REPERT: EFFECT-GROUNDED IDENTITY LEARNING FOR GENERALIZABLE INVERSE PERTURBATION RETRIEVAL

Anonymous authors
Paper under double-blind review

ABSTRACT

Inverse perturbation retrieval seeks to identify the intervention responsible for an observed cellular response, yet current retrievers **generalize poorly** beyond the cellular settings seen during training. Two barriers underlie this failure: (i) *response distributions for the same perturbation can differ substantially across cell lines*, so patterns learned in one cell line may not transfer to another. (ii) *perturbation identity embeddings learned from training cell lines lack direct functional evidence* and transfer poorly to unseen ones. To address these barriers, we introduce **RePert**, a novel **Res**ponse-to-**Pert**urbation framework with functionally-grounded two-stage retrieval. RePert introduces two key innovations: (i) *Effect encoder pretraining* learns transferable effect representations from multiple responses sharing the same perturbation or target. (ii) *Effect-grounded identity retrieval* builds a non-parametric effect memory and uses its prototypes to ground identity embeddings. RePert combines identity and effect matching views through a reliability-calibrated score. We evaluate RePert in six *genetic and chemical* retrieval settings spanning in-domain evaluation, held-out cell-line generalization, and zero-shot cross-platform transfer to external single-cell platforms. For gene-level retrieval, RePert improves recall@10 over strongest baseline by factors of $\mathbf{4.41}\times$, $\mathbf{3.10}\times$, and $\mathbf{1.84}\times$ in single-gene, multiple-gene, and held-out cell-line retrieval, respectively. For compound-level cross-platform transfer without target-platform training, it improves top-1% macro recall against strongest baseline by factors of $\mathbf{1.17}\times$, $\mathbf{1.52}\times$, and $\mathbf{1.15}\times$ on Cellarity-Cancer, Cellarity-Primary, and sciPlex3, respectively. These results establish transferable effect learning and identity grounding as a robust and promising design for **generalizable inverse perturbation retrieval**. Code is available at `https://anonymous.4open.science/r/RePert-DEC3/`.

1 INTRODUCTION

Artificial intelligence for virtual cells (AIVC) seeks to model how cells behave under diverse genetic and chemical interventions Bunne et al. (2024); Rood et al. (2024). The growing scale of perturbational and single-cell transcriptomic datasets Replogle et al. (2022); Srivatsan et al. (2020) has enabled increasingly capable models for perturbation-response prediction Roohani et al. (2024); Theodoris et al. (2023); Bunne et al. (2023), functional screening Dixit et al. (2016); Yao et al. (2024), and drug discovery DeMeo et al. (2025); Gonzalez et al. (2025). Most of these models address the forward problem: given a perturbation, predict the cellular response it will induce Lotfollahi et al. (2019; 2023); Cui et al. (2024). We study the complementary problem of inverse perturbation retrieval: given an observed cellular response, identify the perturbation that produced it.

Existing inverse retrievers learn perturbation identities directly from cellular responses Wang et al. (2025); Gonzalez et al. (2025); DeMeo et al. (2025). However, response distributions for the same perturbation can differ substantially (Figure 1 top left) Nadig et al. (2025); Niepel et al. (2017); Peidli et al. (2024); Hetzel et al. (2022). A response-to-identity reverse perturbation retriever can therefore depend on cell-line-specific patterns and fail on unseen cell lines. As a result, such response distribution shifts directly limit cross-cell-line generalization, motivating our first question: **how can we learn transferable effect representations despite cell-line-dependent response variation?**

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

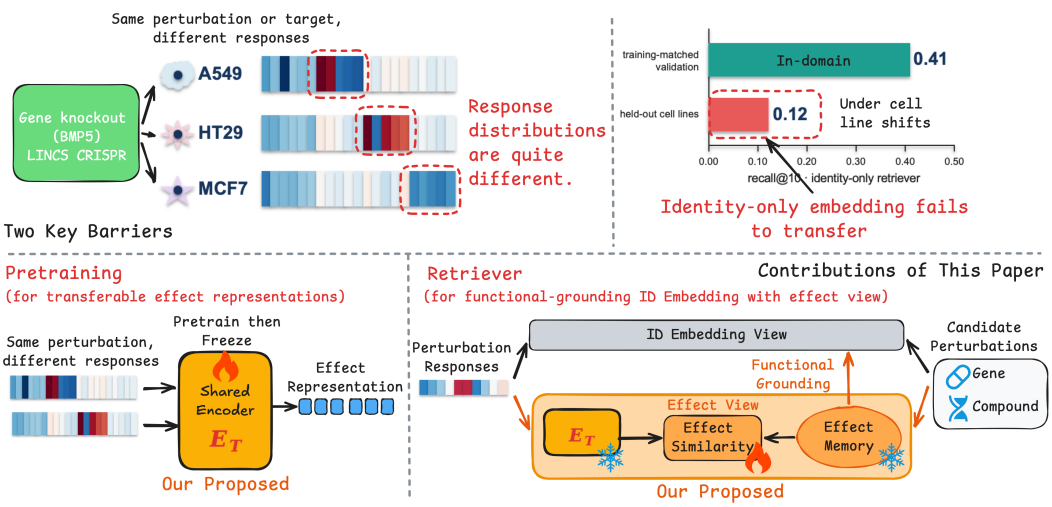


Figure 1: **Barriers and Contributions. Upper row:** The same perturbation exhibits markedly different response distributions across cell lines (left), while task C3 identity-only recall@10 drops from 0.41 on training-matched validation responses to 0.12 on held-out cell lines (right). Both barriers limit generalization. **Lower row:** RePert learns transferable effect representations from multiple responses, then uses an effect memory to functionally ground identity embeddings and combine identity and effect views.

Besides, learning transferable response representations alone is insufficient because existing retrievers still match each response to a trainable identity embedding for every retrieval candidate Gonzalez et al. (2025); DeMeo et al. (2025). Because these embeddings are optimized only against responses from the training cell lines, they can encode cell-line-specific response-to-identity associations and transfer poorly to held-out cell lines (Figure 1 top right). This motivates our second question: **how can identity-based retrieval generalize beyond the cell lines used to learn candidate embeddings?**

To address these challenges, we propose **RePert**, a two-stage framework for generalizable inverse perturbation retrieval. Figure 1 illustrates its two sequential components, which target to improve robustness to response-distribution shifts and the limited transferability of candidate identities. Specifically, *effect encoder pretraining* learns a transferable effect space from multiple responses sharing the same perturbation or target. *Effect-grounded identity retrieval* aggregates each candidate's training responses in this generalizable space into a non-parametric effect memory. This memory provides functional anchors for the trainable identity embeddings. Calibrating the identity and effect views preserves discriminative retrieval while improving robustness to held-out cell lines.

We establish an evaluation pipeline spanning *genetic and chemical* perturbations, in-domain retrieval, held-out cellular settings, and zero-shot cross-platform transfer to external assay platforms. Across six evaluation settings, RePert consistently improves over strong baselines. It raises recall@10 from 0.0179 to 0.0790 for single-gene retrieval and from 0.1246 to 0.3857 for multiple-gene retrieval, while improving recall@10 from 0.0719 to 0.1325 across held-out cell lines. It also transfers without target-platform training or calibration to Cellarity scRNA-seq and sciPlex3 snRNA-seq. Together, these results support a simple principle: **learn effect representations that transfer, then use them to ground perturbation identity retrieval**.

2 Related Work

Single-Cell Perturbation Modeling. Single-cell perturbation modeling studies how genetic and chemical interventions reshape cellular states. Methods such as scGen Lotfollahi et al. (2019), CPA Lotfollahi et al. (2023), and GEARS Roohani et al. (2024) learn forward mappings from cellular context and intervention identity to response, while recent foundation models seek representations that transfer across larger datasets and biological settings Adduri et al. (2025); Theodoris et al. (2023); Cui et al. (2024); Hao et al. (2024). Their central task is to predict what a known perturbation will

do. We instead invert the direction of inference and ask which perturbation produced an observed response, with emphasis on whether that inference transfers to cellular settings absent from training.

Inverse Perturbation Retrieval. Inverse perturbation retrieval identifies the intervention responsible for an observed response, extending the signature-matching principle established by the Connectivity Map and related resources Lamb et al. (2006); Subramanian et al. (2017); Evangelista et al. (2022). Recent approaches scale this inference to large candidate vocabularies through learned response-to-candidate compatibility Gonzalez et al. (2025) or response-based matching DeMeo et al. (2025). These methods demonstrate the promise of inverse retrieval but leave transfer across unseen cellular settings largely dependent on the learned response-to-identity mapping. RePert uses functional grounding to the identity space and complements identity-effect views to generalizable inverse perturbation retrieval.

3 METHOD

3.1 PROBLEM DEFINITION

Figure 2 provides an overview of RePert. Effect pretraining learns transferable effect representations from multiple responses sharing the same perturbation or target. Effect-grounded identity retrieval then grounds candidate identities in an effect memory and combines identity and effect views.

