RELAY: RELATIONAL IN-CONTEXT LEARNING FOR CELLULAR PERTURBATION PREDICTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Cellular responses to perturbations vary across multiple experimental variables, such as perturbation identity, cell type, donor, and dose, defining a multidimensional condition space in which each measured population occupies one coordinate. Existing methods predict unmeasured conditions using explicit representations of these variable values, generalizing unseen values therefore depends on the availability and quality of such representations. We present **RELAY**, a relational in-context model that predicts unmeasured coordinates from measured responses at related coordinates. For each target, RELAY uses measured control–perturbed pairs from conditions that share one or more experimental variables with the target, such as the same perturbation in another cell type. The model receives only which variables are shared, not their values, and uses these demonstrations to predict the target response. We evaluate RELAY on chemical, genetic, and cytokine perturbation datasets under two settings: **unseen combinations**, where test values are individually observed during training but not in their target combinations, and **unseen variable values**, where test values are introduced only through context at inference. RELAY outperforms existing in-context methods on unseen variable values and supervised perturbation-prediction methods on unseen combinations.

1 INTRODUCTION

Accurately predicting cellular responses to perturbations at scale has the potential to reshape target discovery and therapeutic development (Bunne et al., 2024; Rood et al., 2024; Ji et al., 2021; Qi & Chapfuwa, 2026; Dibaeinia et al., 2026). These responses are jointly determined by multiple experimental variables: perturbation identity, such as a compound, cytokine, or genetic intervention; perturbation parameters, such as dose, strength, or duration; and biological context, including cell type, donor, and tissue. These factors define a multidimensional condition space in which each experimentally measured cell population occupies a single coordinate, and the space grows multiplicatively in the number of variables, so only a small fraction of it can ever be measured. Perturbation prediction is therefore fundamentally a problem of predicting responses at unmeasured coordinates from measurements elsewhere in this space.

Existing methods predict at a new coordinate by conditioning on the values of its variables. Perturbations may be represented by learned identifier embeddings, molecular structures, gene embeddings, or knowledge-graph features, while biological covariates are commonly encoded using categorical or learned representations (Lotfollahi et al., 2019; 2023; Roohani et al., 2024; Feng et al., 2026; Guo et al., 2026; Jiang et al., 2026). This paradigm works naturally when the relevant values and representations are available during training, but becomes more restrictive for unseen values. Learned identifier embeddings cannot represent values absent from training, while descriptor-based approaches rely on external priors whose coverage and relevance vary across variable types. Moreover, many experimental variables, such as donor identity or study-specific biological covariates, lack standardized descriptors (Kedzierska et al., 2025; Ahlmann-Eltze et al., 2025; Wu et al., 2026). Thus, generalization to unseen conditions often depends on how well each variable value can be represented.

Measured cell populations provide another way to specify a condition: through demonstrations of how cells actually respond, rather than through a predefined representation of the condition itself. Recent work shows that such in-context prediction is viable in single-cell biology. STACK (Dong

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.

et al., 2026) applies tabular attention over cell sets so that one population specifies a condition applied to another; SCGEN (Lotfollahi et al., 2019) transfers perturbation-induced shifts estimated from measured control and perturbed populations across cell types in a variational latent space; and general-purpose tabular foundation models with no biological pre-training are competitive with specialized architectures on several perturbation benchmarks (Palla et al., 2026; Hollmann et al., 2025; Qu et al., 2025; Zhu et al., 2025). These results suggest that measured responses can serve as informative conditioning signals for perturbation prediction. However, existing approaches largely treat context experiments as generic observations, without explicitly representing how each experiment is related to the target condition. This relation is informative because different context experiments reveal different components of the target coordinate. An experiment sharing the target perturbation shows how that perturbation acts in another biological context, whereas one sharing the target cell type shows how that cellular context responds to another perturbation.

In this paper, we make this relation explicit. For a target coordinate, **RELAY** receives measured control–perturbed pairs from conditions that share one or more variable values with the target, together with their shared-variable relations. Variable values are used only to identify matching demonstrations and are never encoded as model inputs. We implement this formulation with a shared pair encoder for control–perturbed population pairs, relation embeddings that encode which variables each demonstration shares with the target, and target-to-context cross-attention that predicts expression changes in target control cells. The formulation extends naturally to multiple experimental variables by representing each demonstration according to the subset of variables it shares with the target.

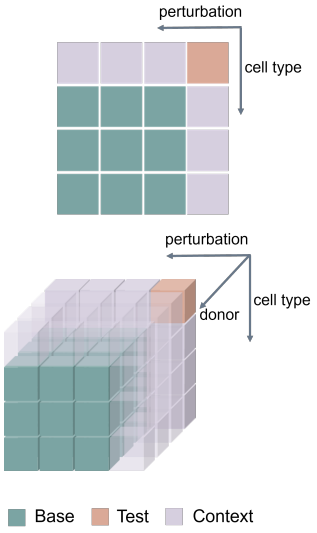

Figure 1: Partition of the condition space into base, context, and test conditions.

We partition the condition space into base, context, and test regions (Figure 1), and evaluate two complementary settings. **Setting 1: unseen combinations.** Training uses both base and context conditions, so the values defining a test condition have been observed individually, but their combination has not. **Setting 2: unseen variable values.** Training uses only base conditions, so every value defining a test condition is unseen during training and is introduced only through context demonstrations at inference, with model parameters fixed. The first tests compositional generalization across previously observed values, whereas the second tests adaptation to entirely new variable values through inference-time evidence, without learning value-specific representations.

