population-level responses vary across contexts (Svensson et al., 2026; Palla et al., 2026). Methods for perturbation generalization instead replace learned perturbation identities with representations derived from prior knowledge, such as molecular structure for compounds (Hetzel et al., 2022) and gene interaction graphs, text descriptions, or pretrained gene embeddings for genetic perturbations (Roohani et al., 2024; Chen & Zou, 2025; Littman et al., 2025; Wenkel et al., 2026; Liang & Singh, 2026). Transfer across datasets remains difficult even for models trained on hundreds of millions of cells (Adduri et al., 2026b).

Methods also differ in the resolution at which they model perturbation effects. Single-cell foundation models such as scGPT, Geneformer, scFoundation, and UCE are pretrained on large observational atlases to learn general-purpose representations of transcriptomic state, and are fine-tuned or prompted to predict perturbed expression profiles (Cui et al., 2024; Theodoris et al., 2023; Hao et al., 2024; Rosen et al., 2023; Dong et al., 2026). Generative and transport-based models learn the distribution of perturbed cells directly (Lotfollahi et al., 2023; Bunne et al., 2023; Klein et al., 2025; Palla et al., 2025). Because the unperturbed state of a perturbed cell is never observed, these models must separate the perturbation effect from heterogeneity within the cell population and from technical variation across experiments, which motivates set-based architectures (Adduri et al., 2026b) and disentangled or sparse latent representations (Bereket & Karaletsos, 2023; Piran et al., 2024).

A smaller set of methods treat the label (context, perturbation) as the sample and predict pseudobulk responses (Littman et al., 2025; Liang & Singh, 2026). Rhaister-O goes further by operating directly on the metrics that biologists use to interpret screens, including log fold changes, p-values, and expression deltas, and matches or exceeds cell-level models on these readouts (Svensson et al., 2026). PIE is likewise trained on differential-expression calls, log fold changes, and deltas in expression at pseudobulk resolution. Unlike Rhaister-O, it represents contexts and perturbations through external knowledge rather than through their observed responses alone, which allows it to predict responses to contexts and perturbations absent from training.

3 Method

3.1 Problem setup

A context c is the cellular background in which a perturbation is applied. In our datasets, a context is a cell line, either unstimulated, or treated with a signaling molecule such as a cytokine. A perturbation p is a CRISPRi knockdown of a target gene or a small molecule at a given dose. An experimental condition is a pair (c, p) measured in a dataset, and we write $\mathcal{C}_{cp}$ for its set of profiled cells. Each dataset also includes unperturbed control cells $\mathcal{C}_{c,\texttt{ctrl}}$ for every c. Different datasets may measure different sets of genes, and we use G to denote the size of the measured gene vocabulary.

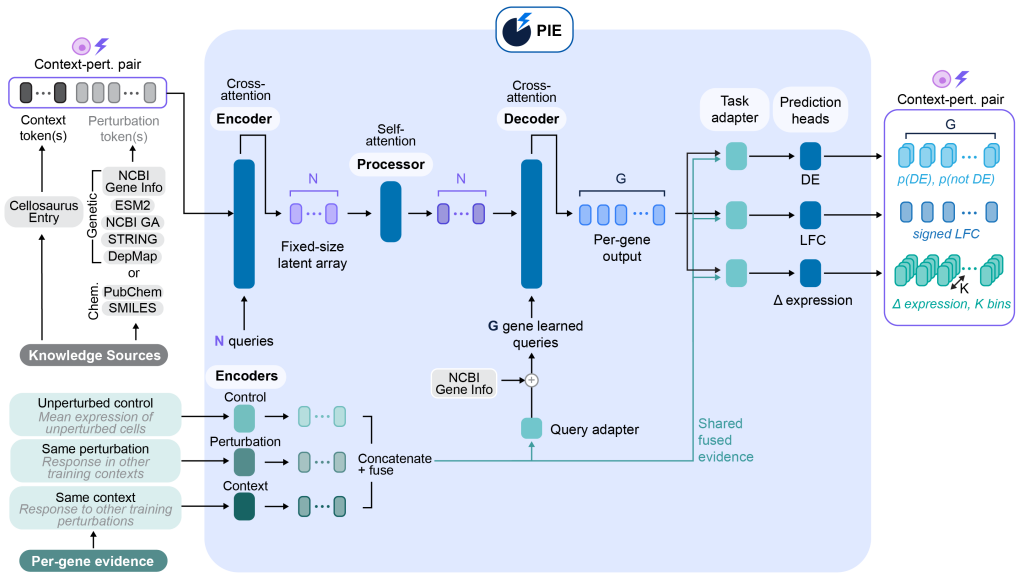

Figure 1: **PIE overview.** A sample is a (context, perturbation) pair. Knowledge Sources describe contexts and perturbations with variable-size sets of embeddings from external databases. A source with no entry for a context or perturbation contributes no tokens. Learned latents query these tokens (Encoder) and subsequently are processed by a bidirectional transformer (Processor). Each measured gene is represented by a text embedding derived from its NCBI Gene record, and is further augmented with experimental evidence, such as the control-cell mean expression of the context, summary statistics of responses to the same perturbation in other contexts, and summary statistics of responses to other perturbations in the same context. These response statistics are derived exclusively from the training set. In the Decoder, each gene queries the latents, producing a per-gene feature vector. Three heads, each preceded by a task adapter, predict a binary DE call, a signed log fold change, and a binned pseudobulk delta expression, for every gene.

For every experimental condition and measured gene g, we derive three labels by comparing $\mathcal{C}_{cp}$ with $\mathcal{C}_{c,\texttt{ctrl}}$ (Appendix A.1). Let n_{ig} be the normalized expression of gene g in cell i, $x_{ig} = \log(1+n_{ig})$ its log-normalized expression, and $\bar{n}_g(\mathcal{C})$ and $\bar{x}_g(\mathcal{C})$ their means over a set of cells $\mathcal{C}$. The DE label $y_{cpg} \in \{0, 1\}$ indicates whether a Wilcoxon rank-sum test yields a Benjamini–Hochberg FDR below 0.05. The LFC and the pseudobulk delta expression are:

$$\ell_{cpg} = \log_2 \frac{\bar{n}_g(\mathcal{C}_{cp})}{\bar{n}_g(\mathcal{C}_{c,\texttt{ctrl}})}, \qquad \Delta_{cpg} = \bar{x}_g(\mathcal{C}_{cp}) - \bar{x}_g(\mathcal{C}_{c,\texttt{ctrl}}). \tag{1}$$

The LFC is defined only for DE genes, and its sign gives the direction of change. All three labels are derived from cell populations rather than individual cells.

During training, the labels are observed for a training set $\mathcal{T}$ of experimental conditions. At test time, the model must predict y_{cpg}, ℓ_{cpg}, and Δ_{cpg} for every measured gene of a condition $(c, p) \notin \mathcal{T}$, using only information that is available without measuring that condition, such as external knowledge about c and p, the control cells of $\mathcal{C}_{c,\texttt{ctrl}}$, and the responses in $\mathcal{T}$. The generalization regime depends on what $\mathcal{T}$ contains. A context is unseen if no condition in $\mathcal{T}$ involves c, and a perturbation is unseen if no condition in $\mathcal{T}$ involves p. The two may also be unseen jointly. In all regimes, model parameters stay fixed at test time.

3.2 Knowledge Sources

To make predictions for contexts and perturbations that have never been measured, PIE represents each context and perturbation using information from external databases, referred to as Knowledge Sources. A Knowledge Source maps a context or perturbation to a single pooled embedding or a

sequence of embeddings (Appendix A.2). We project each embedding to the model width, add a learned embedding that identifies the type of knowledge source, and concatenate the embeddings from all sources of an experimental condition into a single sequence of input embeddings. When a database has no entry for a context or perturbation, that source is omitted from the sequence (Appendix A.2.2).

Contexts are represented by a single source, the Context Description, which is the cell line's Cellosaurus entry (Bairoch, 2018), embedded with a text embedding model. For the stimulated contexts of Jiang et al., we append a description of the stimulus to the entry.

Perturbations are represented by a richer set of Perturbation Knowledge Sources. Every perturbation has a Perturbation Description, embedded in the same way as the Context Description: the NCBI Gene entry of the target gene for genetic perturbations (Brown et al., 2015), and the PubChem entry of the compound, together with its dose, for chemical perturbations (Kim et al., 2025). For genetic perturbations, we add four sources that describe the target gene: NCBI Gene Annotations, which list its Gene Ontology terms (Ashburner et al., 2000; The Gene Ontology Consortium et al., 2025), disease associations (McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University, 2026), and gene-group memberships; ESM2 embeddings of its protein sequence (Lin et al., 2023); STRING embeddings of its protein interaction network (Szklarczyk et al., 2025; Hu et al., 2025); and DepMap profiles of its knockout effect on viability across cell lines (DepMap, Broad, 2025; Dempster et al., 2021). For chemical perturbations, we add ChemBERTa embeddings of the compound's SMILES string (Ahmad et al., 2022) and three measured activity profiles: transcriptional responses from L1000 (Subramanian et al., 2017), viability effects from PRISM (Corsello et al., 2020), and morphological effects from JUMP (Chandrasekaran et al., 2023). Coverage varies across sources (Appendix Table 1), particularly for the measured compound profiles.

3.3 Per-gene evidence

The Knowledge Sources describe a context and a perturbation, but not how individual genes behave in them. To give PIE direct access to measured expression, we supplement each gene with per-gene evidence drawn from the data, organized in three channels. Control Mean Evidence is the gene's mean expression in the control cells of the context, $\frac{1}{|\mathcal{C}_{c,\mathtt{ctrl}}|}\sum_{i\in\mathcal{C}_{c,\mathtt{ctrl}}} x_{ig}$, and is available for every context. Same-pert Evidence summarizes the gene's responses to perturbation p in other contexts, that is, over conditions $(c', p) \in \mathcal{T}$ with $c' \neq c$. Same-context Evidence summarizes its responses to other perturbations in context c, over conditions $(c, p') \in \mathcal{T}$ with $p' \neq p$. Each response channel records the mean and standard deviation of Δ and ℓ across these conditions, together with the fraction of conditions in which the gene is up- or down-regulated (Appendix A.3). When no training condition contributes to a channel, as for Same-context Evidence in an unseen context, we set the channel to zero. Each statistic is passed through its own small encoder, and a fusion MLP combines the three channels into a single evidence vector $\mathbf{e}_g$ per gene.

3.4 Encode–process–decode trunk

PIE uses the encode–process–decode design of Perceiver IO (Jaegle et al., 2022), which decouples the size of the input and the output from the size of the model. Learned latent vectors first cross-attend to the Knowledge Source tokens (encode), and the latents then self-attend to one another (process). Finally, gene embeddings query the latents (decode), producing a per-gene feature vector $\mathbf{h}_g$ (Appendix A.4). A gene's query combines what is known about the gene with what has been measured for it: we obtain a single embedding vector of the gene's NCBI Gene entry, project it to the model width, and add its evidence vector $\mathbf{e}_g$. Gene queries do not attend to one another, so each gene is decoded independently. Because the queries are computed from text rather than learned per gene, PIE has no fixed gene vocabulary, and each dataset is decoded on its own set of measured genes.

3.5 Training formulation

Three linear heads map $\mathbf{h}_g$ to the three labels of Section 3.1, and each head is trained with its own loss. Before each head, $\mathbf{h}_g$ is refined with the gene's evidence vector $\mathbf{e}_g$: an MLP takes the two as input and adds its output back to $\mathbf{h}_g$ (Appendix Figure 5).

DE head. The DE head outputs the probability $\hat{\pi}_{cpg}$ that gene g is differentially expressed, which we use to rank genes and threshold at 0.5 to call DE. It is trained with a weighted binary cross-entropy $\mathcal{L}_{\mathrm{DE}}$ over the genes $\mathcal{G}_{cp}$ that pass the CPM filter, where the weights are capped inverse-frequency class weights computed per condition.

LFC head. The LFC head outputs a real value $\hat{\ell}_{cpg}$, whose sign gives the direction of change. Because the LFC is defined only for the DE genes $\mathcal{D}_{cp} \subseteq \mathcal{G}_{cp}$, we supervise it there, with a Huber loss H_κ (Huber, 1964) on its magnitude and a hinge loss (Cortes & Vapnik, 1995) on its direction:

$$\begin{aligned} \mathcal{L}_{\mathrm{LFC}} &= \frac{1}{\sum_{(c,p)\in\mathcal{B}} |\mathcal{D}_{cp}|} \sum_{(c,p)\in\mathcal{B}} \sum_{g\in\mathcal{D}_{cp}} \lambda_g \, H_\kappa\big(\hat{\ell}_{cpg}, \ell_{cpg}\big), \\ \mathcal{L}_{\mathrm{dir}} &= \frac{1}{\sum_{(c,p)\in\mathcal{B}} |\mathcal{D}_{cp}|} \sum_{(c,p)\in\mathcal{B}} \sum_{g\in\mathcal{D}_{cp}} \max\big(0, -\operatorname{sign}(\ell_{cpg})\, \hat{\ell}_{cpg}/\tau\big), \end{aligned} \tag{2}$$

where λ_g down-weights the target gene of a genetic perturbation, $\mathcal{B}$ is the set of experimental conditions in a minibatch, and τ is a fixed scale.

Delta-expression head. Rather than regressing Δ_{cpg} directly, this head outputs a distribution $\hat{\mathbf{q}}_{cpg}$ over K equal-width bins that tile $[-\Delta_{\max}, \Delta_{\max}]$, with centres $b_1 < \cdots < b_K$, where $\Delta_{\max}$ is the largest per-dataset 99.9th percentile of $|\Delta_{cpg}|$ over the training set (Appendix A.5). We clip Δ_{cpg} to $[b_1, b_K]$ and encode it as a two-hot target $\mathbf{t}_{cpg}$ over the neighbouring centres $b_k \le \Delta_{cpg} < b_{k+1}$, with weight $(b_{k+1} - \Delta_{cpg})/(b_{k+1} - b_k)$ on bin k, weight $(\Delta_{cpg} - b_k)/(b_{k+1} - b_k)$ on bin $k+1$, and zero elsewhere. The head is trained with a weighted cross-entropy over all G measured genes,

$$\mathcal{L}_{\Delta} = -\frac{1}{|\mathcal{B}|\, G} \sum_{(c,p)\in\mathcal{B}} \sum_{g=1}^{G} \sum_{k=1}^{K} \nu_k \, t_{cpgk} \log \hat{q}_{cpgk}, \tag{3}$$

where ν_k are capped inverse-frequency bin weights computed per condition. At inference, the predicted delta expression is the centre of the most probable bin, and adding it to the control mean of the context gives the predicted pseudobulk expression of the perturbed cells.