Let $\mathcal{P} = \{p_1, \ldots, p_N\}$ denote the retrieval-candidate pool and $\Delta \in \mathbb{R}^G$ the transcriptional shift from a matched control. Throughout, p denotes a retrieval candidate: the intervention identity in C1 and C4-C6, and an annotated target gene of the chemical intervention in C2 and C3. Given $\mathcal{D}_{\mathrm{tr}} = \{(\Delta_i, \mathcal{Y}_i)\}_{i=1}^n$, where the nonempty set $\mathcal{Y}_i \subseteq \mathcal{P}$ contains the ground-truth retrieval candidates for Δ_i, we learn $S(\Delta, p)$ to rank every member of $\mathcal{Y}_i$ above the candidates in $\mathcal{P} \setminus \mathcal{Y}_i$. In C1 and C4-C6, $\mathcal{Y}_i$ contains a single intervention identity. In C2 and C3, $\mathcal{Y}_i$ contains one or more annotated target genes, and each observation is expanded during training into one positive pair (Δ_i, p) for every $p \in \mathcal{Y}_i$. Generalization is evaluated on cell lines or assay platforms absent from training. The core notation used in this study is summarized in Appendix A.

3.2 EFFECT ENCODER PRETRAINING

Responses sharing the same perturbation or target can vary within or across cell lines. We use an effect encoder $E_T : \mathbb{R}^G \to \mathbb{R}^d$ to organize repeated responses according to their shared perturbation or target. For positive-pair construction, two responses are matched when they correspond to the same knockout in C1, share the same annotated target gene in C2 or C3, or correspond to the same compound in the drug-level tasks. In C2 and C3, an annotated target is a non-exclusive sampling key: a multi-target response enters the positive pool of every target in $\mathcal{Y}_i$ across training.

The encoder is a three-layer MLP with unit-normalized output

$$e = \mathrm{norm}(E_T(\Delta)) .$$

Each pretraining batch contains one pair of distinct responses for each of B sampled perturbations or targets. The two responses share the corresponding knockout, annotated target, or compound, while the remaining paired responses enter the contrastive normalizer. We minimize the symmetric effect-consistency objective

$$\mathcal{L}_{\mathrm{pre}} = -\frac{1}{2B} \sum_{i=1}^{B} \left[\log \frac{\exp(e_i^\top e_i'/\tau_p)}{\sum_{j=1}^{B} \exp(e_i^\top e_j'/\tau_p)} + \log \frac{\exp(e_i'^\top e_i/\tau_p)}{\sum_{j=1}^{B} \exp(e_i'^\top e_j/\tau_p)} \right] . \tag{1}$$

Here Δ_i' is a second response paired with Δ_i through a shared perturbation or target, e_i and e_i' are their effect representations, j indexes the other paired responses in the batch, and τ_p is the pretraining temperature. Equation 1 pulls the representations within each sampled positive pair together while contrasting them with the remaining paired responses in the batch. After pretraining, we freeze E_T so that retrieval training uses the resulting effect-representation space.

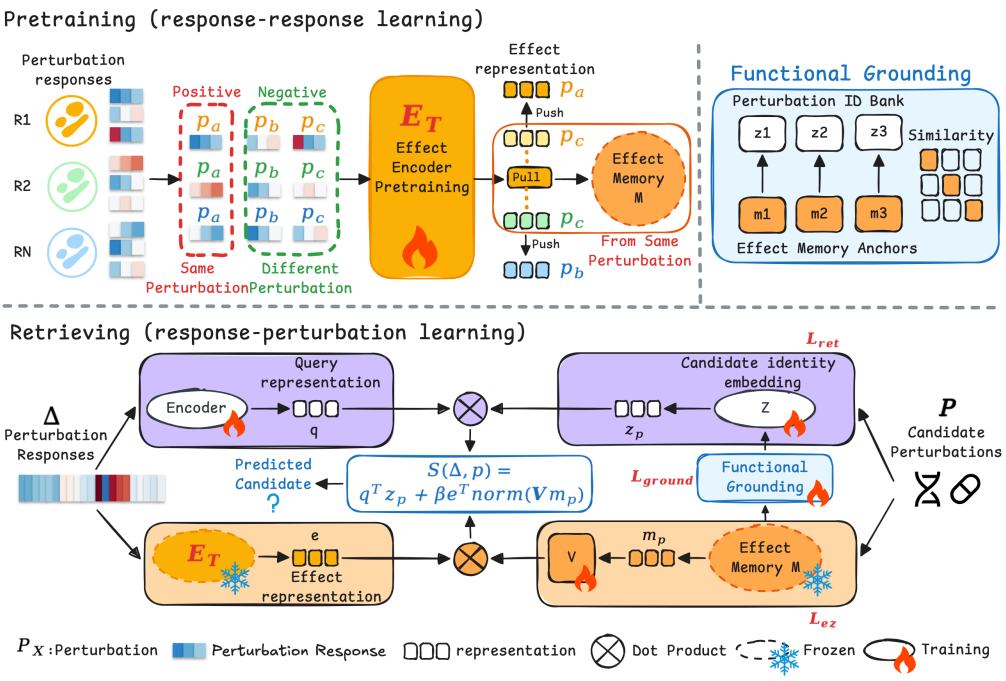


Figure 2: **Overview of effect-grounded inverse perturbation retrieval. Upper left:** A shared encoder E_T learns transferable effect representations from multiple responses sharing the same perturbation or target, and aggregates them into the fixed memory $M = \{m_p\}_{p\in\mathcal{P}}$. **Upper right:** Contrastive alignment functionally grounds each trainable candidate identity z_p in its memory prototype m_p. **Lower:** For response Δ, the identity and frozen effect branches score candidates against z_p and projected effect prototypes, respectively. The branches are trained by $\mathcal{L}_{\text{ret}}$ and $\mathcal{L}_{ez}$, and their calibrated sum gives $S(\Delta, p)$. Flames and snowflakes denote trainable and frozen components.

3.3 Effect-Grounded Identity Retrieval

The pretrained effect space emphasizes shared representation across responses associated with the same perturbation or annotated target, but generalization also requires functional evidence for candidates under cellular-setting shifts. RePert therefore retains the discriminative identity view and grounds it with a candidate-level effect view.

Discriminative identity learning. A trainable query encoder produces $q = \text{norm}(f(\Delta))$, and every retrieval candidate p has a unit-normalized identity embedding z_p. Their cosine $s_p^{\text{id}} = q^\top z_p$ is trained with full-pool cross-entropy:

$$\mathcal{L}_{\text{ret}} = -\log \frac{\exp\left(s_{p^\star}^{\text{id}}/\tau_q\right)}{\sum_{p'\in\mathcal{P}} \exp\left(s_{p'}^{\text{id}}/\tau_q\right)}.$$

Here $p^\star$ is the ground-truth candidate in the current expanded pair, p' ranges over $\mathcal{P}$, and τ_q is the temperature. For C2 and C3, summing the loss over all $p^\star \in \mathcal{Y}_i$ treats every annotated target as a positive. Up to the response weight $|\mathcal{Y}_i|$, this is equivalent to cross-entropy with a uniform target distribution over $\mathcal{Y}_i$, following the multi-positive outside-the-log form Khosla et al. (2020). This preserves direct identity supervision. To reduce input-scale differences, f applies a learned scale-and-clip head based on six log-magnitude statistics, whereas E_T receives the raw shift (Appendix B.1).

Non-parametric candidate effect memory. For every observed candidate, we encode and average its training responses to obtain

$$m_p = \text{norm}\left(\frac{1}{n_p}\sum_{i:p\in\mathcal{Y}_i} \text{norm}(E_T(\Delta_i))\right), \qquad n_p > 0, \tag{2}$$

Here n_p is the number of training responses whose valid-candidate set contains p, and $p \in \mathcal{Y}_i$ selects those responses when constructing its prototype m_p. We set $m_p = \mathbf{0}$ when no observation is available. Stacking the prototypes gives the effect memory $M \in \mathbb{R}^{N \times d}$. Unlike the learned index z_p, m_p is a non-parametric summary of the responses associated with candidate p and can be rebuilt from new observations without retraining the encoder.

For a query effect representation e computed by the frozen E_T, a trainable projection $\mathbf{V}$ adapts the stored prototypes for direct effect matching:

$$s_p^{\text{eff}} = e^\top \text{norm}(\mathbf{V} m_p). \tag{3}$$

We train this effect view with the corresponding full-pool cross-entropy:

$$\mathcal{L}_{ez} = -\log \frac{\exp\big(s_{p^\star}^{\text{eff}}/\tau_e\big)}{\sum_{p' \in \mathcal{P}} \exp\Big(s_{p'}^{\text{eff}}/\tau_e\Big)}.$$

Here $p^\star$ and p' have the same meanings as above, and τ_e is the effect-softmax temperature. The same multi-target interpretation applies to $\mathcal{L}_{ez}$. The query effect representation carries no gradient, so this objective learns only $\mathbf{V}$ and preserves the pretrained representation.