Our contributions are as follows.

- A relational formulation of perturbation prediction as in-context inference over a multidimensional condition space, using measured control–perturbed pairs that share variables with the target rather than explicit value representations.
- A unified in-context model that encodes control–perturbed pairs, represents their shared-variable relations to the target, and predicts target responses through cross-attention.
- An evaluation framework separating generalization to unseen combinations from generalization to unseen variable values, showing that RELAY improves perturbation-response prediction over supervised and in-context baselines under their respective settings.

2 Related Work

Cellular perturbation prediction. Existing methods for cellular perturbation prediction span several modeling paradigms. Early approaches model perturbation effects in latent spaces or through compositional representations of perturbations and biological covariates (Lotfollahi et al., 2019; 2023). Other methods incorporate gene-level, molecular, or mechanistic priors to support generalization across perturbations (Roohani et al., 2024; Feng et al., 2026; Guo et al., 2026; Shi et al., 2026). More recent work models perturbation responses at the population level using optimal transport, flow matching, diffusion, or related distributional objectives (Bunne et al., 2023; Adduri et al.,

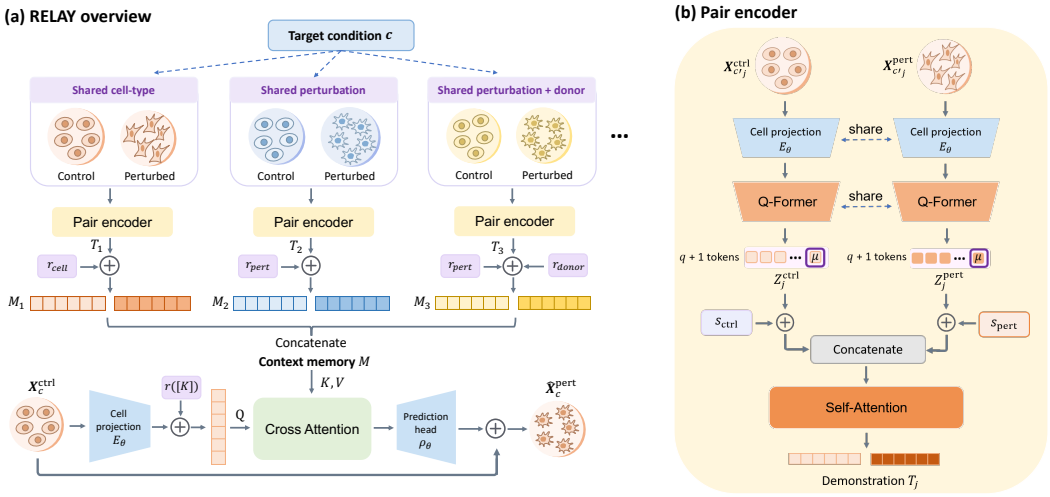

Figure 2: Overview of RELAY. **(a)** Target control cells attend to measured control–perturbed demonstrations to predict the target perturbation response, using shared-variable relations rather than explicit variable-value representations. **(b)** A shared pair encoder summarizes each control and perturbed population and jointly models the pair with self-attention, with state embeddings distinguishing the two population states.

2026; Klein et al., 2025; Yuan et al., 2026; He et al., 2026; Yu et al., 2026; Wang et al., 2026; Chen et al., 2026; Chi et al., 2026; Liu et al., 2026). Despite differences in architecture and training objective, these methods typically predict a target condition by conditioning on explicit representations of its perturbation and biological covariates.

Context-based perturbation prediction. A complementary line of work uses measured cellular responses themselves as conditioning information. scGen transfers perturbation-induced latent shifts estimated from observed control and perturbed populations across cellular contexts (Lotfollahi et al., 2019). STACK (Dong et al., 2026) uses observed cell populations as in-context examples, while MapPFN (Sextro et al., 2026) and PerturbMap (Cui et al., 2026) leverage interventional observations or cross-context response transfer. General-purpose tabular foundation models likewise show that inference from context examples can be competitive on cellular perturbation benchmarks (Hollmann et al., 2025; Qu et al., 2025; Palla et al., 2026). These approaches demonstrate that measured responses can provide informative inference-time evidence. RELAY differs in explicitly representing how each demonstration is related to the prediction target: each control–perturbed population pair is encoded together with the subset of experimental variables it shares with the target, while the variable values themselves are never provided to the model.

3 Method

3.1 Problem Formulation

Problem Setup. We consider cellular perturbation experiments defined over a condition space with K variables, such as perturbation identity, dose, cell type, and donor. A condition is represented as $c = (c_1, \ldots, c_K) \in \mathcal{C}$, where c_k denotes the value of variable k and $\mathcal{C}$ is the space of all such conditions. Only a subset of conditions is experimentally measured. For a measured condition c, let $X_c^{\text{ctrl}}, X_c^{\text{pert}} \in \mathbb{R}^{n \times G}$ denote its control and perturbed cell populations over G genes, where n is the number of cells. Our goal is to predict X_c^{pert} from its matched control population X_c^{ctrl} and measured responses at related conditions.

Experimental Settings. For each variable k, we partition its values into a base set and a held-out set. Among measured conditions, those composed entirely of base values form ($\mathcal{C}_{\text{base}}$), those composed entirely of held-out values form ($\mathcal{C}_{\text{test}}$), and those containing both base and held-out values form ($\mathcal{C}_{\text{ctx}}$) (Figure 1). We consider two settings:

- **Setting 1: unseen combinations.** Training uses $\mathcal{C}_{\text{base}} \cup \mathcal{C}_{\text{ctx}}$: held-out variable values are individually observed during training, but their combinations are not.
- **Setting 2: unseen variable values.** Training uses only $\mathcal{C}_{\text{base}}$: every value composing a test condition is absent from training and is introduced only through context at inference.