The final loss of the model is a composite loss, the unweighted sum

$$\mathcal{L} = \mathcal{L}_{\mathrm{DE}} + \mathcal{L}_{\mathrm{LFC}} + \mathcal{L}_{\mathrm{dir}} + \mathcal{L}_{\Delta}, \tag{4}$$

where each term is averaged over the (condition, gene) pairs it supervises in a batch. During training, we randomly set Same-pert and Same-context Evidence to zero for a subset of samples, so that the model learns to predict when either channel is unavailable, as it is for unseen perturbations and contexts.

3.6 Generalization to unseen contexts, perturbations, and datasets

PIE represents an unseen context or perturbation in the same way as a seen one, through its Knowledge Sources and the evidence available for it (Appendix A.3). For an unseen context, the training set has no responses in that context, so Same-context Evidence is empty. Likewise, for an unseen perturbation, Same-pert Evidence is empty. For a held-out dataset, Same-context Evidence is empty. Its Same-pert Evidence is present only for perturbations that a training dataset also measured. In every setting, the trained model predicts in a single forward pass, without fine-tuning.

4 Experiments

4.1 Experimental Setup

Datasets. Our main benchmark is the Replogle-Nadig CRISPRi Perturb-seq dataset (Replogle et al., 2022; Nadig et al., 2025), comprising single-gene knockdowns in four cell lines (K562, RPE1, HepG2, and Jurkat). For cross-dataset transfer, we additionally use: (i) Tahoe-100M, a small molecule perturbation dataset (Zhang et al., 2025); (ii) the cytokine-stimulation CRISPRi screen of Jiang et al. (2025); (iii) the X-Atlas/Orion CRISPRi screens in HCT116 and HEK293T (Huang et al., 2025); (iv) and the H1 CRISPRi data of the 2025 Virtual Cell Challenge (Roohani et al., 2025). Appendix A.1 describes dataset preprocessing and reference-label construction.

Baselines. We compare PIE to a variety of established methods for context and perturbation generalization. For unseen contexts, we evaluate STATE, which predicts perturbed cells from control cells (Adduri et al., 2026b), and Rhaister-O, which learns perturbation-specific DE statistics (Svensson et al., 2026). For unseen genetic perturbations in seen contexts, we evaluate COMPASS-N (Liang & Singh, 2026), which transfers responses from nearby perturbations in protein-embedding space, and PRESAGE (Littman et al., 2025), which uses gene knowledge graphs. We also evaluate TabICL (Palla et al., 2026) for both unseen-context and unseen-perturbation prediction, using its task-specific procedures. In our baselines, we further include perturbation mean, which averages a perturbation's response across training contexts, and context mean, which averages training perturbation responses within a context (Adduri et al., 2026b). For an unseen context, context mean predicts the mean expression of its control cells, and is the only applicable baseline when both the context and perturbation are unseen. Appendix B.2 reports baseline implementations.

Metrics. We evaluate all models using six metrics drawn from CELL-EVAL and CELL-EVAL2 (Adduri et al., 2026b; Dobin, 2026a): AUPRC, Jaccard score, DE direction match, Spearman LFC, Spearman effect size, and Discrimination Score L1. Together, these evaluate DE prediction, the direction and magnitude of expression changes, the number of DE genes, and discrimination between perturbation responses. Higher values are better for all six metrics. We adapt these metrics computations to accept predicted DE probabilities, LFCs, and pseudobulk delta expression directly. All methods are scored on the same test conditions, reference labels, and genes. Appendix B.1 provides the metric definitions, adaptations, and baseline-output conversions.

Evaluation Protocol. For held-out contexts, unperturbed control cells are available, but their perturbation responses are excluded from training. PIE's response-evidence inputs are computed only from training responses, and its parameters are not updated at test time. We select the PIE checkpoint with the highest validation AUPRC; test responses are used only to evaluate predictions. We report mean performance and sample standard deviation across the four folds for within-dataset experiments and across the held-out dataset's contexts for cross-dataset experiments.

4.2 GENERALIZATION WITHIN DATASET

We evaluate three regimes: unseen contexts with seen perturbations, unseen perturbations in seen contexts, and jointly unseen contexts and perturbations. On Replogle–Nadig, each of four folds holds out one test cell line and a subset of perturbations in the remaining three cell lines. Training uses the remaining perturbations in the three non-heldout cell lines. The three test sets isolate context generalization, perturbation generalization, and their combination. One PIE model is trained per fold, and the same selected checkpoint is evaluated in all three regimes. Appendix B.3 gives the full split and training protocols.

PIE achieves the highest score on all metrics for unseen perturbations and on five of six metrics for unseen contexts (Figure 2). For unseen contexts, its AUPRC is 0.424, compared with 0.344 for Rhaister-O, the strongest baseline, and its Spearman LFC is 0.664 versus 0.591. The exception is Spearman effect size, where Rhaister-O scores 0.741 and PIE scores 0.734. For unseen perturbations, PIE has an AUPRC of 0.304 versus 0.094 for TabICL, a greater than threefold improvement. When both entities are unseen, PIE exceeds context mean, the only applicable model, on every metric, reaching an AUPRC of 0.239 versus 0.088 and a Discrimination Score L1 of 0.643 versus 0.502 (effectively random). For context mean, Spearman effect size is undefined because it predicts no change for every perturbation, and Spearman LFC is undefined for an unseen context because it predicts no DE genes. PIE significantly outperforms the strongest baseline on several metrics including AUPRC in similar experiments conducted on Jiang et al., and Tahoe-100M. Appendix B.3 includes these experiment results.

4.3 GENERALIZATION ACROSS DATASETS

We next test whether PIE can transfer responses to a dataset absent from training. We train on Tahoe-100M, Jiang et al., X-Atlas/Orion, and the Virtual Cell Challenge, and evaluate on Replogle–Nadig. We exclude from training all cell lines shared with the held-out dataset, including derived cell lines, so this experiment requires generalization to both an unseen dataset and unseen contexts.

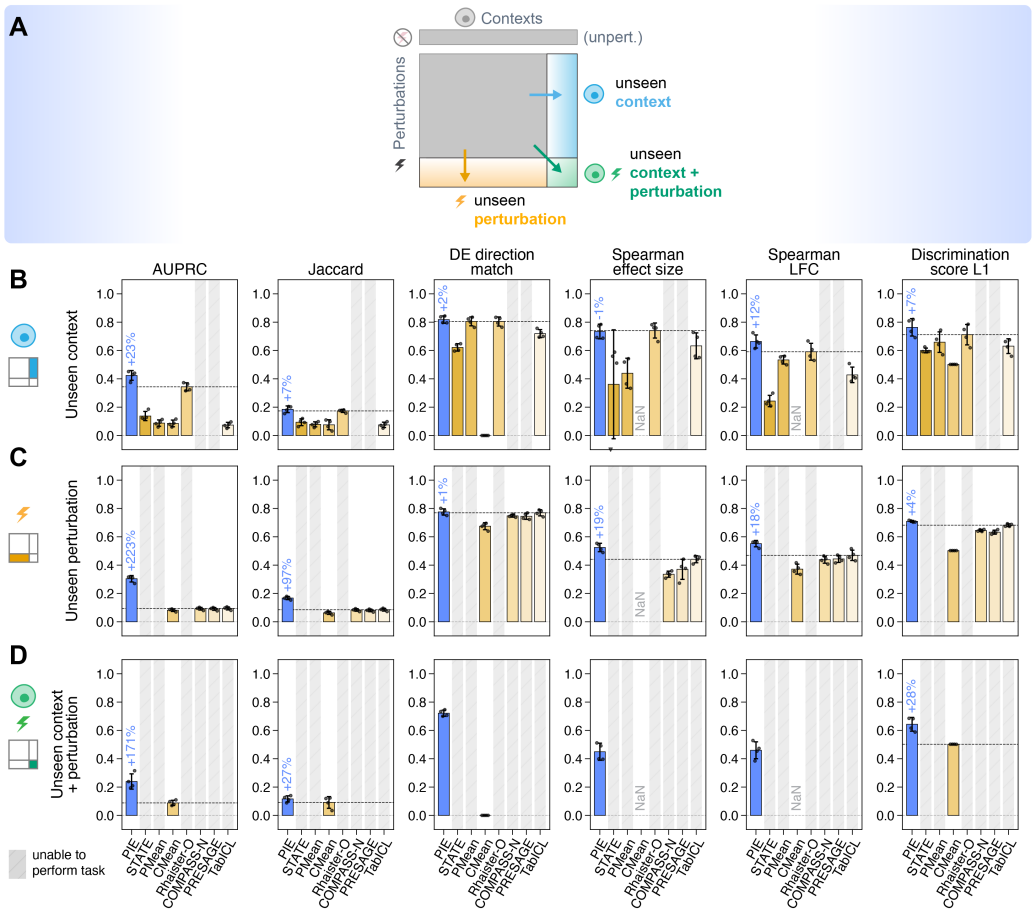

Figure 2: **Generalization within the Replogle–Nadig dataset.** (A) The three regimes. (B–D) Scores for unseen contexts, unseen perturbations, and both unseen. Bars show means and std across four folds, each holding out one cell line; hatched slots mark models that cannot perform the task, and the dashed line marks the strongest baseline. Higher is better for all metrics. PIE has the highest AUPRC in all three regimes, exceeding the strongest baseline by 0.08 for unseen contexts, 0.21 for unseen perturbations, and 0.15 for both unseen.

We separately evaluate perturbations measured in at least one training dataset and perturbations absent from all training datasets. The latter setting requires simultaneous generalization to an unseen dataset, context, and perturbation. A single checkpoint, selected using validation responses from the training datasets, is evaluated in both regimes. Results are averaged across the four held-out cell lines. Appendix B.4 provides the detailed protocol and additional held-out-dataset evaluations.

For seen perturbations, PIE achieves the highest mean score on five of six metrics (Figure 3). Its AUPRC is 0.378 versus 0.283 for Rhaister-O, the strongest baseline, and its Spearman effect size is 0.627 versus 0.547. The exception is DE direction match, where perturbation mean scores 0.733 and PIE scores 0.727.

For unseen perturbations, PIE exceeds context mean on every defined metric, with an AUPRC of 0.252 versus 0.102 and a Jaccard score of 0.123 versus 0.077. As in the within-dataset experiment, both Spearman correlations are undefined for context mean. Performance declines across all metrics when perturbations are also unseen, but PIE retains predictive signal despite the simultaneous shift in dataset, context, and perturbation. PIE also achieves the highest AUPRC in both regimes when Jiang et al., X-Atlas/Orion or the Virtual Cell Challenge 2025 is held out (Appendix B.4).

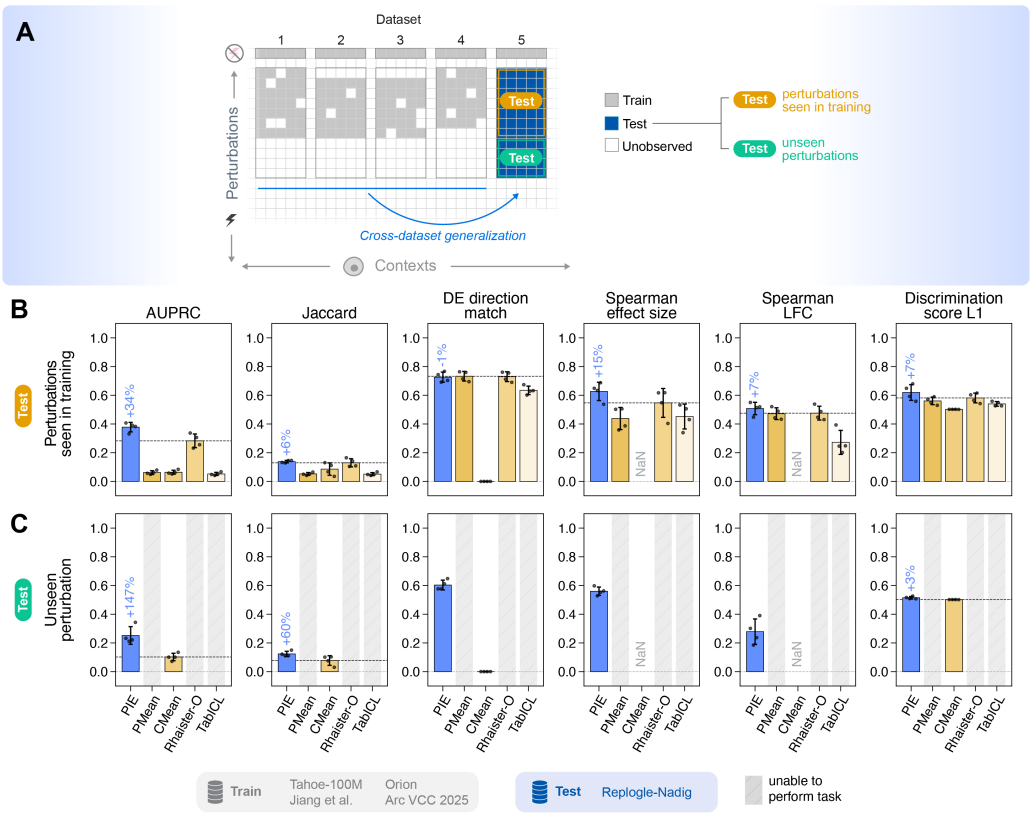

Figure 3: **Generalization to an unseen dataset, Replogle–Nadig.** Models trained on four other datasets predict responses in its four cell lines, for perturbations seen in training (top) and unseen in training (bottom); the latter requires zero-shot generalization to an unseen dataset, context, and perturbation. Bars show means and std across the four cell lines; hatched slots mark models that cannot perform the task, and the dashed line marks the strongest baseline. Higher is better for all metrics. PIE has the highest AUPRC in both regimes, exceeding the strongest baseline by 0.10 for seen perturbations and 0.15 for unseen perturbations.