Functional grounding of candidate identities. Let $w_p = \text{norm}(\mathbf{W} m_p)$ be the aligned effect prototype of candidate p. For a sampled subset $\mathcal{S}$ of observed candidates, we optimize the symmetric contrastive loss

$$\mathcal{L}_{\text{ground}} = -\frac{1}{2|\mathcal{S}|} \sum_{p \in \mathcal{S}} \left[\log \frac{\exp(z_p^\top w_p/\tau_a)}{\sum_{p' \in \mathcal{S}} \exp(z_p^\top w_{p'}/\tau_a)} + \log \frac{\exp(w_p^\top z_p/\tau_a)}{\sum_{p' \in \mathcal{S}} \exp(w_p^\top z_{p'}/\tau_a)} \right]. \tag{4}$$

Here p indexes the anchor candidate in the sampled set $\mathcal{S}$, p' indexes all candidates contrasted with that anchor, $w_p = \text{norm}(\mathbf{W} m_p)$ is its projected effect prototype, and τ_a is the alignment temperature. $\mathbf{W}$ is trainable. Each identity must recognize its own effect prototype, and the contrastive objective transfers their relative organization into the identity embeddings. Gradients reach only z_p and $\mathbf{W}$. The memory and effect encoder remain fixed.

3.4 Dual-View Learning and Reliability-Calibrated Inference

We train identity classification, functional grounding, and effect retrieval jointly:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{ret}} + \lambda \mathcal{L}_{\text{ground}} + \mu \mathcal{L}_{ez}. \tag{5}$$

At inference, we combine the two kinds of evidence in the central RePert score

$$S(\Delta, p) = q^\top z_p + \beta\, e^\top \text{norm}(\mathbf{V} m_p). \tag{6}$$

The effect memory is not equally reliable under every condition: limited candidate support makes a prototype noisy, while a new assay platform can shift the query effect representation away from the prototypes stored in memory. We therefore select one evidence weight β for each trained retrieval model using only validation data from its training distribution. The value selected for the C4 model is reused unchanged for C5 and C6, without calibration on either external dataset. Calibration requires only re-scoring without retraining.

4 Experiments

4.1 Experimental Setup

Datasets. Table 1 summarizes six tasks across three evaluation regimes. C1, C2, and C4 are in-domain LINCS retrieval tasks following PDGrapher (Gonzalez et al., 2025). C3 follows the PDGrapher leave-cell-line-out split, while C5 and C6 evaluate zero-shot cross-platform transfer by applying the C4 model to external Cellarity and sciPlex3 platforms without target-platform training or calibration (DeMeo et al., 2025; Srivatsan et al., 2020). Here, zero-shot refers to the target platform. Both tasks retain C4's closed compound vocabulary.

Table 1: **Datasets and evaluation regimes.**

Task	Dataset	Retrieval target	G	Candidates	Train	Queries	Regime
C1	LINCS-CRISPR	Single gene	10,716	10,676	140,213	43,817	In-domain
C2	LINCS-chemical	Multiple genes	10,716	10,676	111,921	34,981	In-domain
C3	LINCS-chemical	Multiple genes	10,716	10,676	17,489	154,485	Held-out
C4	LINCS L1000	Compound	978	9,597	514,292	135,340	In-domain
C5	Cellarity (scRNA)	Compound	978	9,597	–	994	Cross-platform
C6	sciPlex3 (snRNA)	Compound	978	9,597	–	354	Cross-platform

Baselines. For gene-level retrieval (C1-C3), we compare with the nine methods: Random, Cancer genes, Cancer targets, and Perturbed genes (Gonzalez et al., 2025), scGen (Lotfollahi et al., 2019), ChemCPA (Hetzel et al., 2022), CellOT (Bunne et al., 2023), GEARS (Roohani et al., 2024), and Biolord (Piran et al., 2024), PDGrapher (Gonzalez et al., 2025). For drug-level retrieval (C4-C6), we compare with Random, non-parametric k-NN, FRoGS (Chen et al., 2024), PRnet (Qi et al., 2024) and TranSiGen (Tong et al., 2024), and DrugReflector (DeMeo et al., 2025).

Implementation. C1-C3 report recall@k, nDCG, and APS, whereas C4-C6 report recall@k and top-1% macro recall. Multi-target recall is the fraction of annotated targets retrieved. Results are summarized across evaluation cell lines for C1-C3 and three runs for C4-C6. Model selection and calibration use only validation data from the corresponding training distribution. Hyperparameters are reported in Appendix B.1.

4.2 In-Domain Retrieval

4.2.1 Gene Retrieval

Single-Gene Identification. C1 first tests whether functional grounding preserves discriminative power in domain. Given a LINCS CRISPR expression shift, the model must recover one knocked-out gene across 10 cell lines under the random split. Table 2 shows that RePert reaches recall@10 0.0790, improving over 0.0179 for PDGrapher by $4.41\times$ and over 0.0169 for scGen by $4.7\times$. The advantage grows at deeper cutoffs: recall@100 is 0.5076, compared with 0.1335 for scGen. RePert outperforms PDGrapher in every cell line (11), indicating that the gain is not concentrated in a few settings.

Multiple-Gene Identification. C2 extends the in-domain test from a single knockout to the full target set of a chemical perturbation, using all annotated genes as valid targets. Across cell lines (Table 2), RePert attains recall@10 0.3857, compared with 0.1246 for PDGrapher and 0.0748 for scGen, and raises nDCG from 0.2401 to 0.4634. Its advantage spans the ranking: recall@100 reaches 0.8401, versus 0.4368 for the strongest baseline. RePert again wins in every evaluation cell line (Table 12), showing that reusable effect representations and identity grounding remain effective when a response can correspond to multiple valid gene targets.

4.2.2 Compound Retrieval

C4 tests in-domain retrieval at compound scale, ranking $9{,}597$ compounds for $135{,}340$ LINCS Level-4 queries, of which $8{,}767$ compounds appear among the queries. RePert improves all five metrics over the retrained model by $1.15\times$ to $1.22\times$ (Table 2). It achieves top-1% macro recall 0.6071, compared with 0.4961, and recall@1 0.2086, compared with 0.1704. The gap shows that effect grounding scales beyond gene-level retrieval to a vocabulary of nearly ten thousand compounds.

4.3 Generalization across Cellular Settings

Held-Out Cell-Line Generalization. C3 directly evaluates single-source generalization to held-out cellular settings. Each model is trained on one chemical cell line and evaluated on the other eight, so neither the query encoder nor the effect memory has observed the evaluation cell line. Results are first averaged over the eight models that score each evaluation cell line and then summarized as mean $\pm$ s.d. across the nine evaluation cell lines (Table 3). RePert reaches recall@10 0.1325, compared with 0.0719 for PDGrapher and 0.0552 for the strongest forward-model transfer baseline, CellOT. It also leads on recall@1, recall@100, and nDCG. Whereas the forward models vary sharply across evaluation cell lines, RePert outperforms strong baseline PDGrapher on all cell lines (Table 13).

Table 2: **In-domain retrieval (C1, C2, C4).** C1–C2 report mean ± s.d. across cell lines; C4 reports mean ± s.d. over three runs. Best in **bold red**, second best in *italic blue*.