Relational Context. Rather than encoding variable values, we specify the target through measured responses at related conditions. For a target condition c and a measured condition c', define the set of shared variables as

$$S(c, c') = \{k \in \{1, \dots, K\} : c_k = c'_k\}. \tag{1}$$

We use conditions satisfying $\varnothing \neq S(c, c') \subsetneq \{1, \dots, K\}$, so that each shares at least one variable value with the target while differing in at least one other variable. A *demonstration* consists of the measured control–perturbed pair at such a condition together with its relation to the target. For target c, we select P demonstration conditions and define *context* of the target as

$$\mathcal{D}(c) = \left\{ \left(X^{\text{ctrl}}_{c'_j}, X^{\text{pert}}_{c'_j}, S_j \right) \right\}_{j=1}^{P}, S_j = S(c, c'_j). \tag{2}$$

The prediction is

$$\widehat{X}^{\text{pert}}_{c} = f_\theta \left(X^{\text{ctrl}}_{c}, \mathcal{D}(c) \right). \tag{3}$$

3.2 RELAY

Figure 2 provides an overview of **RELAY**. Given a target control population and a set of measured control–perturbed demonstrations, the model first encodes each demonstration with a shared pair encoder into a compact representation of its observed perturbation-induced transition (Section 3.2.1). It then incorporates the experimental variables shared between each demonstration and the target to form a relation-aware context memory (Section 3.2.2). Target control cells attend to this memory to integrate information across related conditions and predict perturbation-induced expression changes (Section 3.2.3). All components are trained jointly with population-level supervision, without requiring cell-to-cell correspondence between control and perturbed populations (Section 3.3).

3.2.1 Demonstration Encoding

Each demonstration contains a control–perturbed cell population pair, which we encode with a pair encoder (Figure 2b). To represent multiple demonstrations efficiently, the pair encoder first compresses each cell population into a fixed number of tokens and then jointly encodes the two populations to capture the observed perturbation-induced transition. Specifically, each cell is mapped to a d-dimensional representation by a projection layer $E_\theta : \mathbb{R}^G \to \mathbb{R}^d$. A Q-Former (Li et al., 2023) with q learned queries then summarizes each cell population into q tokens, to which we append a projected population-mean token. For demonstration j, this yields

$$Z^s_j \in \mathbb{R}^{(q+1)\times d}, s \in \{\text{ctrl}, \text{pert}\}. \tag{4}$$

The cell projection and Q-Former are shared between the control and perturbed populations, and the resulting summaries are invariant to cell ordering.

To distinguish the control and perturbed populations, we add learnable state embeddings $\mathbf{s}_{\text{ctrl}}, \mathbf{s}_{\text{pert}} \in \mathbb{R}^d$ to their corresponding tokens. The two sets of tokens are then concatenated and jointly processed by L_p self-attention blocks:

$$T_j = \text{SA}_\theta \left([Z^{\text{ctrl}}_j + \mathbf{1}\mathbf{s}^\top_{\text{ctrl}} \,;\, Z^{\text{pert}}_j + \mathbf{1}\mathbf{s}^\top_{\text{pert}}] \right). \tag{5}$$

Here, $[\cdot\,;\,\cdot]$ denotes concatenation along the token dimension, and $\mathbf{1}$ broadcasts the corresponding state embedding to all tokens. By jointly modeling the control and perturbed populations within each demonstration, the pair encoder output T_j represents its observed perturbation-induced transition. The same pair encoder is shared across all demonstrations.

3.2.2 RELATION EMBEDDING

The demonstration representation T_j captures the observed control–perturbed transition, but does not specify how the demonstration is related to the target condition. We therefore encode which experimental variables are shared between each demonstration and the target. For demonstration j, let S_j denote the set of shared variables. Each variable k is associated with a learnable embedding $r_k \in \mathbb{R}^d$, and the target–demonstration relation is represented as:

$$r(S_j) = \sum_{k \in S_j} r_k. \tag{6}$$

For example, if a demonstration shares perturbation identity and donor with the target, its relation embedding is $r_{\text{pert}} + r_{\text{donor}}$.

We add this relation embedding to all tokens of the corresponding demonstration representation:

$$M_j = T_j + \mathbf{1} r(S_j)^\top. \tag{7}$$

The representations of all the P demonstrations are then concatenated along the token dimension to form the context memory,

$$M = [M_1; \ldots; M_P] \in \mathbb{R}^{2P(q+1) \times d}. \tag{8}$$

3.2.3 CONTEXT-CONDITIONED PREDICTION

Given the context memory M, we use the target control cells as queries to attend to relevant information across demonstrations and predict the target perturbed population. Unlike the demonstration populations, which are compressed into a fixed number of tokens, we retain one query token for each target control cell so that the model directly produces a predicted expression profile for each cell. Specifically, the same projection layer E_θ maps the target control cells into d-dimensional representations. Since the target condition matches itself on all K variables, we add the corresponding relation embedding $r([K])$, yielding

$$H = E_\theta(X_c^{\text{ctrl}}) + \mathbf{1} r([K])^\top \in \mathbb{R}^{n \times d}, \tag{9}$$

where $[K] = \{1, \ldots, K\}$. This embedding distinguishes the target queries from demonstration tokens and represents the target as the reference condition.