4.4 Contributions of input signals

We examine the contributions of five input groups: Context Description, Perturbation Knowledge Sources, Control Mean Evidence, Same-context Evidence, and Same-pert Evidence. Using the four Replogle–Nadig folds and three within-dataset regimes, we train separate models for every subset of these groups and compute exact Shapley values: each group's marginal contribution averaged over subsets of the other inputs. Removed inputs are absent during both training and evaluation. Appendix B.5 gives the full protocol.

The contributions differ substantially across regimes and metrics (Figure 4). For unseen contexts, AUPRC increases from 0.238 without the five input groups to 0.422 with all groups. Same-pert Evidence and Perturbation Knowledge Sources account for 0.169 of this 0.184 gain and also explain most of the improvement in Spearman LFC and Discrimination Score L1. For unseen perturbations, the smaller AUPRC gain of 0.045 is distributed across all five groups. In contrast, Perturbation Knowledge Sources account for nearly the entire gain in Discrimination Score L1, contributing 0.205 of 0.207. These inputs therefore contribute more strongly to distinguishing perturbation responses than their AUPRC contribution alone suggests.

When both entities are unseen, the total AUPRC gain is only 0.010. Perturbation Knowledge Sources have a negative mean AUPRC attribution, but account for most of the improvement in DE direction match, Spearman LFC, and Discrimination Score L1. Thus, their contribution depends on the pre-

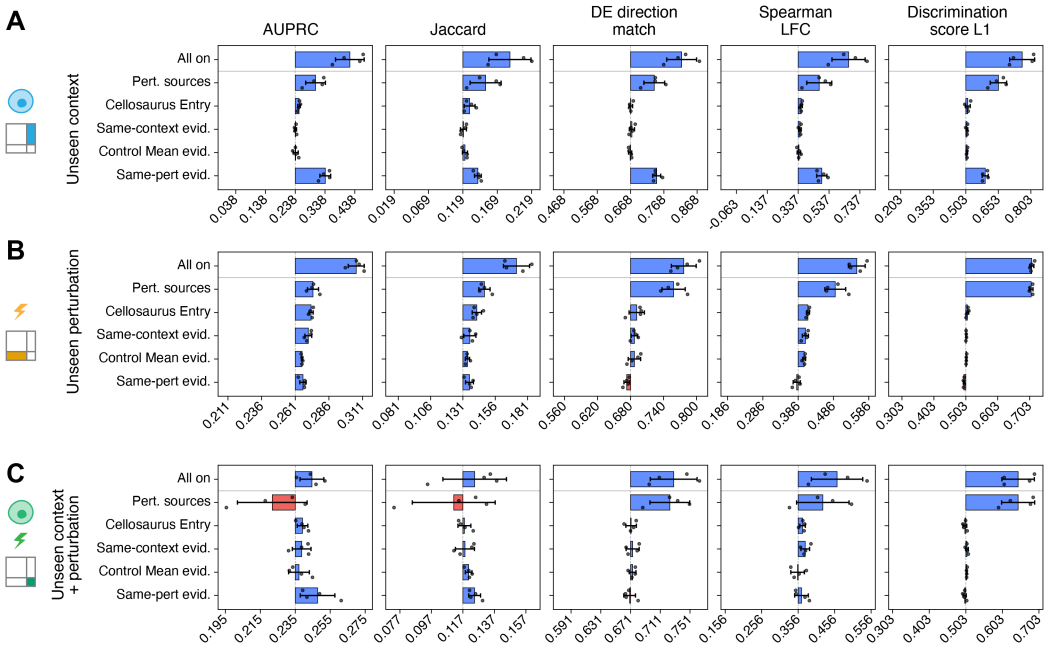

Figure 4: **Contribution of input signals.** Shapley values of five input groups on the within-dataset task, computed from models trained with every subset of the groups: Context Description (Cellosaurus entry), Perturbation Knowledge Sources (Pert. sources), Control Mean Evidence, Same-context Evidence, and Same-pert Evidence. (A), (B), and (C) are the three generalization regimes. Each axis is centered (dashed line) on the model without any group, and the top bar (All on) is the difference between the models with all and with no groups; tick labels are metric values. Bars show means and std across four folds, each holding out one cell line; dots are folds. Same-pert Evidence and the Perturbation Knowledge Sources account for most of the AUPRC gain for unseen contexts.

diction target: improvements in response direction, magnitude, and specificity do not necessarily translate into improved DE ranking. Appendix B.5 further separates the contributions of the individual Perturbation Knowledge Sources.

5 Conclusion

We introduced PIE, a perturbation effect prediction model, and demonstrated its ability to generalize to unseen cellular contexts, perturbations, and experimental datasets. PIE's architecture incorporates external Knowledge Sources, control expression, and response evidence from observed conditions, which together provide a richer representation of each experimental condition than perturbation data alone.

PIE outperformed baselines specialized for unseen-context, unseen-perturbation, or unseen-dataset generalization, and its use of external knowledge allowed it to retain predictive signal when these axes were combined. Ablations revealed that the contributions of its inputs differ substantially across generalization settings and metrics.

A central challenge in predicting perturbation effects in unseen conditions is separating biological signal from experimental variation. Our approach simplified the learning task by supervising on population-level responses while incorporating prior biological knowledge. Although this improved generalization over existing methods, performance declined when contexts and perturbations were jointly unseen, and inputs contributed little to DE ranking in that setting. These results suggest that more diverse perturbation data will be needed to further improve models. We intend PIE to serve as an extensible framework to guide future model development and data collection.

AI USE STATEMENT

In this work, we used generative AI tools to help develop conceptual frameworks, propose and refine hypotheses, design and provide feedback on research methodology and experiments, implement methods, clean and reformat datasets, support qualitative and thematic data analysis, and interpret results. We have not used generative AI tools to generate synthetic datasets. Formulating mathematical claims, providing critical ingredients for proving mathematical claims, assisting in the writing of proofs, and assisting with translation are not applicable to this work. Additionally, we used generative AI tools to create and modify scientific figures, suggest experimental parameters, create and edit software code, draft parts of the paper, summarize or analyse existing literature, discover research topics or identify gaps, brainstorm, source and search for information, edit the paper to improve readability, identify relevant literature, format references, suggest a structure for the paper, and propose its title and keywords.

AI coding agents wrote all project code, including the model, the training and evaluation pipelines, and data preprocessing, under the direction and supervision of the authors. AI agents also executed experiments, orchestrating some of them autonomously under the same direction and supervision. AI tools helped design the templates that render database records as text for the Context Description, Perturbation Description, and NCBI Gene Annotations Knowledge Sources (Appendix A.2). The authors reviewed the code alongside AI review agents, and reviewed all results, hypotheses, and literature produced with AI assistance. We have reviewed all AI-assisted work and take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Abhinav Adduri, Yusuf Roohani, Noam Teyssier, Rishi Verma, Francis Chalissery, and Beatrice Bevilacqua. cell-load: Dataloaders for training models on huge single-cell datasets. https://github.com/ArcInstitute/cell-load, 2026a. Software, MIT license.

Abhinav K. Adduri, Dhruv Gautam, Beatrice Bevilacqua, Mohsen Naghipourfar, Alishba Imran, Rohan Shah, Noam Teyssier, Rishi Verma, Christopher Carpenter, Basak Eraslan, Francis Chalissery, Rajesh Ilango, Vishvak Subramanyam, Chiara Ricci-Tam, Sanjay Nagaraj, Aidan Winters, Mingze Dong, Stefanie Fellinger, Adam Krejci, Tilmann Burckstummer, Sravya Tirukkovular, Jeremy Sullivan, Brian S. Plosky, Nicholas D. Youngblut, Jure Leskovec, Luke A. Gilbert, Silvana Konermann, Patrick D. Hsu, Alexander Dobin, Dave P. Burke, Hani Goodarzi, and Yusuf H. Roohani. Predicting cellular responses to perturbation across diverse contexts with State. *Cell*, August 2026b. ISSN 0092-8674. doi: 10.1016/j.cell.2026.07.052. URL https://doi.org/10.1016/j.cell.2026.07.052.

Walid Ahmad, Elana Simon, Seyone Chithrananda, Gabriel Grand, and Bharath Ramsundar. ChemBERTa-2: Towards chemical foundation models. *CoRR*, abs/2209.01712, 2022. doi: 10.48550/ARXIV.2209.01712. URL https://doi.org/10.48550/arXiv.2209.01712.

Michael Ashburner, Catherine A. Ball, Judith A. Blake, David Botstein, Heather Butler, J. Michael Cherry, Allan P. Davis, Kara Dolinski, Selina S. Dwight, Janan T. Eppig, et al. Gene ontology: tool for the unification of biology. *Nature Genetics*, 25(1):25–29, May 2000. ISSN 1546-1718. doi: 10.1038/75556. URL https://doi.org/10.1038/75556.

Amos Bairoch. The Cellosaurus, a cell-line knowledge resource. *Journal of Biomolecular Techniques : JBT*, 29(2):25–38, July 2018. ISSN 1943-4731. doi: 10.7171/jbt.18-2902-002. URL https://doi.org/10.7171/jbt.18-2902-002.

Michael Bereket and Theofanis Karaletsos. Modelling cellular perturbations with the sparse additive mechanism shift variational autoencoder. *Advances in Neural Information Processing Systems*, 36:1–12, 2023.

Garth R. Brown, Vichet Hem, Kenneth S. Katz, Michael Ovetsky, Craig Wallin, Olga Ermolaeva, Igor Tolstoy, Tatiana Tatusova, Kim D. Pruitt, Donna R. Maglott, and Terence D. Murphy. Gene: a gene-centered information resource at NCBI. *Nucleic Acids Research*, 43(D1):D36–D42, January 2015. ISSN 0305-1048. doi: 10.1093/nar/gku1055. URL https://doi.org/10.1093/nar/gku1055.

Charlotte Bunne, Stefan G Stark, Gabriele Gut, Jacobo Sarabia Del Castillo, Mitch Levesque, Kjong-Van Lehmann, Lucas Pelkmans, Andreas Krause, and Gunnar Rätsch. Learning single-cell perturbation responses using neural optimal transport. *Nature methods*, 20(11):1759–1768, 2023.

Srinivas Niranj Chandrasekaran, Jeanelle Ackerman, Eric Alix, D. Michael Ando, John Arevalo, Melissa Bennion, Nicolas Boisseau, Adriana Borowa, Justin D. Boyd, Laurent Brino, et al. JUMP cell painting dataset: morphological impact of 136,000 chemical and genetic perturbations. *BioRxiv*, March 2023. doi: 10.1101/2023.03.23.534023. URL https://doi.org/10.1101/2023.03.23.534023.

Yiqun Chen and James Zou. Simple and effective embedding model for single-cell biology built from ChatGPT. *Nature Biomedical Engineering*, 9(4):483–493, April 2025. ISSN 2157-846X. doi: 10.1038/s41551-024-01284-6.

Steven M. Corsello, Rohith T. Nagari, Ryan D. Spangler, Jordan Rossen, Mustafa Kocak, Jordan G. Bryan, Ranad Humeidi, David Peck, Xiaoyun Wu, Andrew A. Tang, et al. Discovering the anticancer potential of non-oncology drugs by systematic viability profiling. *Nature Cancer*, 1(2):235–248, January 2020. ISSN 2662-1347. doi: 10.1038/s43018-019-0018-6. URL https://doi.org/10.1038/s43018-019-0018-6.

Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine Learning*, 20(3):273–297, September 1995. ISSN 1573-0565. doi: 10.1007/bf00994018. URL https://doi.org/10.1007/BF00994018.

Haotian Cui, Chloe Wang, Hassaan Maan, Kuan Pang, Fengning Luo, Nan Duan, and Bo Wang. scgpt: toward building a foundation model for single-cell multi-omics using generative ai. *Nature methods*, 21(8):1470–1480, 2024.

Joshua M. Dempster, Isabella Boyle, Francisca Vazquez, David E. Root, Jesse S. Boehm, William C. Hahn, Aviad Tsherniak, and James M. McFarland. Chronos: a cell population dynamics model of CRISPR experiments that improves inference of gene fitness effects. *Genome Biology*, 22(1): 343, December 2021. ISSN 1474-760X. doi: 10.1186/s13059-021-02540-7. URL https://doi.org/10.1186/s13059-021-02540-7.

DepMap, Broad. DepMap public 25Q3. Dataset. https://depmap.org, 2025.

Alexander Dobin. cell-eval2: GPU-accelerated evaluation of perturbation prediction models. https://github.com/ArcInstitute/cell-eval2, 2026a. Software, MIT license.

Alexander Dobin. gpudge: GPU Mann-Whitney differential expression for single-cell CRISPR screens. https://github.com/ArcInstitute/gpudge, 2026b. Software, version 0.9.1, MIT license.

Mingze Dong, Abhinav Adduri, Dhruv Gautam, Christopher Carpenter, Rohan Shah, Chiara Ricci-Tam, Yuval Kluger, Dave P Burke, and Yusuf H Roohani. Stack: In-context learning of single-cell biology. *bioRxiv*, 2026.

Minsheng Hao, Jing Gong, Xin Zeng, Chiming Liu, Yucheng Guo, Xingyi Cheng, Taifeng Wang, Jianzhu Ma, Xuegong Zhang, and Le Song. Large-scale foundation model on single-cell transcriptomics. *Nature methods*, 21(8):1481–1491, 2024.

Dan Hendrycks and Kevin Gimpel. Gaussian error linear units (GELUs). *arXiv preprint arXiv:1606.08415*, 2016. doi: 10.48550/arXiv.1606.08415.

Alex Henry, Prudhvi Raj Dachapally, Shubham Shantaram Pawar, and Yuxuan Chen. Query-key normalization for transformers. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 4246–4253. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.findings-emnlp.379. URL https://doi.org/10.18653/v1/2020.findings-emnlp.379.

Leon Hetzel, Simon Boehm, Niki Kilbertus, Stephan Günnemann, Fabian Theis, et al. Predicting cellular responses to novel drug perturbations at a single-cell resolution. *Advances in Neural Information Processing Systems*, 35:26711–26722, 2022.

Dewei Hu, Damian Szklarczyk, Christian von Mering, and Lars Juhl Jensen. SPACE: STRING proteins as complementary embeddings. *Bioinformatics*, 41(9):btaf496, September 2025. ISSN 1367-4811. doi: 10.1093/bioinformatics/btaf496. URL https://doi.org/10.1093/bioinformatics/btaf496.