Method	Venue	R@1	R@10	R@100	nDCG	APS%
C1 – Single-gene (LINCS CRISPR, 10 cell lines, random split)						
Random	–	$0.0001_{\pm 0.0001}$	$0.0011_{\pm 0.0002}$	$0.0092_{\pm 0.0005}$	$0.0459_{\pm 0.0001}$	$0.01_{\pm 0.01}$
Cancer genes	–	$0.0002_{\pm 0.0001}$	$0.0011_{\pm 0.0005}$	$0.0093_{\pm 0.0011}$	$0.0459_{\pm 0.0003}$	$0.02_{\pm 0.01}$
Cancer targets	–	$0.0003_{\pm 0.0001}$	$0.0024_{\pm 0.0004}$	$0.0122_{\pm 0.0014}$	$0.0468_{\pm 0.0004}$	$0.03_{\pm 0.01}$
Perturbed genes	–	$0.0002_{\pm 0.0001}$	$0.0028_{\pm 0.0005}$	$0.0270_{\pm 0.0013}$	$0.0834_{\pm 0.0017}$	$0.02_{\pm 0.01}$
scGen	NMeth'19	$0.0019_{\pm 0.0011}$	$0.0169_{\pm 0.0078}$	$\mathit{0.1335}_{\pm 0.0594}$	$\mathit{0.1013}_{\pm 0.0166}$	$0.19_{\pm 0.11}$
CellOT	NMeth'23	$0.0003_{\pm 0.0001}$	$0.0039_{\pm 0.0003}$	$0.0382_{\pm 0.0025}$	$0.0594_{\pm 0.0011}$	$0.03_{\pm 0.01}$
GEARS	NBT'24	$0.0004_{\pm 0.0001}$	$0.0041_{\pm 0.0004}$	$0.0397_{\pm 0.0033}$	$0.0668_{\pm 0.0010}$	$0.04_{\pm 0.01}$
Biolord	NBT'24	$0.0005_{\pm 0.0002}$	$0.0045_{\pm 0.0007}$	$0.0421_{\pm 0.0026}$	$0.0750_{\pm 0.0014}$	$0.05_{\pm 0.02}$
PDGrapher	NBME'25	$\mathit{0.0069}_{\pm 0.0042}$	$\mathit{0.0179}_{\pm 0.0100}$	$0.0493_{\pm 0.0207}$	$0.0714_{\pm 0.0110}$	$\mathit{0.69}_{\pm 0.42}$
RePert (ours)	–	$\mathbf{0.0138}_{\pm 0.0067}$	$\mathbf{0.0790}_{\pm 0.0210}$	$\mathbf{0.5076}_{\pm 0.0688}$	$\mathbf{0.1813}_{\pm 0.0139}$	$\mathbf{1.38}_{\pm 0.67}$
gain over PDGrapher		*1.99×*	*4.41×*	*10.29×*	*2.54×*	*1.99×*
C2 – Multiple-gene (LINCS chemical, 9 cell lines, random split)						
Random	–	$0.0001_{\pm 0.0001}$	$0.0010_{\pm 0.0002}$	$0.0092_{\pm 0.0006}$	$0.0640_{\pm 0.0013}$	$0.29_{\pm 0.08}$
Cancer genes	–	$0.0013_{\pm 0.0019}$	$0.0095_{\pm 0.0134}$	$0.0176_{\pm 0.0123}$	$0.0692_{\pm 0.0079}$	$1.76_{\pm 2.18}$
Cancer targets	–	$0.0041_{\pm 0.0029}$	$0.0379_{\pm 0.0263}$	$0.0897_{\pm 0.0614}$	$0.0940_{\pm 0.0197}$	$3.60_{\pm 1.69}$
Perturbed genes	–	$0.0025_{\pm 0.0035}$	$0.0262_{\pm 0.0386}$	$0.2275_{\pm 0.2904}$	$0.1897_{\pm 0.0467}$	$5.20_{\pm 5.37}$
scGen	NMeth'19	$0.0134_{\pm 0.0113}$	$0.0748_{\pm 0.0574}$	$\mathit{0.4368}_{\pm 0.2150}$	$0.1979_{\pm 0.0264}$	$7.86_{\pm 4.46}$
ChemCPA	NeurIPS'22	$0.0084_{\pm 0.0115}$	$0.0493_{\pm 0.0521}$	$0.3387_{\pm 0.2496}$	$0.1664_{\pm 0.0314}$	$6.57_{\pm 4.88}$
CellOT	NMeth'23	$0.0067_{\pm 0.0063}$	$0.0525_{\pm 0.0503}$	$0.3513_{\pm 0.2442}$	$0.1675_{\pm 0.0269}$	$6.60_{\pm 4.10}$
Biolord	NBT'24	$0.0084_{\pm 0.0112}$	$0.0581_{\pm 0.0710}$	$0.3502_{\pm 0.2448}$	$0.1732_{\pm 0.0359}$	$6.88_{\pm 5.19}$
PDGrapher	NBME'25	$\mathit{0.0244}_{\pm 0.0248}$	$\mathit{0.1246}_{\pm 0.0599}$	$0.4262_{\pm 0.1716}$	$\mathit{0.2401}_{\pm 0.0575}$	$\mathit{16.90}_{\pm 3.89}$
RePert (ours)	–	$\mathbf{0.1144}_{\pm 0.0497}$	$\mathbf{0.3857}_{\pm 0.1202}$	$\mathbf{0.8401}_{\pm 0.0665}$	$\mathbf{0.4634}_{\pm 0.0760}$	$\mathbf{33.45}_{\pm 7.49}$
gain over PDGrapher		*4.69×*	*3.10×*	*1.97×*	*1.93×*	*1.98×*

Method	Venue	R@1	R@10	R@100	nDCG	top-1%
C4 – LINCS L1000, in-distribution						
Random	–	$0.0001_{\pm 0.000}$	$0.0010_{\pm 0.000}$	$0.0101_{\pm 0.000}$	$0.0466_{\pm 0.000}$	$0.0098_{\pm 0.001}$
k-NN	–	$0.0157_{\pm 0.000}$	$0.0574_{\pm 0.000}$	$0.1638_{\pm 0.000}$	$0.1158_{\pm 0.000}$	$0.1527_{\pm 0.000}$
FRoGS	NC'24	$0.1023_{\pm 0.001}$	$0.2555_{\pm 0.001}$	$0.4844_{\pm 0.001}$	$0.2657_{\pm 0.001}$	$0.2502_{\pm 0.001}$
PRnet	NC'24	$0.0056_{\pm 0.000}$	$0.0288_{\pm 0.001}$	$0.1042_{\pm 0.001}$	$0.0864_{\pm 0.000}$	$0.0523_{\pm 0.001}$
TranSiGen	NC'24	$0.0198_{\pm 0.000}$	$0.0772_{\pm 0.001}$	$0.2104_{\pm 0.001}$	$0.1284_{\pm 0.001}$	$0.1319_{\pm 0.001}$
DrugReflector	Science'25	$\mathit{0.1704}_{\pm 0.000}$	$\mathit{0.3980}_{\pm 0.001}$	$\mathit{0.6624}_{\pm 0.001}$	$\mathit{0.3636}_{\pm 0.001}$	$\mathit{0.4961}_{\pm 0.002}$
RePert (ours)	–	$\mathbf{0.2086}_{\pm 0.004}$	$\mathbf{0.4838}_{\pm 0.005}$	$\mathbf{0.7738}_{\pm 0.003}$	$\mathbf{0.4199}_{\pm 0.004}$	$\mathbf{0.6071}_{\pm 0.001}$
gain over DrugReflector		*1.22×*	*1.22×*	*1.17×*	*1.15×*	*1.22×*

Table 3: **Held-out cell-line generalization (C3).** Each model trains on one cell line and tests on the other eight. Mean ± s.d. across evaluation cell lines. Best in **bold red**, second best in *italic blue*.

Method	Venue	R@1	R@10	R@100	nDCG	APS%
C3 – Single-source transfer: train on one cell line, evaluate on the other eight						
Random	–	$0.0001_{\pm 0.0000}$	$0.0009_{\pm 0.0001}$	$0.0092_{\pm 0.0003}$	$0.0640_{\pm 0.0013}$	$0.06_{\pm 0.00}$
Cancer genes	–	$0.0009_{\pm 0.0010}$	$0.0070_{\pm 0.0070}$	$0.0153_{\pm 0.0071}$	$0.0675_{\pm 0.0039}$	$0.25_{\pm 0.15}$
Cancer targets	–	$0.0034_{\pm 0.0014}$	$0.0324_{\pm 0.0108}$	$0.0855_{\pm 0.0413}$	$0.0916_{\pm 0.0114}$	$0.70_{\pm 0.13}$
Perturbed genes	–	$0.0014_{\pm 0.0002}$	$0.0138_{\pm 0.0016}$	$0.1350_{\pm 0.0099}$	$0.1649_{\pm 0.0039}$	$0.69_{\pm 0.06}$
scGen	NMeth'19	$0.0068_{\pm 0.0066}$	$0.0531_{\pm 0.0654}$	$\mathit{0.3293}_{\pm 0.2517}$	$0.1570_{\pm 0.0349}$	$6.30_{\pm 5.22}$
ChemCPA	NeurIPS'22	$0.0080_{\pm 0.0101}$	$0.0468_{\pm 0.0497}$	$0.3185_{\pm 0.2559}$	$0.1570_{\pm 0.0342}$	$6.19_{\pm 4.90}$
CellOT	NMeth'23	$\mathit{0.0110}_{\pm 0.0178}$	$0.0552_{\pm 0.0687}$	$0.3229_{\pm 0.2547}$	$0.1618_{\pm 0.0446}$	$6.56_{\pm 6.15}$
Biolord	NBT'24	$0.0086_{\pm 0.0128}$	$0.0508_{\pm 0.0645}$	$0.3164_{\pm 0.2566}$	$0.1587_{\pm 0.0397}$	$6.34_{\pm 5.82}$
PDGrapher	NBME'25	$0.0101_{\pm 0.0015}$	$\mathit{0.0719}_{\pm 0.0077}$	$0.2994_{\pm 0.0169}$	$\mathit{0.1862}_{\pm 0.0090}$	$\mathit{11.66}_{\pm 1.63}$
RePert (ours)	–	$\mathbf{0.0279}_{\pm 0.0099}$	$\mathbf{0.1325}_{\pm 0.0270}$	$\mathbf{0.4579}_{\pm 0.0205}$	$\mathbf{0.2614}_{\pm 0.0198}$	$\mathbf{12.59}_{\pm 1.92}$
gain over PDGrapher		*2.75×*	*1.84×*	*1.53×*	*1.40×*	*1.08×*

Zero-Shot Cross-Platform Transfer. C5 and C6 impose a larger shift: the C4 model is applied without target-platform training or calibration to external single-cell screens profiled on platforms different from LINCS L1000. These are Cellarity scRNA-seq, split into cancer lines (537 queries, 54 compounds) and primary cells (457 queries, 64 compounds), and sciPlex3 snRNA-seq (354 queries, 116 compounds). The retrieval pool remains the full 9,597-compound training vocabulary. Table 4 shows that RePert leads baselines on every metric of all three splits. The separation is clearest on Cellarity-Primary, whose cellular setting has no counterpart in LINCS: gains range from 1.39× to 1.67× across the five metrics. On Cellarity-Cancer, the advantage is strongest at recall@1 of 1.40×, while on sciPlex3 the smaller improvements exhibit.