The target queries are then updated through L_t cross-attention blocks over the context memory M, allowing each target cell to integrate information from the encoded demonstration transitions and their relations to the target. A prediction head $\rho_\theta : \mathbb{R}^d \to \mathbb{R}^G$ maps the resulting representations to perturbation-induced expression changes. We predict the perturbed expression as a residual from the target control state:

$$\widehat{X}_c^{\text{pert}} = X_c^{\text{ctrl}} + \rho_\theta(\text{CA}_\theta(H, M)) \in \mathbb{R}^{n \times G}. \tag{10}$$

CA_θ denotes the stack of cross-attention blocks, including cross-attention, feed-forward layers, normalization, and residual connections. The target controls provide the baseline cell states, while the context memory informs how these states are expected to change under the target condition.

3.3 TRAINING AND INFERENCE

We train all components jointly using an energy-kernel maximum mean discrepancy loss $\mathcal{L}_{\text{MMD}}$:

$$\min_\theta \; \mathbb{E}_{c, \mathcal{D}(c)} \left[\mathcal{L}_{\text{MMD}} \left(\widehat{X}_c^{\text{pert}}, X_c^{\text{pert}} \right) \right]. \tag{11}$$

The population-level objective compares the predicted and observed distributions without requiring cell-to-cell correspondence. Architecture, hyperparameters, and optimization details are provided in Appendix B.2.

4 EXPERIMENTS AND RESULTS

4.1 EXPERIMENTAL SETUP

Benchmarks. We evaluate four benchmarks spanning chemical, genetic, and cytokine perturbations. **Tahoe-100M** (Zhang et al., 2025) varies compound and cell line, using $5\,\mu$M dose and

Table 1: Comparison of methods for unseen combination prediction (Setting 1).