Ann C Huang, Tsung-Han S Hsieh, Jiang Zhu, Jackson Michuda, Ashton Teng, Soohong Kim, Elizabeth M Rumsey, Sharon K Lam, Ikenna Anigbogu, Philip Wright, Mohamed Ameen, Kwontae You, Christopher J Graves, Hyunsung John Kim, Adam J Litterman, Rene V Sit, Alex Blocker, and Ci Chu. X-Atlas/Orion: Genome-wide Perturb-seq datasets via a scalable fix-cryopreserve platform for training dose-dependent biological foundation models. *bioRxiv*, June 2025. doi: 10.1101/2025.06.11.659105. URL https://doi.org/10.1101/2025.06.11.659105. Preprint.

Peter J. Huber. Robust estimation of a location parameter. *The Annals of Mathematical Statistics*, 35(1):73–101, March 1964. ISSN 0003-4851. doi: 10.1214/aoms/1177703732. URL https://doi.org/10.1214/aoms/1177703732.

Andrew Jaegle, Sebastian Borgeaud, Jean-Baptiste Alayrac, Carl Doersch, Catalin Ionescu, David Ding, Skanda Koppula, Daniel Zoran, Andrew Brock, Evan Shelhamer, Olivier J. Hénaff, Matthew M. Botvinick, Andrew Zisserman, Oriol Vinyals, and João Carreira. Perceiver IO: A general architecture for structured inputs & outputs. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net, 2022. URL https://openreview.net/forum?id=fILj7WpI-g.

Longda Jiang, Carol Dalgarno, Efthymia Papalexi, Isabella Mascio, Hans-Hermann Wessels, Huiyoung Yun, Nika Iremadze, Gila Lithwick-Yanai, Doron Lipson, and Rahul Satija. Systematic reconstruction of molecular pathway signatures using scalable single-cell perturbation screens. *Nature Cell Biology*, 27(3):505–517, February 2025. ISSN 1476-4679. doi: 10.1038/s41556-025-01622-z. URL https://doi.org/10.1038/s41556-025-01622-z.

Sunghwan Kim, Jie Chen, Tiejun Cheng, Asta Gindulyte, Jia He, Siqian He, Qingliang Li, Benjamin A Shoemaker, Paul A Thiessen, Bo Yu, Leonid Zaslavsky, Jian Zhang, and Evan E Bolton. PubChem 2025 update. *Nucleic Acids Research*, 53(D1):D1516–D1525, January 2025. ISSN 1362-4962. doi: 10.1093/nar/gkae1059. URL https://doi.org/10.1093/nar/gkae1059.

Dominik Klein, Jonas Simon Fleck, Daniil Bobrovskiy, Lea Zimmermann, Sören Becker, Alessandro Palma, Leander Dony, Alejandro Tejada-Lapuerta, Guillaume Huguet, Hsiu-Chuan Lin, et al. Cellflow enables generative single-cell phenotype modeling with flow matching. *bioRxiv*, pp. 2025–04, 2025.

Huan Liang and Rohit Singh. COMPASS: Component-wise inference of shared and gene-specific perturbation response. *bioRxiv*, August 2026. doi: 10.64898/2026.08.03.742643. URL https://doi.org/10.64898/2026.08.03.742643. Preprint.

Zeming Lin, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Nikita Smetanin, Robert Verkuil, Ori Kabeli, Yaniv Shmueli, Allan dos Santos Costa, Maryam Fazel-Zarandi, Tom Sercu, Salvatore Candido, and Alexander Rives. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379(6637):1123–1130, March 2023. ISSN 1095-9203. doi: 10.1126/science.ade2574. URL https://doi.org/10.1126/science.ade2574.

Russell Littman, Jacob Levine, Sepideh Maleki, Yongju Lee, Vladimir Ermakov, Lin Qiu, Alexander Wu, Kexin Huang, Romain Lopez, Gabriele Scalia, Tommaso Biancalani, David Richmond, Aviv Regev, and Jan-Christian Hütter. Gene-embedding-based prediction and functional evaluation of perturbation expression responses with PRESAGE. *bioRxiv*, June 2025. doi: 10.1101/2025.06.03.657653. URL https://doi.org/10.1101/2025.06.03.657653. Preprint.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=Bkg6RiCqY7.

Mohammad Lotfollahi, Anna Klimovskaia Susmelj, Carlo De Donno, Leon Hetzel, Yuge Ji, Ignacio L Ibarra, Sanjay R Srivatsan, Mohsen Naghipourfar, Riza M Daza, Beth Martin, et al. Predicting cellular responses to complex perturbations in high-throughput screens. *Molecular systems biology*, 19(6):MSB202211517, 2023.

McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University. Online Mendelian inheritance in man, OMIM®. https://omim.org/, 2026. Accessed 2026-02-09.

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(5):1228–1237, May 2025. doi: 10.1038/s41588-025-02169-3.

OpenAI. New embedding models and API updates. https://openai.com/index/new-embedding-models-and-api-updates/, January 2024. Published 2024-01-25.

Giovanni Palla, Sudarshan Babu, Payam Dibaeinia, James D Pearce, Donghui Li, Aly A Khan, Theofanis Karaletsos, and Jakub M Tomczak. Scalable single-cell gene expression generation with latent diffusion models. *arXiv preprint arXiv:2511.02986*, 2025.

Giovanni Palla, Alexander Hillsley, Yang-Joon Kim, and Loic A. Royer. Tabular foundation models are competitive cellular perturbation predictors across biological scales. *bioRxiv*, July 2026. doi: 10.64898/2026.06.28.735106. URL https://doi.org/10.64898/2026.06.28.735106. Preprint.

Zoe Piran, Niv Cohen, Yedid Hoshen, and Mor Nitzan. Disentanglement of single-cell data with biolord. *Nature Biotechnology*, 42(11):1678–1683, 2024.

Qwen Team. Qwen3.5: Towards native multimodal agents. https://qwen.ai/blog?id=qwen3.5, February 2026.

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, July 2022. doi: 10.1016/j.cell.2022.05.013.

Yusuf Roohani, Kexin Huang, and Jure Leskovec. Predicting transcriptional outcomes of novel multigene perturbations with gears. *Nature biotechnology*, 42(6):927–935, 2024.

Yusuf H. Roohani, Tony J. Hua, Po-Yuan Tung, Lexi R. Bounds, Feiqiao B. Yu, Alexander Dobin, Noam Teyssier, Abhinav Adduri, Alden Woodrow, Brian S. Plosky, Reshma Mehta, Benjamin Hsu, Jeremy Sullivan, Chiara Ricci-Tam, Nianzhen Li, Julia Kazaks, Luke A. Gilbert, Silvana Konermann, Patrick D. Hsu, Hani Goodarzi, and Dave P. Burke. Virtual Cell Challenge: Toward a Turing test for the virtual cell. *Cell*, 188(13):3370–3374, June 2025. ISSN 0092-8674. doi: 10.1016/j.cell.2025.06.008. URL https://doi.org/10.1016/j.cell.2025.06.008.

Yanay Rosen, Yusuf Roohani, Ayush Agrawal, Leon Samotorčan, Tabula Sapiens Consortium, Stephen R Quake, and Jure Leskovec. Universal cell embeddings: A foundation model for cell biology. *BioRxiv*, pp. 2023–11, 2023.

L. S. Shapley. *A Value for n-Person Games*, pp. 307–318. Princeton University Press, 1953. ISBN 9781400881970. doi: 10.1515/9781400881970-018.

Aravind Subramanian, Rajiv Narayan, Steven M. Corsello, David D. Peck, Ted E. Natoli, Xiaodong Lu, Joshua Gould, John F. Davis, Andrew A. Tubelli, Jacob K. Asiedu, et al. A next generation connectivity map: L1000 platform and the first 1,000,000 profiles. *Cell*, 171(6):1437–1452.e17, November 2017. ISSN 0092-8674. doi: 10.1016/j.cell.2017.10.049. URL https://doi.org/10.1016/j.cell.2017.10.049.

Valentine Svensson, Umair Khan, Hamed Heydari, Airol A. Ubas, Nicole Thomas, Daniele Merico, Hani Goodarzi, John Yu, Nima Alidoust, and Shreshth Gandhi. Back to basics: Observed statistics are sufficient to predict drug responses. *bioRxiv*, June 2026. doi: 10.64898/2026.06.09.731197. URL https://doi.org/10.64898/2026.06.09.731197. Preprint.

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, January 2025. ISSN 1362-4962. doi: 10.1093/nar/gkae1113. URL https://doi.org/10.1093/nar/gkae1113.

The Gene Ontology Consortium, Suzi A Aleksander, James P Balhoff, Seth Carbon, J Michael Cherry, Dustin Ebert, Marc Feuermann, Pascale Gaudet, Nomi L Harris, David P Hill, et al. The gene ontology knowledgebase in 2026. *Nucleic Acids Research*, 54(D1):D1779–D1792, December 2025. ISSN 1362-4962. doi: 10.1093/nar/gkaf1292. URL https://doi.org/10.1093/nar/gkaf1292.

The UniProt Consortium, Alex Bateman, Maria-Jesus Martin, Sandra Orchard, Michele Magrane, Aduragbemi Adesina, Shadab Ahmad, Emily H Bowler-Barnett, Hema Bye-A-Jee, David Carpentier, et al. UniProt: the universal protein knowledgebase in 2025. *Nucleic Acids Research*, 53(D1):D609–D617, November 2024. ISSN 1362-4962. doi: 10.1093/nar/gkae1010. URL https://doi.org/10.1093/nar/gkae1010.

Christina V Theodoris, Ling Xiao, Anant Chopra, Mark D Chaffin, Zeina R Al Sayed, Matthew C Hill, Helene Mantineo, Elizabeth M Brydon, Zexian Zeng, X Shirley Liu, et al. Transfer learning enables predictions in network biology. *Nature*, 618(7965):616–624, 2023.

Frederik Wenkel, Wilson Tu, Cassandra Masschelein, Hamed Shirzad, Liam Hodgson, Ihab Bendidi, Cian Eastwood, Shawn T Whitfield, Craig Russell, Yassir El Mesbahi, et al. Txpert: using multiple knowledge graphs for prediction of transcriptomic perturbation effects. *Nature Biotechnology*, pp. 1–8, 2026.

Biao Zhang and Rico Sennrich. Root mean square layer normalization. In Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett (eds.), *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pp. 12360–12371, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/1e8a19426224ca89e83cef47f1e7f53b-Abstract.html.

Jesse Zhang, Airol A Ubas, Richard de Borja, Valentine Svensson, Nicole Thomas, Neha Thakar, Ian Lai, Aidan Winters, Umair Khan, Matthew G. Jones, John D. Thompson, Vuong Tran, Joseph Pangallo, Efthymia Papalexi, Ajay Sapre, Hoai Nguyen, Oliver Sanderson, Maria Nigos, Olivia Kaplan, Sarah Schroeder, Bryan Hariadi, Simone Marrujo, Crina Curca Alec Salvino, Guillermo Gallareta Olivares, Ryan Koehler, Gary Geiss, Alexander Rosenberg, Charles Roco, Daniele Merico, Nima Alidoust, Hani Goodarzi, and Johnny Yu. Tahoe-100M: A giga-scale single-cell perturbation atlas for context-dependent gene function and cellular modeling. *bioRxiv*, February 2025. doi: 10.1101/2025.02.20.639398. URL https://doi.org/10.1101/2025.02.20.639398. Preprint.

APPENDIX

A METHOD DETAILS

A.1 DATASETS AND PREPROCESSING

Every dataset starts from raw UMI counts. For genetic perturbation datasets, we apply the on-target knockdown filter of cell-load (Adduri et al., 2026a) in each context. It keeps perturbations that reduce their target gene below 30% of its control expression, keeps the cells within them that reduce it below 50%, and drops perturbations left with fewer than 30 cells. Control cells are never filtered. In every dataset, genetic or chemical, counts are then scaled to 10^4 per cell (CP10K), giving the normalized expression n_{ig} of gene g in cell i, and log-transformed, $x_{ig} = \log(1 + n_{ig})$, which gives the log-normalized expression.

- *Replogle–Nadig* combines the CRISPRi essential-gene screens of Replogle et al. (2022) in K562 and hTERT-RPE1 with those of Nadig et al. (2025) in HepG2 and Jurkat. We use raw counts restricted to the 6,642 genes shared by the four public releases (figshare for Replogle et al., 2022; GEO GSE264667 for Nadig et al., 2025), in which the knockdown filter had already been applied. The data comprise 5,459 experimental conditions over 4 contexts.
- *Tahoe-100M* (Zhang et al., 2025) profiles 379 compounds, each at 0.05, 0.5, and 5 µM, in 50 cancer cell lines, with DMSO-treated control cells per line. We keep the 95.6M of 100.6M cells that pass the release's cell filter, on its 18,151 genes, giving 56,827 experimental conditions.
- *Jiang et al.* (2025) is a CRISPRi screen of 218 signalling regulators in 6 cell lines, each under one of 5 cytokine stimuli, giving 30 contexts. We take raw counts from the five pathway objects on GEO (GSE281048) and concatenate them on their 33,056 shared genes. The knockdown filter keeps 461 of 1,626 experimental conditions.
- *X-Atlas/Orion* (Huang et al., 2025) is a genome-wide dual-guide CRISPRi screen in HCT116 and HEK293T, downloaded from Figshare+ (doi:10.25452/figshare.plus.29190726). We keep the 37,443 genes detected in both lines and, following the release paper, the perturbations with at least 25 cells, at least 30% mean knockdown, and an expressed target. The knockdown filter then keeps 13,284 experimental conditions.
- *Arc VCC 2025* is the H1 hESC CRISPRi data of the 2025 Virtual Cell Challenge (Roohani et al., 2025), on 18,151 genes. Its training and validation files are two contexts, each with its own copy of the control pool. The knockdown filter keeps 146 of 150 and 44 of 50 perturbations, giving 190 experimental conditions.

DE calls. DE is called with gpudge (Dobin, 2026b), separately for each context against its own control cells. For each (perturbation, gene), a two-sided Wilcoxon rank-sum (Mann–Whitney U) test compares the perturbed and control cells. A gene is tested only when its mean CPM exceeds 5 in the perturbed or in the control group; we say it *passes the CPM filter*. The p-values are Benjamini–Hochberg corrected per perturbation across the tested genes, and a gene is DE, $y_{cpg} = 1$, when the adjusted value is below 0.05. No fold-change criterion enters the call. An untested gene is not DE.