Table 4: **Zero-shot cross-platform transfer (C5, C6).** External single-cell datasets are absent from training. The pool remains 9,597 compounds. Mean ± s.d. over three runs. Best in **bold red**, second best in *italic blue*.

Method	Venue	R@1	R@10	R@100	nDCG	top-1%
C5 – Cellarity cancer						
Random	–	$0.0000_{\pm 0.000}$	$0.0006_{\pm 0.001}$	$0.0137_{\pm 0.002}$	$0.0451_{\pm 0.002}$	$0.0140_{\pm 0.001}$
k-NN	–	$0.0410_{\pm 0.000}$	$0.1490_{\pm 0.000}$	$0.2756_{\pm 0.000}$	$0.1633_{\pm 0.000}$	$0.2281_{\pm 0.000}$
FRoGS	NC'24	*$0.0658_{\pm 0.005}$*	$0.2210_{\pm 0.008}$	$0.3892_{\pm 0.019}$	*$0.2194_{\pm 0.007}$*	$0.3685_{\pm 0.017}$
PRnet	NC'24	$0.0074_{\pm 0.002}$	$0.0236_{\pm 0.004}$	$0.1850_{\pm 0.001}$	$0.0905_{\pm 0.001}$	$0.1499_{\pm 0.000}$
TranSiGen	NC'24	*$0.0658_{\pm 0.005}$*	$0.1633_{\pm 0.004}$	$0.2564_{\pm 0.007}$	$0.1783_{\pm 0.003}$	$0.2150_{\pm 0.003}$
DrugReflector	Science'25	$0.0577_{\pm 0.013}$	*$0.2439_{\pm 0.023}$*	*$0.4078_{\pm 0.008}$*	$0.2177_{\pm 0.011}$	*$0.3855_{\pm 0.006}$*
RePert (ours)	–	**$0.0807_{\pm 0.005}$**	**$0.2471_{\pm 0.014}$**	**$0.4662_{\pm 0.080}$**	**$0.2525_{\pm 0.015}$**	**$0.4496_{\pm 0.087}$**
gain over DrugReflector		*1.40×*	*1.01×*	*1.14×*	*1.16×*	*1.17×*
C5 – Cellarity primary						
Random	–	$0.0000_{\pm 0.000}$	$0.0000_{\pm 0.000}$	$0.0102_{\pm 0.003}$	$0.0467_{\pm 0.001}$	$0.0078_{\pm 0.004}$
k-NN	–	$0.0197_{\pm 0.000}$	$0.0985_{\pm 0.000}$	$0.1904_{\pm 0.000}$	$0.1208_{\pm 0.000}$	$0.1794_{\pm 0.000}$
FRoGS	NC'24	*$0.0416_{\pm 0.004}$*	*$0.1255_{\pm 0.003}$*	*$0.2597_{\pm 0.010}$*	*$0.1674_{\pm 0.002}$*	*$0.2850_{\pm 0.013}$*
PRnet	NC'24	$0.0015_{\pm 0.001}$	$0.0109_{\pm 0.004}$	$0.0999_{\pm 0.023}$	$0.0743_{\pm 0.003}$	$0.0940_{\pm 0.022}$
TranSiGen	NC'24	$0.0372_{\pm 0.004}$	$0.0970_{\pm 0.003}$	$0.1685_{\pm 0.004}$	$0.1277_{\pm 0.003}$	$0.1590_{\pm 0.005}$
DrugReflector	Science'25	$0.0263_{\pm 0.002}$	$0.1028_{\pm 0.013}$	$0.2166_{\pm 0.003}$	$0.1359_{\pm 0.003}$	$0.2274_{\pm 0.012}$
RePert (ours)	–	**$0.0423_{\pm 0.009}$**	**$0.1714_{\pm 0.015}$**	**$0.3268_{\pm 0.063}$**	**$0.1889_{\pm 0.010}$**	**$0.3447_{\pm 0.065}$**
gain over DrugReflector		*1.61×*	*1.67×*	*1.51×*	*1.39×*	*1.52×*
C6 – sciPlex3						
Random	–	$0.0000_{\pm 0.000}$	$0.0019_{\pm 0.003}$	$0.0113_{\pm 0.003}$	$0.0450_{\pm 0.002}$	$0.0105_{\pm 0.002}$
k-NN	–	$0.0028_{\pm 0.000}$	$0.0593_{\pm 0.000}$	$0.1328_{\pm 0.000}$	$0.0956_{\pm 0.000}$	$0.1307_{\pm 0.000}$
FRoGS	NC'24	$0.0132_{\pm 0.004}$	$0.1073_{\pm 0.013}$	$0.2853_{\pm 0.006}$	$0.1464_{\pm 0.005}$	$0.2763_{\pm 0.005}$
PRnet	NC'24	$0.0009_{\pm 0.002}$	$0.0264_{\pm 0.004}$	$0.1299_{\pm 0.006}$	$0.0788_{\pm 0.001}$	$0.1221_{\pm 0.006}$
TranSiGen	NC'24	$0.0151_{\pm 0.002}$	$0.0725_{\pm 0.002}$	$0.1620_{\pm 0.004}$	$0.1074_{\pm 0.001}$	$0.1533_{\pm 0.005}$
DrugReflector	Science'25	*$0.0169_{\pm 0.010}$*	*$0.1243_{\pm 0.014}$*	*$0.3164_{\pm 0.024}$*	*$0.1537_{\pm 0.008}$*	*$0.3147_{\pm 0.022}$*
RePert (ours)	–	**$0.0386_{\pm 0.017}$**	**$0.1685_{\pm 0.042}$**	**$0.3672_{\pm 0.049}$**	**$0.1894_{\pm 0.026}$**	**$0.3616_{\pm 0.048}$**
gain over DrugReflector		*2.28×*	*1.36×*	*1.16×*	*1.23×*	*1.15×*

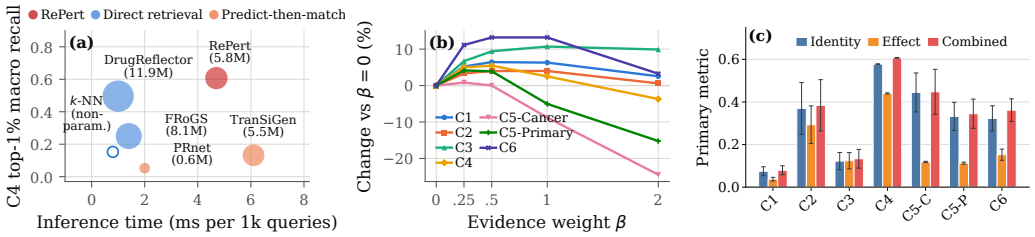


Figure 3: **Efficiency analysis, sensitivity analysis, and identity-effect ablation study. (a)** C4 accuracy versus inference time on one A800. Bubble area denotes learned parameters. **(b)** Relative change in the primary metric as β varies from the identity-score reference ($\beta = 0$). **(c)** Identity-score only ($\beta = 0$), effect-score only, and combined performance from the same trained model.

4.4 Ablation Study

Model Efficiency Analysis. Figure 3(a) relates compound-level accuracy to inference cost on C4. RePert reaches top-1% macro recall 0.607 with 5.8M learned parameters, approximately half the 11.9M parameters of DrugReflector. The effect view adds one encoder pass and one matrix product over identity-only retrieval, yielding 4.7 ms per 1,000 queries. Unlike predict-then-match baselines, it does not need to synthesize a response for all 9,597 candidates before ranking.

Sensitivity Analysis. Figure 3(b) shows that performance is stable for $\beta \in [0.25, 1]$ in domain. Exact task-wise scores and selected values are reported in Table 9. Across tasks, the validation-selected β achieves a primary-metric score within 0.007 of the best score in the tested grid. Overweighting effect evidence is most harmful under the largest platform shift, supporting calibrated combination with identity evidence.

Ablation Study on Components. Table 5 isolates the roles of effect pretraining, effect retrieval branch, and identity-effect grounding. The full model uses all three components. Specifically,

Table 5: **RePert component ablations.** C1 and C4 represent gene- and compound-level retrieval.