		Non-learning		Supervised methods					Ours
Data	Metric	Untreated	Mean perturbation	STATE	CellFlow	CPA	scDFM	PerturbDiff	RELAY
Tahoe-100M	MSE ↓	$0.007_{\pm 2e-5}$	$0.011_{\pm 1e-3}$	$0.007_{\pm 7e-6}$	$\underline{0.004}_{\pm 3e-6}$	$0.005_{\pm 5e-6}$	$0.035_{\pm 7e-6}$	$0.009_{\pm 2e-5}$	$\mathbf{0.002}_{\pm 7e-6}$
	MAE ↓	$0.034_{\pm 3e-5}$	$0.049_{\pm 2e-3}$	$0.042_{\pm 1e-5}$	$\underline{0.027}_{\pm 1e-5}$	$0.031_{\pm 9e-6}$	$0.106_{\pm 2e-5}$	$0.050_{\pm 5e-5}$	$\mathbf{0.019}_{\pm 2e-5}$
	PDCorr ↑	$-0.003_{\pm 9e-4}$	$0.259_{\pm 6e-3}$	$0.481_{\pm 2e-4}$	$0.453_{\pm 6e-4}$	$0.294_{\pm 9e-4}$	$0.186_{\pm 2e-4}$	$\underline{0.559}_{\pm 3e-4}$	$\mathbf{0.789}_{\pm 1e-3}$
	PDS ↑	$0.493_{\pm 3e-3}$	$\underline{0.950}_{\pm 7e-3}$	$0.814_{\pm 5e-4}$	$0.816_{\pm 2e-3}$	$0.764_{\pm 2e-3}$	$0.502_{\pm 2e-4}$	$0.925_{\pm 2e-4}$	$\mathbf{0.966}_{\pm 1e-3}$
	REACH ↑	$0.002_{\pm 2e-4}$	$0.038_{\pm 3e-3}$	$\underline{0.324}_{\pm 2e-4}$	$0.099_{\pm 1e-3}$	$0.057_{\pm 4e-4}$	$0.056_{\pm 1e-3}$	$0.303_{\pm 7e-4}$	$\mathbf{0.588}_{\pm 2e-3}$
Replogle–Nadig	MSE ↓	$0.011_{\pm 2e-5}$	$0.018_{\pm 4e-4}$	$0.024_{\pm 1e-3}$	$0.011_{\pm 7e-4}$	$\underline{0.010}_{\pm 1e-5}$	$0.032_{\pm 4e-5}$	$0.011_{\pm 1e-4}$	$\mathbf{0.009}_{\pm 1e-4}$
	MAE ↓	$\underline{0.065}_{\pm 4e-5}$	$0.091_{\pm 7e-4}$	$0.115_{\pm 2e-3}$	$0.069_{\pm 1e-3}$	$\mathbf{0.063}_{\pm 5e-6}$	$0.125_{\pm 7e-5}$	$0.067_{\pm 2e-4}$	$0.066_{\pm 3e-4}$
	PDCorr ↑	$-0.001_{\pm 4e-3}$	$0.210_{\pm 5e-3}$	$\underline{0.230}_{\pm 8e-3}$	$0.193_{\pm 6e-2}$	$0.063_{\pm 3e-3}$	$0.107_{\pm 1e-3}$	$0.158_{\pm 2e-3}$	$\mathbf{0.336}_{\pm 3e-3}$
	PDS ↑	$0.489_{\pm 7e-3}$	$\mathbf{0.709}_{\pm 9e-3}$	$0.536_{\pm 2e-2}$	$0.567_{\pm 2e-2}$	$0.657_{\pm 6e-3}$	$0.497_{\pm 4e-3}$	$0.558_{\pm 2e-3}$	$\underline{0.684}_{\pm 2e-3}$
	REACH ↑	$0.064_{\pm 5e-3}$	$0.291_{\pm 9e-3}$	$\underline{0.304}_{\pm 3e-2}$	$0.246_{\pm 5e-2}$	$0.170_{\pm 3e-3}$	$0.098_{\pm 4e-3}$	$0.295_{\pm 6e-3}$	$\mathbf{0.480}_{\pm 2e-3}$
PBMC-2D	MSE ↓	$0.009_{\pm 5e-5}$	$0.020_{\pm 6e-3}$	$0.016_{\pm 2e-3}$	$\underline{0.007}_{\pm 8e-4}$	$\underline{0.007}_{\pm 9e-5}$	$0.039_{\pm 8e-5}$	$0.017_{\pm 2e-4}$	$\mathbf{0.005}_{\pm 2e-4}$
	MAE ↓	$0.035_{\pm 8e-5}$	$0.061_{\pm 8e-3}$	$0.063_{\pm 2e-3}$	$0.036_{\pm 6e-4}$	$\mathbf{0.032}_{\pm 2e-4}$	$0.100_{\pm 1e-4}$	$0.067_{\pm 4e-4}$	$\underline{0.033}_{\pm 2e-4}$
	PDCorr ↑	$0.006_{\pm 5e-3}$	$0.182_{\pm 3e-2}$	$0.241_{\pm 2e-2}$	$0.275_{\pm 1e-2}$	$\underline{0.317}_{\pm 5e-3}$	$0.040_{\pm 7e-4}$	$0.281_{\pm 4e-3}$	$\mathbf{0.441}_{\pm 1e-2}$
	PDS ↑	$0.495_{\pm 2e-2}$	$0.807_{\pm 3e-2}$	$0.693_{\pm 1e-3}$	$0.682_{\pm 6e-3}$	$0.780_{\pm 4e-3}$	$0.501_{\pm 2e-3}$	$\underline{0.811}_{\pm 1e-2}$	$\mathbf{0.868}_{\pm 3e-2}$
	REACH ↑	$0.099_{\pm 2e-2}$	$0.267_{\pm 1e-2}$	$\mathbf{0.611}_{\pm 8e-3}$	$0.218_{\pm 8e-3}$	$0.326_{\pm 4e-3}$	$0.096_{\pm 4e-3}$	$0.293_{\pm 2e-2}$	$\underline{0.550}_{\pm 1e-2}$
PBMC-3D	MSE ↓	$\underline{0.004}_{\pm 2e-5}$	$0.012_{\pm 3e-3}$	$0.012_{\pm 7e-4}$	$\underline{0.004}_{\pm 2e-4}$	$0.005_{\pm 9e-6}$	$0.048_{\pm 4e-5}$	$0.025_{\pm 8e-5}$	$\mathbf{0.002}_{\pm 3e-5}$
	MAE ↓	$0.031_{\pm 2e-5}$	$0.054_{\pm 4e-3}$	$0.061_{\pm 1e-3}$	$\underline{0.028}_{\pm 8e-4}$	$0.035_{\pm 2e-5}$	$0.126_{\pm 5e-5}$	$0.090_{\pm 1e-4}$	$\mathbf{0.024}_{\pm 1e-4}$
	PDCorr ↑	$-0.005_{\pm 4e-3}$	$0.079_{\pm 2e-2}$	$\underline{0.225}_{\pm 4e-3}$	$0.087_{\pm 2e-2}$	$0.201_{\pm 5e-4}$	$0.089_{\pm 5e-4}$	$0.190_{\pm 1e-3}$	$\mathbf{0.495}_{\pm 5e-3}$
	PDS ↑	$0.498_{\pm 1e-2}$	$\underline{0.714}_{\pm 3e-2}$	$0.587_{\pm 2e-2}$	$0.568_{\pm 8e-3}$	$0.626_{\pm 4e-3}$	$0.502_{\pm 1e-3}$	$0.668_{\pm 5e-3}$	$\mathbf{0.833}_{\pm 5e-3}$
	REACH ↑	$0.017_{\pm 5e-3}$	$0.053_{\pm 2e-2}$	$\mathbf{0.545}_{\pm 7e-3}$	$0.035_{\pm 5e-3}$	$0.054_{\pm 3e-3}$	$0.075_{\pm 1e-3}$	$0.096_{\pm 1e-3}$	$\underline{0.420}_{\pm 1e-2}$

matched DMSO controls. **Replogle–Nadig** (Replogle et al., 2022; Nadig et al., 2024) varies CRISPRi target and cell line, with non-targeting guides as controls. From the Parse cytokine dataset (Oesinghaus et al., 2025), **PBMC-2D** varies cytokine and cell type within Donor9, while **PBMC-3D** additionally varies donor; both use matched PBS controls. Dataset sizes and condition-level cell counts are summarized in Appendix A.1.

Evaluation protocol. We follow the partitions and settings defined in Section 3.1. **Setting 1** trains on base and context conditions to predict unseen combinations, whereas **Setting 2** trains only on base conditions and introduces context at inference without parameter updates. Both settings use the same test conditions and inference context. On Tahoe-100M, base and held-out compounds are additionally separated by Bemis–Murcko scaffold (Bemis & Murcko, 1996). All methods use identical data splits and preprocessing, including library-size normalization to 10^4, log transformation, and a common highly variable gene panel selected using training-visible cells only. Split construction and preprocessing are detailed in Appendices A and B.1, respectively.