LFC. The LFC target is the log2 fold change that gpudge reports with each test. With $\bar{n}_g$ the arithmetic mean CP10K expression over a set of cells,

$$\ell_{cpg} \;=\; \log_2 \frac{\bar{n}_g(\mathcal{C}_{cp}) + \epsilon}{\bar{n}_g(\mathcal{C}_{c,\texttt{ctrl}}) + \epsilon}, \qquad \epsilon = 10^{-9}. \tag{5}$$

It is defined for DE genes with a finite fold change and is not clipped. A DE gene is up-regulated when $\ell_{cpg} > 0$ and down-regulated otherwise.

Pseudobulk delta expression. For context c, perturbation p, and gene g,

$$\Delta_{cpg} \;=\; \frac{1}{|\mathcal{C}_{cp}|} \sum_{i \in \mathcal{C}_{cp}} x_{ig} \;-\; \frac{1}{|\mathcal{C}_{c,\texttt{ctrl}}|} \sum_{i \in \mathcal{C}_{c,\texttt{ctrl}}} x_{ig}, \tag{6}$$

where x_{ig} is the log-normalized expression of cell i, $\mathcal{C}_{cp}$ the perturbed cells, and $\mathcal{C}_{c,\texttt{ctrl}}$ the control cells of context c. It is defined for every measured gene, whether or not the gene passes the CPM filter.

Measured genes. When a run trains on several datasets, each experimental condition is decoded and supervised only on the G genes of its own dataset.

A.2 Knowledge Sources

A.2.1 Sources

Context Knowledge Sources. The *Context Description* is the only context source. For a cell line, it is rendered from the Cellosaurus entry of the line (Bairoch, 2018):

```
Context Name: {context} ;
Recommended Name: {name} ;
Synonyms: {synonyms} ;
Cell Line Category: {category} ;
Organism: {organism} ;
Sex: {sex} ;
Age: {age} ;
Population: {population} ;
Disease: {disease} ;
Derived From Site: {site} ;
Cell Type: {cell type} ;
Parent Cell Line: {parent} ;
Genetic Integration: {label} via {method} ;
Sequence Variation: {genes}: {type}, {description} ({zygosity}; {note})
```

For Jiang et al., where a context is a cell line under a stimulus, the Cellosaurus entry is followed by a record of the stimulus. We derived one record for each of the five stimuli from the UniProt entry of its ligand (The UniProt Consortium et al., 2024), giving its name, family, receptors, canonical signalling pathway, and a short description:

```
{Cellosaurus entry} ;
Stimulation: {name} ({abbreviation}) ;
Stimulation Family: {family} ;
Stimulation Receptors: {receptors} ;
Stimulation Signaling: {pathway} ;
Stimulation Description: {description}
```

The rendered text is embedded with `text-embedding-3-large` (OpenAI, 2024) into one 3,072-dimensional vector. A text longer than the model's 8,191-token limit is split into windows with 50% overlap, and the window embeddings are averaged.

Genetic Perturbation Knowledge Sources.

- The *Perturbation Description* is rendered from the NCBI Gene entry of the target gene (Brown et al., 2015):

  ```
  Gene Name: {perturbation} ;
  HGNC Gene Symbol: {symbol} ;
  Full Name: {full name} ;
  Synonyms: {synonyms} ;
  Gene Type: {type} ;
  Organism: {organism} ;
  Lineage: {lineage} ;
  Map Location: {cytoband} ;
  Alt Designations: {designations} ;
  Gene Summary: {summary}
  ```

  It is embedded like the Context Description.

- *NCBI Gene Annotations* list, in full, the Gene Ontology terms (Ashburner et al., 2000; The Gene Ontology Consortium et al., 2025), OMIM disease associations (McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University, 2026), and gene-group memberships of the NCBI Gene record. GO terms are sorted by evidence count and diseases by confidence. The gene summary is left out, since the Perturbation Description carries it:

  ```
  Gene: {symbol}

  === GO BIOLOGICAL PROCESS ===
  ```

```
Gene Ontology biological process annotations, sorted by evidence count.
- {GO ID}: {term} (evidence: {count})

=== GO MOLECULAR FUNCTION ===
(same layout)

=== GO CELLULAR COMPONENT ===
(same layout)

=== DISEASE ASSOCIATIONS ===
Known disease associations with confidence scores and OMIM (Online Mendelian
Inheritance in Man) identifiers.
- {disease} (confidence: {score}, MIM: {MIM ID})

=== GENE GROUPS ===
Gene family and group memberships by relationship type.
- {relationship}: {members}
```

An empty section reads "None". The text is encoded by `Qwen3.5-35B-A3B-Base` (Qwen Team, 2026), and the final hidden state of every token is kept, one 2,048-dimensional vector per token. This is the only per-token source. Texts run from 122 to 5,874 tokens, below the 8,192-token limit.

- *ESM2* (Lin et al., 2023): the `esm2_t33_650M_UR50D` embedding of the protein sequence, mean-pooled over residues. Sequences longer than 1,022 residues are embedded in windows of 1,022 with stride 511.
- *STRING* (Szklarczyk et al., 2025): the 512-dimensional SPACE embedding (Hu et al., 2025) of the STRING v12.0 protein network.
- *DepMap* (DepMap, Broad, 2025; Dempster et al., 2021): the Chronos gene effect (release 25Q3) across 1,186 cell lines, with missing values set to 0.

Chemical Perturbation Knowledge Sources.

- The *Perturbation Description* is rendered from the PubChem entry of the compound (Kim et al., 2025), followed by the dose:

```
Compound Name: {name} ;
Synonyms: {first 10 synonyms} ;
Molecular Formula: {formula} ;
Molecular Weight: {weight} g/mol ;
SMILES: {SMILES} ;
Connectivity SMILES: {connectivity SMILES} ;
XLogP: {XLogP} ;
Topological Polar Surface Area: {TPSA} square angstroms ;
Complexity: {complexity} ;
Formal Charge: {charge} ;
Hydrogen Bond Donors: {donors} ;
Hydrogen Bond Acceptors: {acceptors} ;
Rotatable Bonds: {bonds} ;
Compound Description: {description} ;
Dose: {dose} micromolar
```

  It is embedded like the Context Description.
- *SMILES* (Ahmad et al., 2022): the `ChemBERTa-77M-MTR` embedding of the SMILES string, mean-pooled over tokens (384 dimensions).
- *L1000* (Subramanian et al., 2017): the transcriptional activity score of the compound across 83 cell lines (GEO GSE92742 and GSE70138). For each line, the signature nearest the perturbation's dose and a 24 h time point is used.
- *PRISM* (Corsello et al., 2020): secondary-screen $\log_2$ viability across 480 cell lines, at the measured dose nearest the perturbation's.
- *JUMP* (Chandrasekaran et al., 2023): 737 processed Cell Painting features. JUMP does not vary the dose, so every dose of a compound receives the same vector.

The three profiles enter unstandardized, and a cell line without a measurement is set to 0. No source contains measurements from the datasets of Appendix A.1. DepMap includes some of the same cell lines, such as K562, but from independent CRISPR screens.

Table 1: **Knowledge Source coverage.** Contexts or perturbations for which each source resolves, out of the total in the dataset. Tahoe-100M counts drug–dose perturbations (379 compounds × 3 doses); a compound is covered at all three doses or at none. –, the source does not apply to the dataset's perturbation type.

Source	Replogle–Nadig	Jiang et al.	X-Atlas/Orion	Arc VCC 2025	Tahoe-100M
Contexts	4	30	2	2	50
Perturbations	1,857	146	10,007	190	1,137
Context Description	100%	100%	100%	100%	100%
Perturbation Description	100%	100%	100%	100%	100%
NCBI Gene Annotations	97.5%	99.3%	98.9%	100%	–
ESM2	97.5%	99.3%	98.6%	100%	–
STRING	98.8%	99.3%	98.3%	100%	–
DepMap	97.1%	99.3%	97.4%	99.5%	–
SMILES	–	–	–	–	99.5%
L1000	–	–	–	–	58.3%
PRISM	–	–	–	–	30.6%
JUMP	–	–	–	–	71.0%

A.2.2 Integration into PIE

Let source i be present for an experimental condition with m_i vectors of width d_i. Each vector passes through an RMSNorm over d_i and a bias-free linear map to the model width $d = 768$. A learned source-type embedding, indexed from a fixed registry of source names, is added to every token of the source, with element-wise dropout of 0.05, and a per-source RMSNorm follows. The present sources are concatenated into one key-value set of $M = \sum_i m_i$ tokens, with no positional encoding, so an absent source contributes no tokens. Within a batch, each source is zero-padded to its longest sequence, and cross-attention masks the padded positions. A condition for which no source resolves receives a single all-zero carrier token, so the key-value set is never empty. A run uses the sources that fit its perturbations: the four gene sources for Replogle–Nadig and Jiang et al., the four compound sources for Tahoe-100M, and all ten in the cross-dataset recipe, each with both descriptions.

A.3 Per-gene evidence

Channels. Each response statistic below comes with a presence flag and the log number of contributing responses. Delta expression and LFC are divided by a global training standard deviation.

- *Control Mean Evidence*: the gene's mean log-normalized expression over control cells of the context, given as the value, its log, and a not-expressed flag. It is always present.
- *Same-pert Evidence*: over the training responses to the same perturbation, the mean and standard deviation of the gene's pseudobulk delta expression and of its LFC, and the fractions of tested responses in which it is up- or down-regulated. In the cross-dataset recipe, a provenance block also records which training datasets measured the perturbation. The channel is empty for an unseen perturbation, and for a held-out dataset unless another dataset measured the perturbation.
- *Same-context Evidence*: the same statistics over the training responses in the same context. It is empty for an unseen context and for a held-out dataset.

Donors. All response evidence comes from the training set.

Fusion. Each statistic has its own small encoder. The output of a response encoder is multiplied by its presence flag, so an empty channel contributes exactly zero. The encoded blocks are concatenated and fused by an MLP into a 128-dimensional evidence vector per gene (Figure 5). The vector is added to the gene's query and conditions the task adapter before each head (Figure 1).

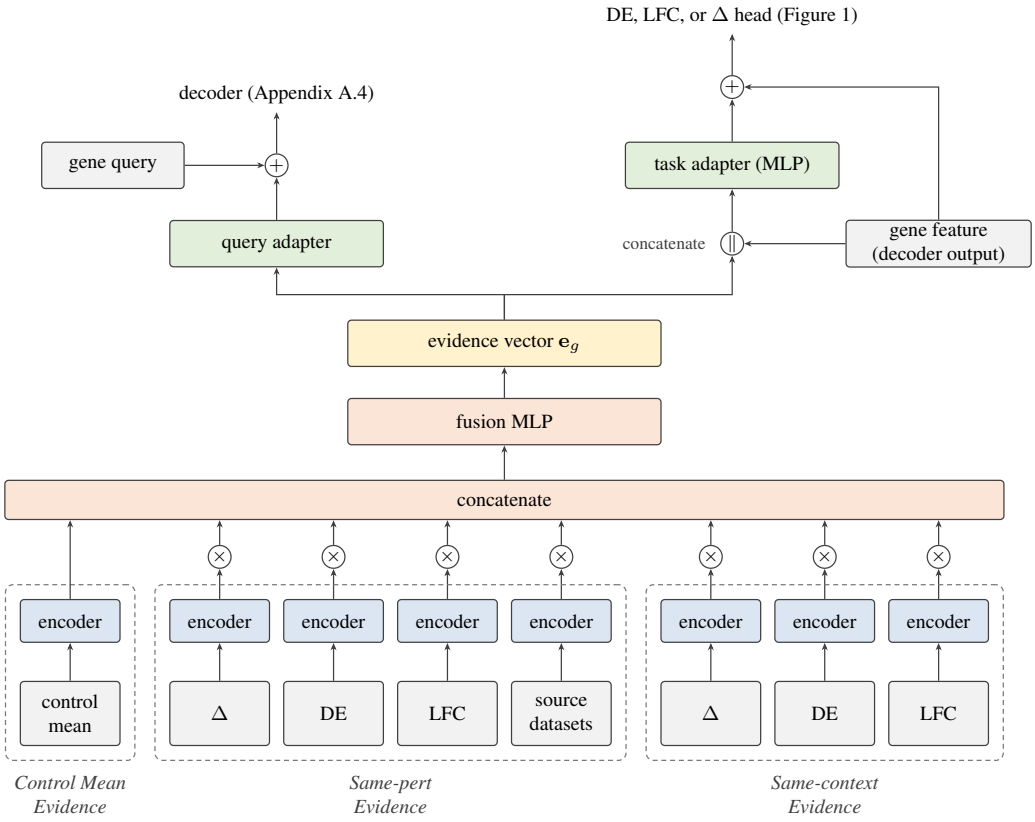

Figure 5: **Evidence fusion.** Each block of per-gene evidence is encoded separately, and the response blocks are gated by their presence flags. The concatenated blocks are fused into the evidence vector $\mathbf{e}_g$, which a query adapter adds to the gene query and task adapters pass to the three heads.

A.4 Encode–process–decode trunk

Blocks. The trunk has three stages (Figure 6). A learned array of latents cross-attends the Knowledge Source tokens, the latents then pass through self-attention blocks, and the gene queries cross-attend the latents. Each stage stacks several blocks. Every block applies attention and a two-layer GELU feed-forward (Hendrycks & Gimpel, 2016), each pre-normalized with RMSNorm (Zhang & Sennrich, 2019) and wrapped in a residual connection. Attention applies query–key normalization with a learned per-head temperature (Henry et al., 2020). Cross-attention also normalizes its keys and values. Gene queries do not attend to each other, so each gene is decoded independently of the others.

Gene queries. A gene's query is built from its NCBI Gene entry, the same text as the Perturbation Description of a genetic perturbation (Section 3.2; Appendix A.2), embedded with `text-embedding-3-large`. A query MLP maps the embedding to the model width, the query adapter adds the gene's evidence vector (Appendix A.3), and an RMSNorm follows. The Perturbation Description has its own projection, separate from the query MLP. Because a query depends only on the gene's text, there is no learned per-gene parameter, and the gene vocabulary can differ across datasets. Both in training and at inference, each experimental condition is decoded on the full gene vocabulary of its dataset.