Identity retrieval $\mathcal{L}_{\text{ret}}$	Effect pretraining $\mathcal{L}_{\text{pre}}$	Effect retrieval $\mathcal{L}_{ez}$	Identity–effect grounding $\mathcal{L}_{\text{ground}}$	C1 (single-gene) R@10	C1 (single-gene) R@100	C4 (compound) R@10	C4 (compound) R@100
✓	✓	✓	✓	$0.0790_{\pm 0.0210}$	$0.5076_{\pm 0.0688}$	$0.4838_{\pm 0.0055}$	$0.7738_{\pm 0.0031}$
✓	✗	✓	✓	$0.0729_{\pm 0.0219}$	$0.4357_{\pm 0.0705}$	$0.4888_{\pm 0.0005}$	$0.7632_{\pm 0.0009}$
✓	✓	✗	✓	$0.0785_{\pm 0.0211}$	$0.5046_{\pm 0.0654}$	$0.4845_{\pm 0.0005}$	$0.7713_{\pm 0.0014}$
✓	✓	✓	✗	$0.0708_{\pm 0.0204}$	$0.4581_{\pm 0.0665}$	$0.4818_{\pm 0.0011}$	$0.7699_{\pm 0.0005}$
✓	✗	✗	✗	$0.0603_{\pm 0.0221}$	$0.3682_{\pm 0.0757}$	$0.4895_{\pm 0.0011}$	$0.7615_{\pm 0.0010}$

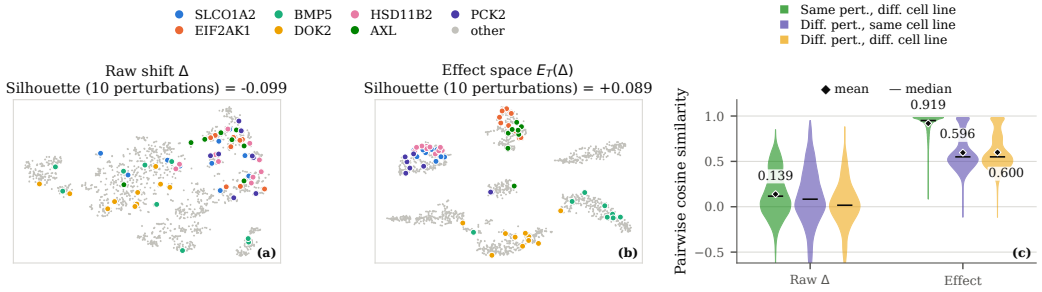


Figure 4: **Analysis of effect pretraining. (a,b)** Raw held-out C1 responses and their learned effect representations. Silhouette scores are computed over all 10 rare perturbations. **(c)** Pairwise cosine similarity in raw and effect spaces for pairs sharing a perturbation, a cell line, or neither, with 20,000 held-out pairs per category.

removing effect pretraining discards the pretrained effect space; removing effect retrieval branch keeps the pretrained effect space only for identity-effect grounding; removing grounding retains both identity and effect retrieval branches but eliminates their representation alignment. The final row is identity-only retrieval, removing all effect-based components. On C1, the full model improves over this identity-only baseline from 0.0603 to 0.0790 at recall@10 and 0.3682 to 0.5076 at recall@100. Grounding contributes most at recall@10, whereas effect pretraining contributes most at recall@100. On C4, removing pretraining or grounding lowers recall@100 from 0.7738 to 0.7632 or 0.7699.

Ablation Study on Identity and Effect Evidence. Figure 3(c) scores the same trained model using the identity view ($\beta = 0$), the effect view, and their combination, with exact values in Appendix Table 10. Across seven splits, the combination matches or exceeds the stronger single view by 0.7% to 12%. The effect score matches the identity score on C3 with 0.1240 versus 0.1224. Although the effect score weakens under cross-platform shift, calibration still improves the combination.

4.5 What the Transferable Effect Encoder Learns

Figure 4 evaluates the learned effect-representation space on held-out responses in C1. Raw responses form dispersed perturbation groups with silhouette -0.099, whereas their effect representations cluster across cell lines and raise the silhouette to $+0.089$. Across 20,000 held-out pairs per comparison category, mean same-perturbation similarity increases from 0.139 to 0.919. After encoding, the mean similarity of 0.596 between different perturbations in the same cell line is nearly identical to the background mean of 0.600 for pairs sharing neither perturbation nor cell line. These results indicate that effect pretraining preserves perturbation signal while suppressing cell-line-specific similarity.

5 Conclusion

We presented RePert, a two-stage framework for generalizable inverse perturbation retrieval. RePert first learns transferable effect representations, and then uses candidate effect prototypes to ground identity embeddings and complement identity matching. Across genetic and chemical tasks, RePert improves in-domain retrieval, generalizes to held-out cell lines, and transfers to external single-cell platforms. Representation analyses and ablations support the contributions of effect pretraining and effect-grounded retrieval. Together, these results show that learning effects before identities provides a practical foundation for generalizable inverse retrieval.

AI USE STATEMENT

In this work, generative AI tools were used to assist with language polishing. No AI-generated output was treated as experimental evidence or used to determine the reported results. The authors reviewed, revised, and verified all AI-assisted material and take full responsibility for the final text, claims, and artifacts in this work.

ETHICS STATEMENT

This study uses only previously published, publicly available datasets and does not collect new human-subject data or personally identifiable information. This study therefore raises no ethical concerns.

REPRODUCIBILITY STATEMENT

The model formulation and training objectives are described in Section 3, and the datasets, evaluation protocols, and metrics are summarized in Section 4.1. Hyperparameters, training configurations, and detailed results are provided in the appendix. Key source codes are available through the `https://anonymous.4open.science/r/RePert-DEC3/`.

REFERENCES

Abhinav K Adduri, Dhruv Gautam, Beatrice Bevilacqua, Mohsen Naghipourfar, Alishba Imran, Rohan Shah, Noam Teyssier, Rishi Verma, Christopher Carpenter, Basak Eraslan, et al. Predicting cellular responses to perturbation across diverse contexts with state. *Cell*, 2025.

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.

Charlotte Bunne, Yusuf Roohani, Yanay Rosen, Ankit Gupta, Xikun Zhang, Marcel Roed, Theo Alexandrov, Mohammed AlQuraishi, Patricia Brennan, Daniel B Burkhardt, et al. How to build the virtual cell with artificial intelligence: Priorities and opportunities. *Cell*, 187(25):7045–7063, 2024.

Hao Chen, Frederick J King, Bin Zhou, Yu Wang, Carter J Canedy, Joel Hayashi, Yang Zhong, Max W Chang, Lars Pache, Julian L Wong, et al. Drug target prediction through deep learning functional representation of gene signatures. *Nature communications*, 15(1):1853, 2024.

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.

Benjamin DeMeo, Charlotte Nesbitt, Samuel A Miller, Daniel B Burkhardt, Inna Lipchina, Doris Fu, Peter Holderrieth, David Kim, Sergey Kolchenko, Artur Szalata, et al. Active learning framework leveraging transcriptomics identifies modulators of disease phenotypes. *Science*, 390(6776): eadi8577, 2025.

Atray Dixit, Oren Parnas, Biyu Li, Jenny Chen, Charles P Fulco, Livnat Jerby-Arnon, Nemanja D Marjanovic, Danielle Dionne, Tyler Burks, Raktima Raychowdhury, et al. Perturb-seq: dissecting molecular circuits with scalable single-cell rna profiling of pooled genetic screens. *cell*, 167(7): 1853–1866, 2016.

John Erol Evangelista, Daniel JB Clarke, Zhuorui Xie, Alexander Lachmann, Minji Jeon, Kerwin Chen, Kathleen M Jagodnik, Sherry L Jenkins, Maxim V Kuleshov, Megan L Wojciechowicz, et al. Sigcom lincs: data and metadata search engine for a million gene expression signatures. *Nucleic acids research*, 50(W1):W697–W709, 2022.

Guadalupe Gonzalez, Xiang Lin, Isuru Herath, Kirill Veselkov, Michael Bronstein, and Marinka Zitnik. Combinatorial prediction of therapeutic perturbations using causally inspired neural networks. *Nature Biomedical Engineering*, pp. 1–18, 2025.

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.

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.

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. *Advances in neural information processing systems*, 33:18661–18673, 2020.

Justin Lamb, Emily D Crawford, David Peck, Joshua W Modell, Irene C Blat, Matthew J Wrobel, Jim Lerner, Jean-Philippe Brunet, Aravind Subramanian, Kenneth N Ross, et al. The connectivity map: using gene-expression signatures to connect small molecules, genes, and disease. *science*, 313(5795):1929–1935, 2006.

Mohammad Lotfollahi, F Alexander Wolf, and Fabian J Theis. scgen predicts single-cell perturbation responses. *Nature methods*, 16(8):715–721, 2019.

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.

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, 2025.

Mario Niepel, Marc Hafner, Qiaonan Duan, Zichen Wang, Evan O Paull, Mirra Chung, Xiaodong Lu, Joshua M Stuart, Todd R Golub, Aravind Subramanian, et al. Common and cell-type specific responses to anti-cancer drugs revealed by high throughput transcript profiling. *Nature communications*, 8(1):1186, 2017.