Baselines. We evaluate broadly applicable baselines across chemical, genetic, and cytokine perturbation benchmarks. Setting 1 includes five supervised methods: **STATE** (Adduri et al., 2026) models perturbation responses at the population level. **CellFlow** (Klein et al., 2025) learns covariate-conditioned transport maps from control to perturbed populations. **CPA** (Lotfollahi et al., 2023) combines a basal cell state with perturbation and covariate embeddings. We also include two recent baselines: **scDFM** (Yu et al., 2026) generates perturbed cell distributions from control states and perturbation embeddings via conditional flow matching with an MMD objective. **PerturbDiff** (Yuan et al., 2026) models perturbation responses through functional diffusion over cell-population distributions embedded in a Hilbert space.

Setting 2 includes three context-based methods: **STACK** (Dong et al., 2026) uses attention over measured cell populations to predict responses in a target population. **TabPFN** (Hollmann et al., 2025; Palla et al., 2026) uses a pretrained tabular model to infer target responses from context examples. **scGen** (Lotfollahi et al., 2019) is adapted to transfer perturbation-induced latent shifts from context populations to target controls. Both settings also include two non-learning baselines: **Untreated** directly uses target controls as the prediction. **Mean perturbation** adds the mean perturbed-minus-control shift for the same perturbation in other cellular contexts to target controls.

Metrics. We evaluate expression accuracy and perturbation-response recovery, averaging each metric equally over test conditions. **MSE** and **MAE** measure differences between predicted and observed mean expression profiles. **PDCorr** measures Pearson correlation coefficient between predicted and ground-truth relative perturbation effects. **PDS** measures perturbation specificity by

Table 2: Comparison of methods for prediction with unseen variable values (Setting 2).

		Non-learning		In-context methods			Ours
Dataset	Metric	Untreated	Mean perturbation	STACK	scGen	TabPFN	RELAY
Tahoe-100M	MSE ↓	$\underline{0.007}_{\pm 2e-5}$	$0.011_{\pm 1e-3}$	$0.177_{\pm 1e-5}$	$0.022_{\pm 3e-6}$	$0.013_{\pm 1e-4}$	$\mathbf{0.004}_{\pm 7e-6}$
	MAE ↓	$\underline{0.034}_{\pm 3e-5}$	$0.049_{\pm 2e-3}$	$0.157_{\pm 1e-5}$	$0.067_{\pm 1e-5}$	$0.050_{\pm 2e-4}$	$\mathbf{0.028}_{\pm 6e-5}$
	PDCorr ↑	$-0.003_{\pm 9e-4}$	$0.259_{\pm 6e-3}$	$0.087_{\pm 2e-5}$	$\underline{0.268}_{\pm 2e-4}$	$0.254_{\pm 1e-3}$	$\mathbf{0.467}_{\pm 2e-3}$
	PDS ↑	$0.493_{\pm 3e-3}$	$\mathbf{0.950}_{\pm 7e-3}$	$0.500_{\pm 8e-5}$	$0.610_{\pm 4e-4}$	$0.594_{\pm 2e-4}$	$\underline{0.905}_{\pm 1e-3}$
	REACH ↑	$0.002_{\pm 2e-4}$	$0.038_{\pm 3e-3}$	$0.061_{\pm 3e-4}$	$0.052_{\pm 4e-4}$	$\underline{0.057}_{\pm 1e-3}$	$\mathbf{0.194}_{\pm 6e-3}$
Replogle–Nadig	MSE ↓	$\mathbf{0.011}_{\pm 2e-5}$	$\underline{0.018}_{\pm 4e-4}$	$0.410_{\pm 1e-4}$	$0.043_{\pm 6e-4}$	$0.026_{\pm 2e-4}$	$\underline{0.012}_{\pm 7e-5}$
	MAE ↓	$\mathbf{0.065}_{\pm 4e-5}$	$0.091_{\pm 7e-4}$	$0.363_{\pm 2e-4}$	$0.127_{\pm 9e-4}$	$0.100_{\pm 5e-4}$	$\underline{0.075}_{\pm 1e-4}$
	PDCorr ↑	$-0.001_{\pm 4e-3}$	$\underline{0.210}_{\pm 5e-3}$	$0.092_{\pm 1e-4}$	$0.177_{\pm 1e-3}$	$0.114_{\pm 6e-3}$	$\mathbf{0.245}_{\pm 2e-3}$
	PDS ↑	$0.489_{\pm 7e-3}$	$\mathbf{0.709}_{\pm 9e-3}$	$0.508_{\pm 4e-4}$	$0.611_{\pm 6e-3}$	$0.532_{\pm 1e-2}$	$\underline{0.620}_{\pm 3e-3}$
	REACH ↑	$0.064_{\pm 5e-3}$	$\underline{0.291}_{\pm 9e-3}$	$0.065_{\pm 2e-3}$	$0.059_{\pm 5e-3}$	$0.044_{\pm 3e-3}$	$\mathbf{0.363}_{\pm 5e-3}$
PBMC-2D	MSE ↓	$\underline{0.009}_{\pm 5e-5}$	$0.020_{\pm 6e-3}$	$0.236_{\pm 3e-4}$	$0.021_{\pm 3e-4}$	$0.016_{\pm 3e-4}$	$\mathbf{0.007}_{\pm 1e-4}$
	MAE ↓	$\underline{0.035}_{\pm 8e-5}$	$0.061_{\pm 8e-3}$	$0.181_{\pm 1e-4}$	$0.052_{\pm 2e-4}$	$0.046_{\pm 3e-4}$	$\mathbf{0.034}_{\pm 5e-5}$
	PDCorr ↑	$0.006_{\pm 5e-3}$	$0.182_{\pm 3e-2}$	$0.001_{\pm 1e-4}$	$\underline{0.215}_{\pm 1e-2}$	$0.152_{\pm 5e-3}$	$\mathbf{0.222}_{\pm 7e-3}$
	PDS ↑	$0.495_{\pm 2e-2}$	$\mathbf{0.807}_{\pm 3e-2}$	$0.500_{\pm 5e-4}$	$0.715_{\pm 1e-2}$	$0.635_{\pm 2e-2}$	$\underline{0.731}_{\pm 2e-2}$
	REACH ↑	$0.099_{\pm 2e-2}$	$\underline{0.267}_{\pm 1e-2}$	$\underline{0.165}_{\pm 0e0}$	$0.174_{\pm 3e-2}$	$0.119_{\pm 1e-2}$	$\mathbf{0.313}_{\pm 3e-3}$
PBMC-3D	MSE ↓	$\mathbf{0.004}_{\pm 2e-5}$	$0.012_{\pm 3e-3}$	$0.324_{\pm 1e-4}$	$0.078_{\pm 7e-4}$	$0.007_{\pm 7e-5}$	$\underline{0.005}_{\pm 3e-5}$
	MAE ↓	$\mathbf{0.031}_{\pm 2e-5}$	$0.054_{\pm 4e-3}$	$0.240_{\pm 3e-5}$	$0.121_{\pm 5e-4}$	$\underline{0.039}_{\pm 1e-4}$	$\underline{0.039}_{\pm 1e-4}$
	PDCorr ↑	$-0.005_{\pm 4e-3}$	$0.079_{\pm 2e-2}$	$0.061_{\pm 9e-5}$	$0.096_{\pm 2e-3}$	$\underline{0.110}_{\pm 1e-3}$	$\mathbf{0.124}_{\pm 6e-3}$
	PDS ↑	$0.498_{\pm 1e-2}$	$\underline{0.714}_{\pm 3e-2}$	$0.500_{\pm 5e-4}$	$0.523_{\pm 3e-3}$	$0.562_{\pm 1e-2}$	$\mathbf{0.738}_{\pm 2e-2}$
	REACH ↑	$0.017_{\pm 5e-3}$	$\underline{0.053}_{\pm 2e-2}$	$0.035_{\pm 5e-3}$	$0.049_{\pm 3e-4}$	$0.030_{\pm 5e-3}$	$\mathbf{0.064}_{\pm 1e-3}$