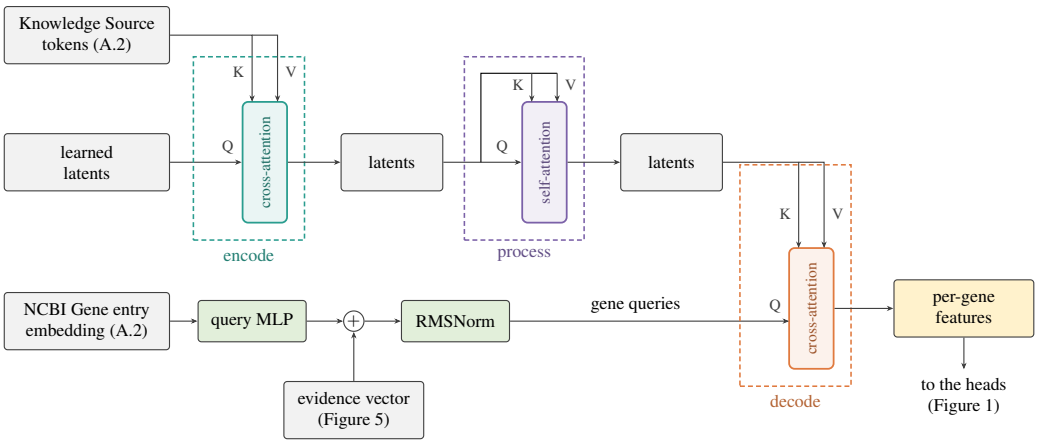

Figure 6: **Encode–process–decode trunk.** Following Perceiver IO (Jaegle et al., 2022), learned latents cross-attend the Knowledge Source tokens (encode), attend to one another (process), and are read out through cross-attention by one query per gene (decode). Each gene query is the gene's embedded NCBI Gene entry, passed through a query MLP, summed with its evidence vector, and normalized.

A.5 Training formulation

Each head is a linear layer applied to $\mathbf{h}_g + a(\mathbf{e}_g, \mathbf{h}_g)$, where $\mathbf{h}_g$ is the decoded gene feature, $\mathbf{e}_g$ the evidence vector, and a a head-specific MLP on their concatenation (Figure 5). As in Section 3.5, $\mathcal{G}_{cp}$ is the set of genes of an experimental condition (c, p) that pass the CPM filter (Appendix A.1), $\mathcal{D}_{cp} \subseteq \mathcal{G}_{cp}$ its DE genes, and G the number of measured genes. We drop the subscript cp from per-gene quantities.

DE head. The label is the binary DE call y_g of Appendix A.1. The head outputs two logits, and the DE probability $\hat{\pi}_g$ is the softmax probability of the DE class; a gene is called DE when $\hat{\pi}_g \geq 0.5$. $\mathcal{L}_{\mathrm{DE}}$ is the cross-entropy over $\mathcal{G}_{cp}$, with each gene weighted by the inverse frequency of its class in the condition.

LFC head. The label is the $\log_2$ fold change ℓ_g of Appendix A.1, defined for DE genes. The head outputs one value $\hat{\ell}_g$, whose sign gives the predicted direction. $\mathcal{L}_{\mathrm{LFC}}$ and $\mathcal{L}_{\mathrm{dir}}$ are given in Equation 2. $\mathcal{L}_{\mathrm{LFC}}$ is the Huber loss H_κ with $\kappa = 1$ between $\hat{\ell}_g$ and ℓ_g over $\mathcal{D}_{cp}$; for a genetic perturbation, the target gene's term is down-weighted. The direction loss is a hinge loss (Cortes & Vapnik, 1995) with zero margin on the sign of the prediction; it is zero when $\hat{\ell}_g$ has the sign of ℓ_g.

Delta-expression head. The label is the pseudobulk delta expression Δ_g of Appendix A.1, discretized on a grid fixed once per run from its training rows. For each training dataset s, let R_s be the 99.9th percentile of $|\Delta|$ and φ_s its minimum fold change, a resolution constant of 1.1 for Replogle–Nadig and Tahoe-100M and 1.01 for the other datasets. With $\Delta_{\max} = \max_s R_s$ and $w = \min_s \ln \varphi_s$, the grid has the smallest odd number of bins $K \geq 2\Delta_{\max}/w$ over $[-\Delta_{\max}, \Delta_{\max}]$, so one bin is centred on zero; values beyond $\pm\Delta_{\max}$ are clamped. The target is two-hot: its mass is split between the two bins whose centres bracket the value, so that the expected centre equals the value. The head outputs K logits, and the predicted delta expression is the centre of the most probable bin. Adding the predicted delta expression to the control mean gives the predicted pseudobulk expression. $\mathcal{L}_\Delta$ is the cross-entropy against the two-hot target over the G measured genes with a finite Δ_g, with each gene weighted by the inverse frequency of its target bin in the condition. The grid is computed from training data only, and validation and test values are binned on this same grid.

Training. Parameters are trained from random initialization with AdamW (Loshchilov & Hutter, 2019), with a learning rate that warms up linearly, stays constant, and then decays on a cosine

schedule. The Knowledge Source embeddings are precomputed and fixed. Each response-evidence channel is emptied, independently, for a random subset of experimental conditions. The checkpoint with the best validation AUPRC is kept.

A.6 HYPERPARAMETERS

Table 2: **Hyperparameters.** Shared by the within-dataset (Section 4.2) and cross-dataset (Section 4.3) runs except where noted.

Group	Setting	Value
Trunk	model width d	768
	latents	512
	attention heads (width)	8 (96)
	encoder / process / decoder blocks	4 / 6 / 4
	feed-forward width	1,536
	dropout	0.1
Knowledge Sources	source-type embedding dropout	0.05
Evidence	per-block encoder width	64
	evidence vector width	128
	fusion dropout	0.1
	response-channel dropout	0.25
Losses	DE class-weight cap	10
	LFC Huber threshold	1
	target-gene LFC weight	0.001
	direction scale τ	0.25
	delta-expression bin-weight cap	5
	delta-expression range percentile	99.9
Optimization	optimizer	AdamW, $\beta = (0.9, 0.999)$
	training steps	5e3 within dataset; 1e4 cross dataset
	learning rate	10^{-4}
	weight decay	0.01
	schedule	5% linear warmup, 4,250 steps at peak, cosine decay to 10^{-6} over the rest
	effective batch size	128
	precision	bf16 mixed
	gradient clipping	10

B EXPERIMENT DETAILS

B.1 EVALUATION METRICS

Notation. For a test experimental condition (c, p) of context c and perturbation p, dropping the subscript cp, let $\mathcal{D}$ be the reference DE genes among the G genes measured in the test dataset, and ℓ_g and Δ_g the reference LFC and pseudobulk delta expression of gene g (Appendix A.1). A model predicts a DE probability $\hat{\pi}_g$, an LFC $\hat{\ell}_g$, and a pseudobulk delta expression $\hat{\Delta}_g$ for every measured gene, and its predicted DE set is $\hat{\mathcal{D}} = \{g : \hat{\pi}_g \geq 0.5\}$. For a genetic perturbation, $\mathcal{D}^-$ and $\hat{\mathcal{D}}^-$ denote these sets without the target gene of p; otherwise they equal $\mathcal{D}$ and $\hat{\mathcal{D}}$.

Metrics.

- **AUPRC** is the average precision of the measured genes ranked by $\hat{\pi}_g$, with the genes in $\mathcal{D}$ as positives. An experimental condition whose genes are all positive or all negative scores 0.
- **Jaccard score** is $|\mathcal{D}^- \cap \hat{\mathcal{D}}^-|/|\mathcal{D}^- \cup \hat{\mathcal{D}}^-|$, and 1 when the union is empty.
- **DE direction match** is the fraction of genes in $\mathcal{D}$ whose $\hat{\ell}_g$ has the sign of ℓ_g, and 0 when $\mathcal{D}$ is empty. It does not depend on $\hat{\mathcal{D}}$.

- **Spearman effect size** is the Spearman correlation between $|\mathcal{D}|$ and $|\hat{\mathcal{D}}|$ across the perturbations of a context with $|\mathcal{D}| > 0$.
- **Spearman LFC** is the Spearman correlation between $\hat{\ell}_g$ and ℓ_g over the genes in $\mathcal{D}$.
- **Discrimination Score L1.** Let N be the number of perturbations in the context and r_p the number of other perturbations q whose reference profile $\Delta^{(q)}$ is closer in L1 distance to the prediction $\hat{\Delta}^{(p)}$ than p's own reference $\Delta^{(p)}$ is. For a genetic perturbation, the target gene of p is removed from all profiles. The score is $1 - r_p/N$, so 1 is perfect and random profiles score about 0.5. Predicted deltas are clipped so that control expression plus delta is nonnegative.

Aggregation. Spearman effect size is computed once per context; the other five metrics are computed per experimental condition and averaged over the perturbations of a context. Each context thus contributes one value per metric, and we report the unweighted mean over contexts. A Spearman correlation is undefined when fewer than two values enter it or either side is constant; such values are excluded from the means.

Adaptation. CELL-EVAL and CELL-EVAL2 score single-cell predictions. They run a Wilcoxon rank-sum test of predicted perturbed cells against control cells, call a gene DE at Benjamini–Hochberg adjusted $p < 0.05$, rank genes by $-\log_{10}$ adjusted p for AUPRC, take LFCs from the test, and compute the Discrimination Score on pseudobulk expression. Our adaptation replaces the predicted significance with $\hat{\pi}_g$: DE calls are $\hat{\pi}_g \geq 0.5$, and AUPRC ranks by $\hat{\pi}_g$. The LFC metrics take $\hat{\ell}_g$, and the Discrimination Score takes $\hat{\Delta}_g$. Otherwise the metrics follow the v2 definitions of CELL-EVAL2, including its values for degenerate experimental conditions, with one exception: an undefined Spearman correlation is excluded from the mean instead of scoring -1. Discrimination Score L1 uses the CELL-EVAL implementation of Adduri et al. (2026b).

Rescoring baselines. Each baseline's output is converted to $\hat{\pi}_g$, $\hat{\ell}_g$ and $\hat{\Delta}_g$ and scored by the same code as PIE.

- *Single cells* (STATE; TabICL for unseen contexts). We run gpudge (Dobin, 2026b), with the settings used for the reference labels, on the predicted perturbed cells against the control cells of the model's output, which for STATE are its own predicted controls. This gives an adjusted p-value q_g and $\hat{\ell}_g$. We set $\hat{\pi}_g = s_g/(s_g + s_0)$ with $s_g = -\log_{10} \max(q_g, 10^{-10})$ and $s_0 = -\log_{10} 0.05$. $\hat{\pi}_g$ decreases with q_g: $q_g = 0.05$ gives $\hat{\pi}_g = 0.5$, $q_g < 0.05$ gives $\hat{\pi}_g \in (0.5, 0.885]$, and $q_g > 0.05$ gives $\hat{\pi}_g \in [0, 0.5)$, with $\hat{\pi}_g = 0$ at $q_g = 1$. The floor of 10^{-10} on q_g sets the upper bound of 0.885; without it, $q_g = 0$ would make s_g infinite and $\hat{\pi}_g$ undefined. A gene is therefore called DE ($\hat{\pi}_g \geq 0.5$) exactly when $q_g \leq 0.05$, the threshold of the reference labels. $\hat{\Delta}_g$ is the difference in mean log-normalized expression between predicted perturbed and control cells. Genes removed by the gpudge expression filter receive $q_g = 1$ and $\hat{\ell}_g = 0$.
- *Pseudobulk shifts* (perturbation mean, context mean, COMPASS-N, PRESAGE; TabICL for unseen perturbations). We draw 500 control cells of the test context once without replacement and add the predicted shift to each, clipping at zero. The resulting cells and the 500 controls are then converted as single cells.
- *DE statistics* (Rhaister-O). Rhaister-O predicts an adjusted p-value, an LFC, and a delta for each gene. The adjusted p-value is mapped to $\hat{\pi}_g$ as above, and the other two are used directly.

For an unseen context, context mean predicts a zero shift. It therefore calls no gene DE, predicts $\hat{\ell}_g = \hat{\Delta}_g = 0$, and leaves both Spearman correlations undefined.

B.2 BASELINES

Each baseline is run with its authors' code where available, and its predictions are converted and scored as in Appendix B.1.

- STATE (Adduri et al., 2026b) is a transformer over sets of cells, trained with an energy-distance loss, that maps the control cells of a context to perturbed cells. It represents each perturbation by a learned embedding and so applies only to perturbations seen in training.
- *Perturbation mean* (Adduri et al., 2026b) predicts the control mean of the test context plus the perturbation's pseudobulk delta expression averaged over the training contexts that measured it. It applies only to seen perturbations.
- *Context mean* (Adduri et al., 2026b) predicts the control mean of the test context plus the mean pseudobulk delta expression of that context's training perturbations, the same for every perturbation. For a context without training responses, it predicts the control mean.
- *Rhaister-O* (Svensson et al., 2026) is a zero-shot model. It predicts a perturbation's DE statistics in a new context from that context's control pseudobulk, through a per-perturbation, per-gene intercept and slope fit on the training DE statistics. Perturbations measured in too few training contexts fall back to the additive model of the same work. It applies only to seen perturbations.
- *COMPASS-N* (Liang & Singh, 2026) decomposes the training responses in the test context into shared axes, loadings, and residuals, and predicts an unseen perturbation from its nearest neighbours in the SPACE protein embedding (Hu et al., 2025), with 5 neighbours for the loadings and 10 for the residual. Perturbations without a SPACE embedding receive the training mean. It requires training responses in the test context and a gene target.
- *PRESAGE* (Littman et al., 2025) embeds each perturbation from 40 published gene knowledge graphs and two co-expression graphs computed from the training responses, and is trained per context. It requires training responses in the test context and a gene target.
- *TabICL* (Palla et al., 2026) uses the frozen TabICL regressor in two modes. For unseen perturbations, it regresses the principal components of pseudobulk delta expression on those of perturbation features, with the training responses of the test context as the in-context examples. The features are CellProfiler morphology, BioGPT, ESM2, STRING, and DepMap for genetic perturbations, and PIE's small-molecule Knowledge Sources (perturbation text, SMILES, L1000, PRISM, and JUMP) for *Tahoe-100M*; a missing source enters as zeros. For unseen contexts, it maps perturbed cells of the training contexts to the test context by optimal-transport matching of control cells, which uses no perturbation features.