Stefan Peidli, Tessa D Green, Ciyue Shen, Torsten Gross, Joseph Min, Samuele Garda, Bo Yuan, Linus J Schumacher, Jake P Taylor-King, Debora S Marks, et al. scperturb: harmonized single-cell perturbation data. *Nature Methods*, 21(3):531–540, 2024.

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

Xiaoning Qi, Lianhe Zhao, Chenyu Tian, Yueyue Li, Zhen-Lin Chen, Peipei Huo, Runsheng Chen, Xiaodong Liu, Baoping Wan, Shengyong Yang, et al. Predicting transcriptional responses to novel chemical perturbations using deep generative model for drug discovery. *Nature Communications*, 15(1):9256, 2024.

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, et al. Mapping information-rich genotype-phenotype landscapes with genome-scale perturb-seq. *Cell*, 185(14):2559–2575, 2022.

Jennifer E Rood, Anna Hupalowska, and Aviv Regev. Toward a foundation model of causal cell and tissue biology with a perturbation cell and tissue atlas. *Cell*, 187(17):4520–4545, 2024.

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

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

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, 2017.

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.

Xiaochu Tong, Ning Qu, Xiangtai Kong, Shengkun Ni, Jingyi Zhou, Kun Wang, Lehan Zhang, Yiming Wen, Jiangshan Shi, Sulin Zhang, et al. Deep representation learning of chemical-induced transcriptional profile for phenotype-based drug discovery. *Nature Communications*, 15(1):5378, 2024.

Tianze Wang, Yan Pan, Fusong Ju, Shuxin Zheng, Chang Liu, Yaosen Min, Qun Jiang, Xinwei Liu, Huanhuan Xia, Guoqing Liu, et al. Cellnavi predicts genes directing cellular transitions by learning a gene graph-enhanced cell state manifold. *Nature Cell Biology*, 27(10):1863–1874, 2025.

Douglas Yao, Loic Binan, Jon Bezney, Brooke Simonton, Jahanara Freedman, Chris J Frangieh, Kushal Dey, Kathryn Geiger-Schuller, Basak Eraslan, Alexander Gusev, et al. Scalable genetic screening for regulatory circuits using compressed perturb-seq. *Nature biotechnology*, 42(8): 1282–1295, 2024.

A NOTATION

Table 6 illustrates core notation used in this study.

Table 6: **Core notation.**

Variable	Meaning
Δ	Perturbation response: the expression shift from the matched control to the perturbed profile.
p	Retrieval candidate: an intervention identity in C1 and C4-C6, or an annotated target gene in C2 and C3.
$\mathcal{P}$	Retrieval-candidate pool.
$\mathcal{Y}_i$	Set of ground-truth retrieval candidates for response Δ_i; it may contain multiple annotated target genes in C2 and C3.
q	Query embedding produced by the trainable retrieval encoder f.
z_p	Candidate identity embedding for retrieval candidate p.
e	Effect representation extracted from response Δ by the shared effect encoder: $e = \mathrm{norm}(E_T(\Delta))$.
m_p	Candidate effect prototype obtained by aggregating training responses whose valid-candidate sets contain p.
$M = \{m_p\}_{p\in\mathcal{P}}$	Effect memory containing the candidate effect prototypes.

B IMPLEMENTATION DETAILS AND ADDITIONAL RESULTS

B.1 TWO-STAGE CONFIGURATION

All trained models are implemented in PyTorch and optimized on a single A800. C5 and C6 reuse the C4-trained drug model without target-platform training or calibration. Tables 7 and 8 detail the two stages: transferable effect-representation learning followed by effect-grounded retrieval. The only tuned values are the alignment weight λ and inference-time evidence weight β, both selected exclusively on validation data from the training distribution.

Table 7: **Stage 1: transferable effect-representation pretraining.**

	Gene-level (C1-C3)	Drug-level (trained on C4, evaluated on C4-C6)
Input dimension G	10,716	978
Encoder E_T	MLP $G \to 1024 \to 512 \to 256$	MLP $G \to 1024 \to 512 \to 256$
Normalisation, activation	BatchNorm, SiLU	BatchNorm, SiLU
Dropout	0.3	0.3
Output	ℓ_2-normalised, $d = 256$	ℓ_2-normalised, $d = 256$
Positive pair	same knockout (C1) or target gene (C2, C3)	same compound
Eligible candidates	$n_p \geq 2$	$n_p \geq 2$
Loss	symmetric InfoNCE	symmetric InfoNCE
Temperature τ_p	0.07	0.07
Batch	256 pairs	256 pairs
Optimiser	AdamW	AdamW
Learning rate	10^{-3}	10^{-3}
Weight decay	10^{-4}	10^{-4}
Schedule	cosine	cosine
Epochs	30	30

Table 8: **Stage 2: effect-grounded retrieval.**

	Gene-level (C1-C3)	Drug-level (trained on C4, evaluated on C4-C6)
Input dimension G	10,716	978
Scale head	6 statistics $\to$ MLP $6 \to 64 \to 1$, clip ± 2	6 statistics $\to$ MLP $6 \to 64 \to 1$, clip ± 2
Query encoder f	MLP $G \to 1024 \to 512 \to 256$	MLP $G \to 1024 \to 512 \to 256$
Dropout	0.5	0.5
Identity z_p	$N \times 256$, $\mathcal{N}(0, 0.02^2)$ init	$N \times 256$, $\mathcal{N}(0, 0.02^2)$ init
$\mathbf{V}$	256×256, identity init	256×256, identity init
$\mathbf{W}$	256×256, no bias	256×256, no bias
τ_q	0.07	0.07
τ_e	0.07	0.07
τ_a	0.07	0.07
$\lvert\mathcal{S}\rvert$	1,024	1,024
λ	1.0 (C1, C2); 0.25 (C3)	0.25
μ	0.25	0.25
Batch	256	512
Optimiser	AdamW	AdamW
Learning rate	10^{-3}	10^{-3}
Weight decay	10^{-4}	10^{-4}
Schedule	cosine	cosine
Epochs	30 (last)	50 (best validation)
Precision	fp32	bf16
β grid	$\{0, 0.25, 0.5, 1, 2\}$	$\{0, 0.25, 0.5, 1, 2\}$
Validation / reuse	PDGrapher (C1, C2); random 10% (C3)	C4: random 5%; model and β reused unchanged for C5/C6

B.2 Sensitivity Analysis

The sensitivity analysis of β is shown in Table 9.

Table 9: **Sensitivity to effect weight β.** Primary metric after re-scoring. C1–C4 select β using validation data from their respective training distributions; C5 and C6 inherit the C4 value without external-data calibration. Best in **bold red**, second best in *italic blue*.

Task	$\beta = 0$	$\beta = 0.25$	$\beta = 0.5$	$\beta = 1$	$\beta = 2$	β used
C1	0.0746	0.0785	**0.0794**	*0.0793*	0.0765	1
C2	0.3697	0.3820	*0.3844*	**0.3845**	0.3719	0.5
C3	0.1224	0.1306	0.1339	**0.1355**	*0.1345*	0.25
C4	0.5776	*0.6064*	**0.6091**	0.5921	0.5563	0.25
C5-Cancer	*0.4447*	**0.4484**	*0.4447*	0.4057	0.3360	0.25 (C4)
C5-Primary	0.3321	**0.3459**	*0.3451*	0.3154	0.2816	0.25 (C4)
C6	0.3228	*0.3587*	**0.3654**	**0.3654**	0.3333	0.25 (C4)

B.3 IDENTITY-EFFECT COMPLEMENTARITY

The identity-effect complementarity analysis is shown in Table 10.

Table 10: **Identity-effect complementarity.** The same trained model is scored by either view alone and by their calibrated combination. C1-C3 use recall@10 and C4-C6 use top-1% macro recall.

Task	Identity only ($\beta = 0$)	Effect only	Combined	Gain
C1	*0.0746*$_{\pm 0.0210}$	0.0376$_{\pm 0.0084}$	**0.0791**$_{\pm 0.0211}$	+6.1%
C2	*0.3697*$_{\pm 0.1218}$	0.2931$_{\pm 0.0885}$	**0.3842**$_{\pm 0.1206}$	+3.9%
C3	0.1224$_{\pm 0.0405}$	*0.1240*$_{\pm 0.0384}$	**0.1331**$_{\pm 0.0439}$	+7.3%
C4	*0.5776*$_{\pm 0.0018}$	0.4401$_{\pm 0.0027}$	**0.6071**$_{\pm 0.0017}$	+5.1%
C5-Cancer	*0.4447*$_{\pm 0.0915}$	0.1185$_{\pm 0.0028}$	**0.4476**$_{\pm 0.1057}$	+0.7%
C5-Primary	*0.3321*$_{\pm 0.0664}$	0.1127$_{\pm 0.0049}$	**0.3448**$_{\pm 0.0685}$	+3.8%
C6	*0.3228*$_{\pm 0.0594}$	0.1523$_{\pm 0.0264}$	**0.3616**$_{\pm 0.0533}$	+12.0%

B.4 PER-CELL-LINE EVALUATION

Tables 11–13 break down the gene-level in-domain and held-out-cell-line results by evaluation cell line. PDGrapher values are recomputed from the published Source Data: means over the 5 cross-validation folds for C1 and C2, and over the 8 single-cell-line models that score each held-out evaluation cell line for C3. They are the only baseline values reported per cell line by the original work. Table 14 provides the corresponding C4 comparison against retrained DrugReflector for the 81 LINCS cell lines with at least 50 test signatures. RePert wins on all 81, while both methods are weakest on the same few cell types (PHH, NPC, and VCAP).