ranking the correct measured response against other perturbations sharing the same non-perturbation variables. **REACH** is based on differential expression (DE) analysis and measures the fraction of reference DE genes recovered with reliable response directions, prioritizing genes with stronger predicted statistical evidence (Welch, 1947; Benjamini & Hochberg, 1995). Lower MSE and MAE and higher PDCorr, PDS, and REACH indicate better performance. Formal definitions of the training objective and evaluation metrics are provided in Appendix B.3.

4.2 Main Results

Tables 1 and 2 summarize performance in the two evaluation settings. We report means and standard deviations over three seeds.

Setting 1: unseen combinations. RELAY achieves the highest PDCorr and lowest MSE on all four benchmarks, while also outperforming all supervised baselines in PDS. The largest absolute PDCorr margin over the strongest supervised baseline occurs on PBMC-3D, where RELAY reaches 0.495 compared with 0.225 for STATE, highlighting its effectiveness in a condition space spanning perturbation, cell type, and donor. Beyond these consistent gains, RELAY also achieves the highest REACH on Tahoe-100M and Replogle–Nadig, supporting its ability to recover differential-expression response directions. These results show that relational context supports generalization to unobserved combinations of previously seen variable values, improving both perturbation-effect recovery and mean expression accuracy.

Setting 2: unseen variable values. RELAY outperforms the learned context-based baselines STACK, scGen, and TabPFN in PDCorr, PDS, and REACH on all four benchmarks. The largest PDCorr margin occurs on Tahoe-100M, where RELAY reaches 0.467 compared with 0.268 for scGen. Performance is generally lower than in Setting 1 when the target variable values are entirely absent from training, reflecting the greater difficulty of inference from context alone. Simple baselines nevertheless remain competitive on some metrics: Mean perturbation achieves the highest PDS on the three 2D benchmarks, while Untreated obtains lower MSE and MAE than RELAY on Replogle–Nadig and PBMC-3D despite near-zero PDCorr. This contrast highlights that low absolute expression error does not necessarily imply accurate recovery of perturbation effects.

4.3 Ablation and Model Analysis

We conduct the following analyses on PBMC-3D under Setting 1, except where noted otherwise.

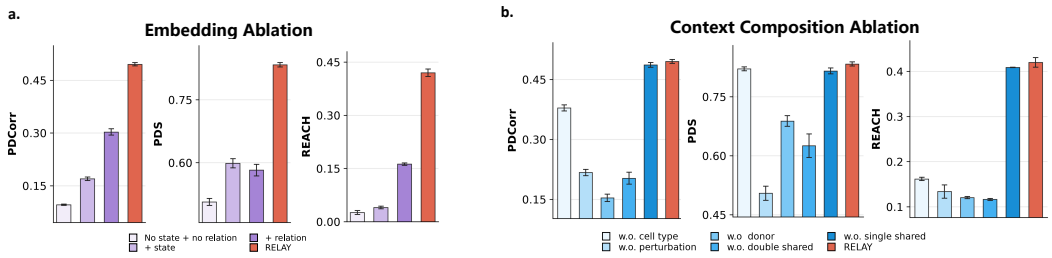

Figure 3: Effects of model components and context composition. (a) Removal of state and relation embeddings. (b) Removal of context groups defined by the variables shared with the target. Bars show means; error bars indicate standard deviations.