B.3 GENERALIZATION WITHIN DATASET

In all three datasets, the validation set is 10% of each training context's training experimental conditions, and PIE is trained at learning rate 10^{-4} for 5,000 optimizer steps. Test-set sizes below are for unseen contexts, unseen perturbations, and both unseen, in that order.

Replogle–Nadig (Replogle et al., 2022; Nadig et al., 2025). The four cell lines contribute 611,211 cells. Each of four folds has one test cell line and 1,259–1,273 randomly drawn training perturbations, and every validation perturbation keeps a training response. The test sets hold 867–983, 1,172–1,252, and 386–470 experimental conditions and, with the training and validation sets, cover every experimental condition exactly once. There are 17 delta-expression bins (Appendix A.5), or 13 when RPE1 is the test line. The selected checkpoints lie at steps 1,518–2,631, and we report the mean and sample standard deviation across folds.

Jiang et al. (2025). The 30 contexts are six cell lines (A549, BxPC-3, HAP1, HT-29, K562, MCF7) under five stimuli (IFN-β, IFN-γ, insulin, TGF-β, TNF-α). On-target knockdown filtering leaves 461 experimental conditions over 146 perturbations. Each of six folds has one test cell line, whose five contexts are the test contexts, and 92–102 randomly drawn training perturbations. The test sets hold 21–76, 92–128, and 6–43 experimental conditions. The selected checkpoints lie at steps 800–2,171. A fold's score averages the contexts of its test set, and we report the mean and sample standard deviation across folds. Spearman effect size and Discrimination Score L1 are undefined in a context with one test perturbation.

Tahoe-100M (Zhang et al., 2025). The 50 contexts are cancer cell lines, and the 1,137 perturbations are 379 small molecules at 0.05, 0.5, and 5 μM, with DMSO-treated cells as controls. A

single fold has five test cell lines and 265 training molecules at all three doses (795 perturbations), giving 32,173 training and 3,588 validation experimental conditions and test sets of 3,377, 13,055, and 1,451. The SMILES embedding and the measured compound profiles replace the gene-target Knowledge Sources (Appendix A.2). The selected checkpoint is at step 4,914. We report the mean and sample standard deviation across cell lines: the five test lines for unseen contexts and both unseen, and the 45 training lines for unseen perturbations.

Baselines. Context mean is evaluated in all three regimes and is the only baseline when both entities are unseen.

- STATE is trained per fold on the cells of the training experimental conditions and the control cells of all contexts, for 50,000 steps at learning rate 10^{-4}. We select the checkpoint with the lowest validation loss on Replogle–Nadig and use the final checkpoint on Jiang et al. It is evaluated on unseen contexts and not run on *Tahoe-100M*.
- *Perturbation mean* is computed from each fold's training set and evaluated on unseen contexts.
- *Rhaister-O* is fit per fold on the training DE statistics, without a validation set, and evaluated on unseen contexts. Its full model, rather than the fallback, covers about a third of the Replogle–Nadig test experimental conditions, none of Jiang et al., and 5 of 3,377 of *Tahoe-100M*.
- *COMPASS-N* has no trainable parameters and is evaluated on unseen perturbations in Replogle–Nadig and Jiang et al.; it is not run on *Tahoe-100M*, whose perturbations have no gene target.
- *PRESAGE* is trained once per training context and fold, with early stopping on validation loss and, on Replogle–Nadig, the per-cell-line learning rates of the original paper. It is evaluated on unseen perturbations in Replogle–Nadig and Jiang et al. and, like COMPASS-N, not run on *Tahoe-100M*.
- *TabICL* is evaluated on unseen perturbations and unseen contexts in all three datasets. For the Jurkat context of Replogle–Nadig, only the STRING and DepMap features are available. On *Tahoe-100M*, it is fit once per training cell line on PIE's small-molecule Knowledge Sources. Neither mode applies when both entities are unseen.

Tables 3–5 give the scores of all three datasets.

Table 3: **Within-dataset generalization, unseen contexts.** Mean ± sample standard deviation across folds (Replogle–Nadig, four; Jiang et al., six) or, for *Tahoe-100M*, across the five test cell lines of its single fold. Bold marks the best mean per dataset and metric; –, unable to perform the task; NaN, the metric is undefined. Higher is better for all metrics.

Model	AUPRC	Jaccard	DE direction match	Spearman effect size	Spearman LFC	Discrimination Score L1
Replogle–Nadig						
STATE	0.138 ± 0.033	0.094 ± 0.024	0.622 ± 0.026	0.362 ± 0.385	0.244 ± 0.040	0.600 ± 0.017
Perturbation mean	0.088 ± 0.022	0.079 ± 0.018	0.806 ± 0.031	0.440 ± 0.106	0.534 ± 0.029	0.659 ± 0.073
Context mean	0.085 ± 0.023	0.076 ± 0.035	0.000 ± 0.000	NaN	NaN	0.502 ± 0.000
Rhaister-O	0.344 ± 0.028	0.173 ± 0.011	0.805 ± 0.032	**0.741 ± 0.053**	0.591 ± 0.060	0.712 ± 0.072
COMPASS-N	–	–	–	–	–	–
PRESAGE	–	–	–	–	–	–
TabICL	0.074 ± 0.019	0.075 ± 0.019	0.719 ± 0.029	0.633 ± 0.091	0.428 ± 0.055	0.631 ± 0.052
PIE	**0.424 ± 0.034**	**0.185 ± 0.023**	**0.819 ± 0.026**	0.734 ± 0.052	**0.664 ± 0.046**	**0.764 ± 0.061**
Jiang et al.						
STATE	0.084 ± 0.038	0.082 ± 0.083	0.525 ± 0.079	0.218 ± 0.176	0.252 ± 0.160	0.588 ± 0.020
Perturbation mean	0.020 ± 0.010	0.018 ± 0.009	0.538 ± 0.032	−0.032 ± 0.220	0.350 ± 0.152	0.586 ± 0.023
Context mean	0.022 ± 0.012	**0.119 ± 0.102**	0.000 ± 0.000	NaN	NaN	0.569 ± 0.026
Rhaister-O	0.186 ± 0.060	0.094 ± 0.066	0.578 ± 0.047	0.344 ± 0.179	0.280 ± 0.191	0.599 ± 0.022
COMPASS-N	–	–	–	–	–	–
PRESAGE	–	–	–	–	–	–
TabICL	0.017 ± 0.009	0.022 ± 0.011	0.443 ± 0.047	−0.137 ± 0.228	0.199 ± 0.093	0.584 ± 0.011
PIE	**0.214 ± 0.065**	0.077 ± 0.042	**0.704 ± 0.045**	**0.371 ± 0.153**	**0.523 ± 0.130**	**0.647 ± 0.039**
Tahoe-100M						
Perturbation mean	0.367 ± 0.136	0.264 ± 0.074	0.744 ± 0.021	0.307 ± 0.068	0.478 ± 0.058	0.813 ± 0.048
Context mean	0.373 ± 0.132	0.000 ± 0.000	0.000 ± 0.000	NaN	NaN	0.501 ± 0.000
Rhaister-O	0.665 ± 0.125	0.372 ± 0.045	0.684 ± 0.022	**0.830 ± 0.131**	0.403 ± 0.029	0.970 ± 0.019
TabICL	0.302 ± 0.122	0.375 ± 0.130	0.582 ± 0.028	−0.049 ± 0.121	0.372 ± 0.038	0.866 ± 0.042
PIE	**0.680 ± 0.094**	**0.400 ± 0.112**	**0.872 ± 0.053**	0.759 ± 0.090	**0.721 ± 0.104**	**0.982 ± 0.013**

Table 4: **Within-dataset generalization, unseen perturbations.** Mean ± sample standard deviation across folds (Replogle–Nadig, four; Jiang et al., six) or, for *Tahoe-100M*, across the 45 training cell lines of its single fold. Bold marks the best mean per dataset and metric; –, unable to perform the task; NaN, the metric is undefined. Higher is better for all metrics.

Model	AUPRC	Jaccard	DE direction match	Spearman effect size	Spearman LFC	Discrimination Score L1
Replogle–Nadig						
STATE	–	–	–	–	–	–
Perturbation mean	–	–	–	–	–	–
Context mean	0.082 ± 0.009	0.063 ± 0.011	0.675 ± 0.025	NaN	0.372 ± 0.035	0.503 ± 0.000
Rhaister-O	–	–	–	–	–	–
COMPASS-N	0.092 ± 0.010	0.083 ± 0.009	0.748 ± 0.009	0.336 ± 0.022	0.438 ± 0.024	0.645 ± 0.008
PRESAGE	0.091 ± 0.010	0.080 ± 0.009	0.745 ± 0.022	0.370 ± 0.071	0.444 ± 0.025	0.633 ± 0.014
TabICL	0.094 ± 0.010	0.085 ± 0.010	0.769 ± 0.022	0.442 ± 0.025	0.469 ± 0.036	0.682 ± 0.010
PIE	**0.304 ± 0.023**	**0.168 ± 0.011**	**0.775 ± 0.022**	**0.525 ± 0.030**	**0.552 ± 0.022**	**0.709 ± 0.006**
Jiang et al.						
STATE	–	–	–	–	–	–
Perturbation mean	–	–	–	–	–	–
Context mean	0.026 ± 0.004	0.024 ± 0.004	0.632 ± 0.030	0.051 ± 0.110	0.273 ± 0.054	0.623 ± 0.010
Rhaister-O	–	–	–	–	–	–
COMPASS-N	0.026 ± 0.004	0.024 ± 0.004	0.613 ± 0.027	0.060 ± 0.106	0.263 ± 0.081	0.623 ± 0.011
PRESAGE	0.025 ± 0.004	0.024 ± 0.004	0.625 ± 0.036	−0.032 ± 0.084	0.290 ± 0.059	0.624 ± 0.008
TabICL	0.026 ± 0.004	0.024 ± 0.004	0.601 ± 0.033	−0.027 ± 0.072	0.195 ± 0.054	0.630 ± 0.009
PIE	**0.112 ± 0.027**	**0.058 ± 0.019**	**0.707 ± 0.019**	**0.240 ± 0.133**	**0.501 ± 0.058**	**0.640 ± 0.013**
Tahoe-100M						
Perturbation mean	–	–	–	–	–	–
Context mean	0.306 ± 0.135	0.222 ± 0.093	0.630 ± 0.068	−0.104 ± 0.166	0.287 ± 0.149	0.502 ± 0.000
Rhaister-O	–	–	–	–	–	–
TabICL	0.309 ± 0.136	0.221 ± 0.091	0.630 ± 0.071	−0.041 ± 0.082	0.273 ± 0.128	0.519 ± 0.008
PIE	**0.528 ± 0.109**	**0.316 ± 0.080**	**0.682 ± 0.094**	**0.031 ± 0.082**	**0.454 ± 0.159**	**0.531 ± 0.010**

Table 5: **Within-dataset generalization, both unseen.** Mean ± sample standard deviation across folds (Replogle–Nadig, four; Jiang et al., six) or, for *Tahoe-100M*, across the five test cell lines of its single fold. Bold marks the best mean per dataset and metric; –, unable to perform the task; NaN, the metric is undefined. Higher is better for all metrics.

Model	AUPRC	Jaccard	DE direction match	Spearman effect size	Spearman LFC	Discrimination Score L1
Replogle–Nadig						
STATE	–	–	–	–	–	–
Perturbation mean	–	–	–	–	–	–
Context mean	0.088 ± 0.018	0.091 ± 0.041	0.000 ± 0.000	NaN	NaN	0.502 ± 0.000
Rhaister-O	–	–	–	–	–	–
COMPASS-N	–	–	–	–	–	–
PRESAGE	–	–	–	–	–	–
TabICL	–	–	–	–	–	–
PIE	**0.239 ± 0.054**	**0.116 ± 0.022**	**0.721 ± 0.022**	**0.450 ± 0.061**	**0.460 ± 0.059**	**0.643 ± 0.049**
Jiang et al.						
STATE	–	–	–	–	–	–
Perturbation mean	–	–	–	–	–	–
Context mean	0.024 ± 0.013	**0.121 ± 0.108**	0.000 ± 0.000	NaN	NaN	0.621 ± 0.065
Rhaister-O	–	–	–	–	–	–
COMPASS-N	–	–	–	–	–	–
PRESAGE	–	–	–	–	–	–
TabICL	–	–	–	–	–	–
PIE	**0.077 ± 0.030**	0.043 ± 0.025	**0.615 ± 0.070**	**0.173 ± 0.600**	**0.359 ± 0.114**	**0.636 ± 0.064**
Tahoe-100M						
Perturbation mean	–	–	–	–	–	–
Context mean	0.396 ± 0.127	0.000 ± 0.000	0.000 ± 0.000	NaN	NaN	0.502 ± 0.000
Rhaister-O	–	–	–	–	–	–
TabICL	–	–	–	–	–	–
PIE	**0.536 ± 0.105**	**0.288 ± 0.082**	**0.612 ± 0.024**	**−0.005 ± 0.066**	**0.323 ± 0.057**	**0.523 ± 0.006**

B.4 GENERALIZATION ACROSS DATASETS

In every held-out-dataset experiment, PIE is trained on the other four datasets for 10,000 optimizer steps at learning rate 10^{-4} with a global batch of 128 experimental conditions and validated every 100 steps. The held-out dataset contributes no training or validation response, and contexts of a held-out cell line are removed from the training and validation sets. A context derived from a held-out cell line is discarded as well. Validation AUPRC averages over the contexts of each training dataset and then with equal weight over the four datasets. A perturbation is seen if it is measured in any context of any training dataset. The delta-expression bins have width $\ln 1.01$, with the clipping range set by the largest 99.9th percentile among the training datasets (Appendix A.5). Test scores are computed per context of the held-out dataset, and we report the mean and sample standard deviation across contexts. Test-set sizes below are for seen and unseen perturbations, in that order.

Replogle–Nadig held out. Training samples *Tahoe-100M* (weight 0.68; 51,144 experimental conditions, 45 cell lines, 379 small molecules, 1,137 perturbations), X-Atlas/Orion (0.30; 12,628, two lines, 9,679 perturbations), Jiang et al. (0.01; 342, 23 contexts, 139 perturbations), and the Virtual Cell Challenge (0.01; 146, one context). Removed: the five Jiang et al. K562 contexts and *Tahoe-100M* Hep-G2/C3A, derived from HepG2. Validation: 957 experimental conditions (*Tahoe-100M* 224 from four lines, X-Atlas/Orion 656, Jiang et al. 33 from two HT-29 contexts, Virtual Cell Challenge 44). Test: 2,286 (723 perturbations) and 3,173 (1,134), with 1,253, 1,439, 1,314, and 1,453 in HepG2, Jurkat, K562, and RPE1. Bins: 121; selected step: 4,700.