Table 11: **C1 (single-gene) per cell line**, recall@10.

	A375	A549	AGS	BICR6	ES2	HT29	MCF7	PC3	U251MG	YAPC
PDGrapher	0.0139	0.0240	0.0080	0.0254	0.0366	0.0099	0.0142	0.0091	0.0285	0.0094
RePert	**0.0806**	**0.0881**	**0.0712**	**0.0777**	**0.1059**	**0.0780**	**0.0664**	**0.0410**	**0.1155**	**0.0661**
gain	5.78×	3.66×	8.92×	3.05×	2.89×	7.89×	4.68×	4.49×	4.06×	7.04×

Table 12: **C2 (multiple-gene) per cell line**, recall@10.

	A375	A549	BT20	HELA	HT29	MCF7	MDAMB231	PC3	VCAP
PDGrapher	0.1029	0.1055	0.2517	0.0988	0.0789	0.0843	0.1961	0.0772	0.1255
RePert	**0.3373**	**0.4046**	**0.6835**	**0.2738**	**0.3120**	**0.3830**	**0.4075**	**0.3167**	**0.3532**
gain	3.28×	3.83×	2.72×	2.77×	3.95×	4.54×	2.08×	4.10×	2.81×

Table 13: **C3 (cross-context, leave-cell-out) per evaluation cell line**, recall@10. Each column is an evaluation cell line, averaged over the eight models trained on each of the other cell lines. The evaluation cell line is never seen in training.

	A375	A549	BT20	HELA	HT29	MCF7	MDAMB231	PC3	VCAP
PDGrapher	0.0708	0.0669	0.0561	0.0689	0.0815	0.0754	0.0711	0.0783	0.0784
RePert	**0.1243**	**0.1431**	**0.1919**	**0.0898**	**0.1387**	**0.1335**	**0.1197**	**0.1235**	**0.1275**
gain	1.76×	2.14×	3.42×	1.30×	1.70×	1.77×	1.68×	1.58×	1.63×

Table 14: **C4 (drug-level) per cell line**, top-1% macro recall computed within each of the 81 LINCS cell lines with at least 50 test signatures (n). DR is the DrugReflector architecture retrained on our split. RePert is the mean over three runs. RePert wins on 81 of 81 cell lines (median gain 1.16×). Best per cell line in **bold**.

Cell line	n	DR	Ours	gain	Cell line	n	DR	Ours	gain	Cell line	n	DR	Ours	gain
A375	12221	0.605	**0.687**	1.14×	HT29	12032	0.601	**0.686**	1.14×	PL21	99	0.797	**0.949**	1.19×
A549	8769	0.498	**0.604**	1.21×	HUES3	175	0.630	**0.969**	1.54×	RKO	101	0.742	**0.908**	1.22×
A673	107	0.867	**0.935**	1.08×	HUH7	223	0.946	**0.960**	1.01×	RMGI	66	0.895	**0.965**	1.08×
AGS	85	0.916	**0.972**	1.06×	HUVEC	534	0.751	**0.962**	1.28×	RMUGS	151	0.769	**0.895**	1.16×
ASC	1807	0.439	**0.525**	1.20×	JHUEM2	90	0.800	**0.931**	1.16×	SKB	1105	0.483	**0.574**	1.19×
ASC.C	525	0.813	**0.994**	1.22×	JURKAT	407	0.510	**0.943**	1.85×	SKBR3	1727	0.752	**0.961**	1.28×
BT20	1702	0.774	**0.969**	1.25×	LNCAP	357	0.935	**0.994**	1.06×	SKL	521	0.848	**0.981**	1.16×
CD34	573	0.784	**0.966**	1.23×	LOVO	105	0.947	**0.988**	1.04×	SKL.C	549	0.838	**0.987**	1.18×
CL34	100	0.797	**0.930**	1.17×	MCF10A	1489	0.793	**0.972**	1.23×	SKLU1	108	0.902	**0.954**	1.06×
CORL23	88	0.783	**0.910**	1.16×	MCF7	18579	0.586	**0.668**	1.14×	SKM1	100	0.907	**0.970**	1.07×
COV644	102	0.667	**0.885**	1.33×	MDAMB231	1617	0.780	**0.971**	1.24×	SKMEL1	87	0.681	**0.838**	1.23×
DV90	85	0.890	**0.981**	1.10×	MDST8	104	0.930	**0.970**	1.04×	SKMEL28	95	0.799	**0.927**	1.16×
EFO27	89	0.854	**1.000**	1.17×	MNEU.E	191	0.543	**0.930**	1.71×	SNGM	76	0.831	**0.975**	1.17×
FIBRNPC	332	0.492	**0.623**	1.27×	NCIH1694	76	0.619	**0.823**	1.33×	SNU1040	66	0.879	**0.946**	1.08×
H1299	84	0.947	**0.982**	1.04×	NCIH1836	88	0.707	**0.878**	1.24×	SNUC4	96	0.871	**0.970**	1.11×
HA1E	9377	0.720	**0.815**	1.13×	NCIH2073	88	0.933	**0.967**	1.04×	SNUC5	86	0.773	**0.935**	1.21×
HCC15	169	0.775	**0.859**	1.11×	NCIH508	83	0.796	**0.901**	1.13×	SW480	104	0.777	**0.855**	1.10×
HCC515	4550	0.619	**0.703**	1.14×	NCIH596	86	0.890	**0.981**	1.10×	SW620	164	0.915	**0.969**	1.06×
HCT116	92	0.831	**0.934**	1.12×	NEU	1645	0.425	**0.540**	1.27×	SW948	173	0.842	**0.928**	1.10×
HEC108	119	0.572	**0.837**	1.46×	NKDBA	284	0.564	**0.728**	1.29×	T3M10	92	0.714	**0.937**	1.31×
HEK293T	361	0.703	**0.946**	1.35×	NOMO1	170	0.732	**0.892**	1.22×	THP1	204	0.860	**0.921**	1.07×
HELA	4400	0.774	**0.920**	1.19×	NPC	2537	0.407	**0.471**	1.16×	TYKNU	106	0.722	**0.890**	1.23×
HEPG2	3569	0.652	**0.724**	1.11×	NPC.CAS9	519	0.857	**0.990**	1.16×	U266	312	0.793	**0.995**	1.26×
HL60	83	0.669	**0.898**	1.34×	NPC.TAK	561	0.828	**0.970**	1.17×	U937	122	0.901	**0.976**	1.08×
HME1	893	0.862	**0.971**	1.13×	OV7	74	0.924	**0.960**	1.04×	VCAP	12170	0.443	**0.524**	1.18×
HS578T	1514	0.770	**0.956**	1.24×	PC3	17427	0.591	**0.672**	1.14×	WSUDLCL2	95	0.786	**0.893**	1.14×
HT115	102	0.882	**0.944**	1.07×	PHH	956	0.346	**0.391**	1.13×	YAPC	4359	0.762	**0.896**	1.18×

Table 15: **C5/C6 zero-shot cross-platform transfer per cell line**, top-1% macro recall.

	C5 Cancer lines						C5 Primary cells			
	A549	A375	PC3	HA1E	HEK293T	HepG2	CD34+	CD8+	PAD	HBEC
# queries	115	114	115	115	39	39	133	90	119	115
RePert	0.5602	0.3426	0.5463	0.5417	0.4737	0.4737	0.2968	0.3864	0.3591	0.3102

C6 sciPlex3	A549	K562	MCF7
# queries	118	118	118
RePert	0.3793	0.3190	0.4612

C Limitations and Future Directions

Interpretability. While RePert provides ranked perturbation candidates through complementary identity and effect evidence, it does not explicitly identify which response genes, pathways, or biological programs contribute most strongly to an individual prediction. Incorporating gene-level attribution, pathway-aware decomposition, or prototype-level explanations could help connect retrieval scores to the underlying biological mechanisms and make the predictions more informative for downstream experimental interpretation.

Transcriptomic scope. Our current evaluation focuses on transcriptional responses. Extending effect representation learning to complementary modalities, such as proteomic, epigenomic, or morphological profiles, would test whether the learned notion of perturbation effect generalizes beyond gene expression. Multimodal perturbation datasets may further enable joint effect representations that capture biological responses not fully reflected at the transcriptomic level.