Jiang et al. held out. Training samples *Tahoe-100M* (0.584), X-Atlas/Orion (0.258), Replogle–Nadig (0.15), and the Virtual Cell Challenge (0.009): 65,597 experimental conditions, 49 contexts, 11,952 perturbations. Removed: Replogle–Nadig K562 and *Tahoe-100M* A-549 and HT-29. Validation: 1,172. Test: 227 (71 perturbations, 27 contexts) and 234 (75, 29). Bins: 155; selected step: 9,700.

X-Atlas/Orion held out. Training samples *Tahoe-100M* (0.826), Replogle–Nadig (0.15), Jiang et al. (0.012), and the Virtual Cell Challenge (0.012): 56,921 experimental conditions, 78 contexts,

3,161 perturbations; nothing removed. Validation: 613. Test: 841 (775 perturbations) and 12,443 (9,232), with 8,987 in HCT116 and 4,297 in HEK293T. Bins: 147; selected step: 8,300.

Virtual Cell Challenge held out. Training samples *Tahoe-100M* (0.584), X-Atlas/Orion (0.258), Replogle–Nadig (0.15), and Jiang et al. (0.009): 69,403 experimental conditions, 79 contexts, 12,017 perturbations; nothing removed. Validation: 1,225. Test: 121 and 69 in the single H1 context, so no standard deviation is reported. Bins: 147; selected step: 8,600.

Baselines. In every experiment, perturbation mean, context mean, Rhaister-O, and TabICL are evaluated on seen perturbations, and context mean alone on unseen perturbations.

- *Perturbation mean* averages, for each gene, the perturbation's pseudobulk delta expression over the training contexts, from any training dataset, that measure it and the gene.
- *Context mean* predicts the control mean of the held-out context, since no training response is measured there.
- *Rhaister-O* is fit on the DE statistics of the training set. Its full model covers no test experimental condition in any experiment, so every prediction comes from the fallback.
- *TabICL* uses its unseen-context mode, with the training contexts that measure the test context's seen perturbations as sources.

Tables 6 and 7 give the scores for all four held-out datasets.

Table 6: **Generalization across datasets, seen perturbations.** Mean ± sample standard deviation across the contexts of each held-out dataset: four cell lines for Replogle–Nadig, two for X-Atlas/Orion, and 27 cell line–stimulus contexts for Jiang et al.; the Virtual Cell Challenge has a single context and no standard deviation. Bold marks the best mean per dataset and metric; –, unable to perform the task; NaN, the metric is undefined. Higher is better for all metrics.

Model	AUPRC	Jaccard	DE direction match	Spearman effect size	Spearman LFC	Discrimination Score L1
Replogle–Nadig						
Perturbation mean	0.062 ± 0.013	0.052 ± 0.010	**0.733 ± 0.034**	0.439 ± 0.077	0.472 ± 0.041	0.561 ± 0.026
Context mean	0.064 ± 0.014	0.086 ± 0.043	0.000 ± 0.000	NaN	NaN	0.502 ± 0.000
Rhaister-O	0.283 ± 0.048	0.130 ± 0.029	0.730 ± 0.034	0.547 ± 0.101	0.476 ± 0.048	0.582 ± 0.034
TabICL	0.052 ± 0.011	0.051 ± 0.011	0.634 ± 0.030	0.452 ± 0.087	0.272 ± 0.084	0.540 ± 0.016
PIE	**0.378 ± 0.033**	**0.138 ± 0.009**	0.727 ± 0.035	**0.627 ± 0.064**	**0.508 ± 0.043**	**0.620 ± 0.055**
Jiang et al.						
Perturbation mean	0.025 ± 0.023	0.019 ± 0.018	0.497 ± 0.099	−0.017 ± 0.478	0.187 ± 0.167	0.581 ± 0.058
Context mean	0.027 ± 0.025	**0.123 ± 0.172**	0.000 ± 0.000	NaN	NaN	0.579 ± 0.045
Rhaister-O	0.073 ± 0.048	0.086 ± 0.119	0.445 ± 0.081	**0.197 ± 0.502**	0.141 ± 0.174	0.585 ± 0.051
TabICL	0.023 ± 0.022	0.026 ± 0.025	0.431 ± 0.175	−0.071 ± 0.556	0.139 ± 0.276	0.582 ± 0.070
PIE	**0.162 ± 0.069**	0.055 ± 0.078	**0.615 ± 0.115**	0.180 ± 0.494	**0.425 ± 0.269**	**0.586 ± 0.050**
X-Atlas/Orion						
Perturbation mean	0.016 ± 0.008	0.015 ± 0.008	0.620 ± 0.054	0.190 ± 0.111	0.405 ± 0.064	0.560 ± 0.024
Context mean	0.012 ± 0.007	0.134 ± 0.089	0.000 ± 0.000	NaN	NaN	0.504 ± 0.004
Rhaister-O	0.207 ± 0.046	**0.145 ± 0.032**	0.650 ± 0.055	0.576 ± 0.070	0.471 ± 0.081	0.571 ± 0.032
TabICL	0.010 ± 0.006	0.010 ± 0.006	0.590 ± 0.023	0.135 ± 0.206	0.376 ± 0.004	0.568 ± 0.039
PIE	**0.307 ± 0.070**	0.080 ± 0.000	**0.809 ± 0.073**	**0.628 ± 0.086**	**0.709 ± 0.065**	**0.657 ± 0.045**
Virtual Cell Challenge						
Perturbation mean	0.281	0.124	0.549	0.076	0.132	0.524
Context mean	0.289	0.000	0.000	NaN	NaN	0.506
Rhaister-O	0.299	0.005	0.550	0.536	0.133	0.524
TabICL	0.271	**0.202**	**0.571**	0.429	0.142	0.509
PIE	**0.395**	0.076	0.559	**0.604**	**0.147**	**0.678**

Table 7: **Generalization across datasets, unseen perturbations.** Mean ± sample standard deviation across the contexts of each held-out dataset: four cell lines for Replogle–Nadig, two for X-Atlas/Orion, and 29 cell line–stimulus contexts for Jiang et al.; the Virtual Cell Challenge has a single context and no standard deviation. Bold marks the best mean per dataset and metric; –, unable to perform the task; NaN, the metric is undefined. Higher is better for all metrics.

Model	AUPRC	Jaccard	DE direction match	Spearman effect size	Spearman LFC	Discrimination Score L1
Replogle–Nadig						
Perturbation mean	–	–	–	–	–	–
Context mean	0.102 ± 0.026	0.077 ± 0.034	0.000 ± 0.000	NaN	NaN	0.502 ± 0.000
Rhaister-O	–	–	–	–	–	–
TabICL	–	–	–	–	–	–
PIE	**0.252 ± 0.061**	**0.123 ± 0.018**	**0.603 ± 0.034**	**0.560 ± 0.029**	**0.278 ± 0.089**	**0.515 ± 0.009**
Jiang et al.						
Perturbation mean	–	–	–	–	–	–
Context mean	0.024 ± 0.037	**0.110 ± 0.170**	0.000 ± 0.000	NaN	NaN	**0.578 ± 0.050**
Rhaister-O	–	–	–	–	–	–
TabICL	–	–	–	–	–	–
PIE	**0.089 ± 0.077**	0.019 ± 0.017	**0.629 ± 0.136**	**0.066 ± 0.425**	**0.341 ± 0.256**	0.577 ± 0.075
X-Atlas/Orion						
Perturbation mean	–	–	–	–	–	–
Context mean	0.001 ± 0.001	**0.409 ± 0.082**	0.000 ± 0.000	NaN	NaN	0.500 ± 0.000
Rhaister-O	–	–	–	–	–	–
TabICL	–	–	–	–	–	–
PIE	**0.065 ± 0.016**	0.173 ± 0.066	**0.510 ± 0.141**	**0.096 ± 0.055**	**0.502 ± 0.064**	**0.504 ± 0.002**
Virtual Cell Challenge						
Perturbation mean	–	–	–	–	–	–
Context mean	0.196	0.000	0.000	NaN	NaN	0.510
Rhaister-O	–	–	–	–	–	–
TabICL	–	–	–	–	–	–
PIE	**0.273**	**0.031**	**0.541**	**0.409**	**0.088**	**0.516**

B.5 ABLATING SIGNAL COMPONENTS

Shapley values. Let N be the set of $n = 5$ players. A subset $S \subseteq N$ is a set of players present in a model, with the other players removed, and $v(S)$ is the test metric of the model trained with exactly the players in S. The Shapley value of player i sums over the subsets that exclude i (Shapley, 1953):

$$\phi_i = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!\,(n - |S| - 1)!}{n!} \big(v(S \cup \{i\}) - v(S)\big), \tag{7}$$

and the values satisfy the efficiency property of the same work,

$$\sum_{i \in N} \phi_i = v(N) - v(\emptyset). \tag{8}$$

In the experiment of Section 4.4, $v(\emptyset)$ is the model with all five groups removed. We compute ϕ_i exactly in each fold, where the values add up to that fold's difference.

Ablation of Perturbation Knowledge Sources. A second experiment attributes the metrics to the five Perturbation Knowledge Sources individually: the Perturbation Description, NCBI Gene Annotations, ESM2, STRING, and DepMap. The Context Description and the three evidence channels are present in every model, so the model without any source is the first experiment's model with only the Perturbation Knowledge Sources removed. The protocol is otherwise that of Section 4.4.

Removal. A removed input is absent in training and at test. A removed Knowledge Source contributes no tokens, and a sample for which no source remains receives the zero carrier token (Section 3.2). Removing Control Mean Evidence drops the control channel from the per-gene evidence, and removing Same-context or Same-pert Evidence sets that channel and its flag to zero in every training and test sample, the encoding of a channel with no contributing response (Section 3.3).

Models and training. Each experiment has $2^5 = 32$ models per fold, and the two experiments share two: the full model, and the model without Perturbation Knowledge Sources, which is the first experiment's model with only that group removed. This gives 62 models per fold and 248 in total. Each model follows the recipe of Section 4.2 with one seed and at most 5,000 optimizer steps, and the checkpoint with the highest validation AUPRC is used for all metrics and regimes. Every model stops early once validation AUPRC has not improved by 0.001 for 1,000 optimizer steps; the median model stops at step 3,282 (range 1,822 to 5,000, with 22 models reaching the step limit).

Results. The five sources add the most AUPRC for unseen perturbations, raising it from 0.281 to 0.306 (Figure 7). Of this gain, DepMap contributes 0.013, STRING 0.008, and NCBI Gene Annotations 0.006; ESM2 contributes 0.000 and the Perturbation Description -0.002. For Discrimination Score L1 in this regime, every source is positive, from 0.020 for ESM2 to 0.066 for DepMap, and the five together account for 0.206. For unseen contexts, the sources add 0.013 AUPRC, most of it from ESM2 (0.007) and DepMap (0.004). With both entities unseen, the sources leave AUPRC unchanged (-0.001). They raise DE direction match by 0.075, with every source contributing between 0.010 and 0.021, Spearman LFC by 0.101, and Discrimination Score L1 by 0.143. The Perturbation Description has a negative AUPRC value in all three regimes.

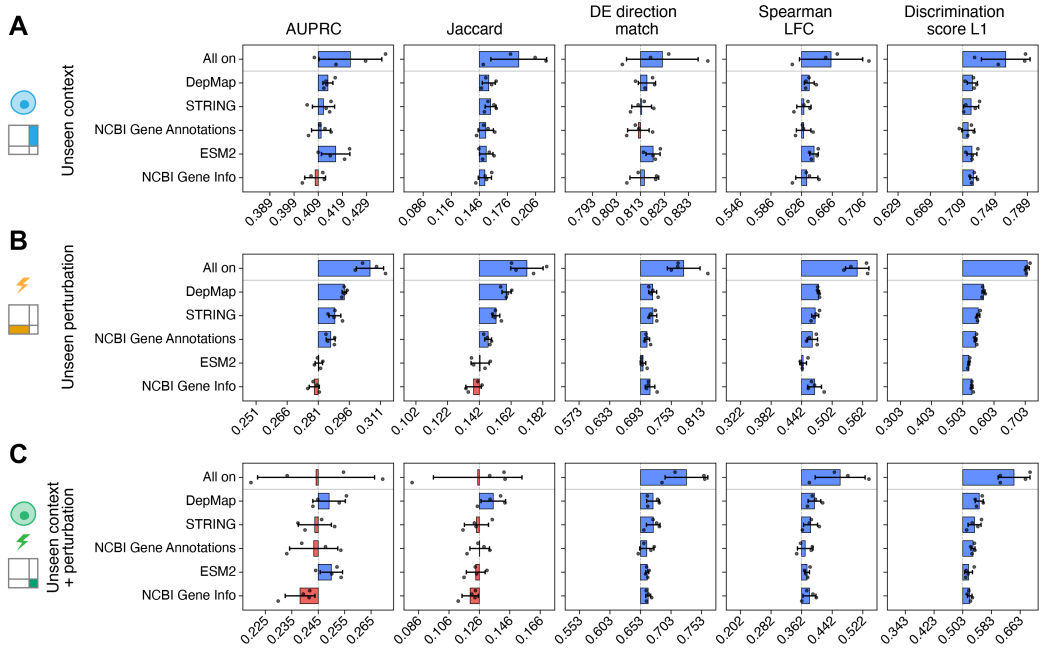

Figure 7: **Contribution of Perturbation Knowledge Sources.** Shapley values of the five Perturbation Knowledge Sources on the within-dataset task, computed from models trained with every subset of the sources: the Perturbation Description (NCBI Gene Info), NCBI Gene Annotations, ESM2, STRING, and DepMap, with the other four input groups present in every model. Rows are the three regimes. Each axis is centred (dashed line) on the model without any source, and the top bar (All on) is the difference between the models with all and with no sources; tick labels are metric values. Bars show means and std across four folds, each holding out one cell line; dots are folds. DepMap contributes the most AUPRC among the sources for unseen perturbations.