State and relation embeddings. Combining both embeddings gives the highest mean scores on all three metrics (Figure 3a): PDCorr, PDS, and REACH reach 0.495, 0.833, and 0.420, respectively. Removing relation embedding reduces these values to 0.170, 0.599, and 0.040, while removing state embedding gives 0.303, 0.582, and 0.162. Removing both reduces performance further to 0.096, 0.507, and 0.026. These results suggest that state and relation embeddings capture complementary aspects of the context, with relation embeddings contributing more strongly to PDCorr and REACH.

Context composition. Context groups contribute differently across metrics (Figure 3b). Removing perturbation-matching groups brings PDS close to chance (0.505), whereas removing donor-matching groups causes the largest PDCorr drop (to 0.154). Removing double-shared groups is more damaging than removing single-shared groups: retaining only single-shared groups gives PDCorr 0.203, PDS 0.626, and REACH 0.116, whereas retaining only double-shared groups gives 0.486, 0.816, and 0.409. These results show that both the identity and the number of target-matching variables determine context utility, while the retained performance with missing groups suggests applicability when related measurements are sparse.

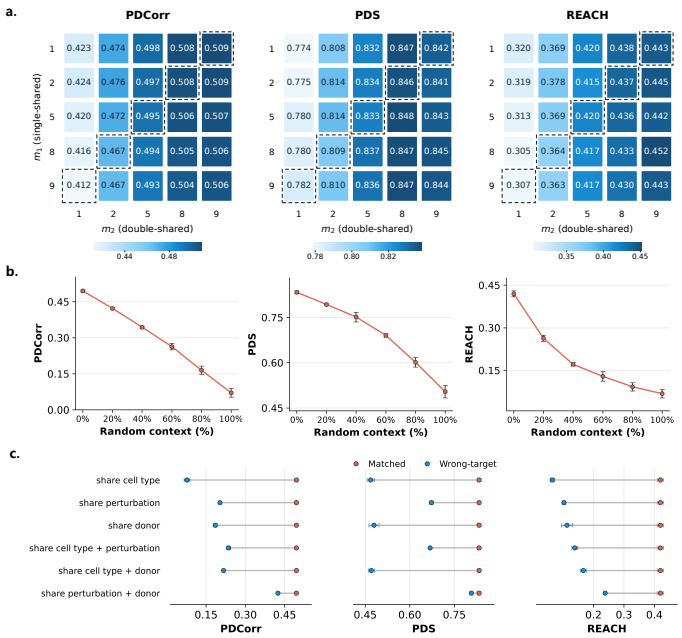

Figure 4: Sensitivity to context quantity and relevance. (a) Varying the numbers of demonstrations per single-shared group (m_1) and double-shared group (m_2). (b) Random replacement of context pairs. (c) Replacement with the complete context of another target sharing the listed variables.

Context size. Here, m_1 and m_2 denote the numbers of demonstrations sampled per context group sharing exactly one and two target variables, respectively. The results (Figure 4a) show that performance is driven primarily by m_2, whereas changing m_1 produces only small differences. At fixed m_1, increasing m_2 from 1 to 5 raises PDCorr, PDS, and REACH by roughly 0.075, 0.053, and 0.107, respectively; increasing m_2 from 5 to 9 adds only about 0.012, 0.010, and 0.022. Under comparable context budgets, allocating more examples to double-shared groups is therefore preferable, although the gains become small once $m_2 \geq 5$.

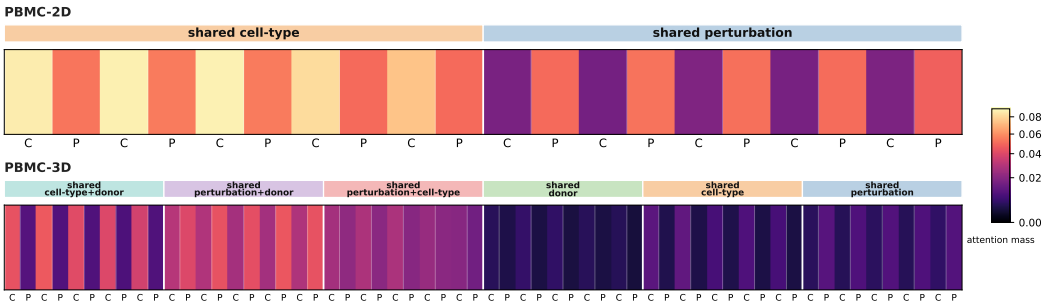

Figure 5: Cross-attention to context in PBMC-2D (top) and PBMC-3D (bottom), averaged over 256 samples. C/P denote control/perturbed tokens; group labels indicate target-shared variables. Colors show attention mass summed over each population's compressed tokens.

Random context replacement. To test whether the model relies on the correspondence between the provided context and the target condition, we replace demonstrations at inference while keeping the target controls, context size, and group sizes unchanged (Figure 4b). Random replacement preserves each control–perturbed pair but leaves the original relation embeddings unchanged, disrupting the correspondence between demonstrations and the target. All three metrics decrease monotonically with the replacement fraction: PDCorr falls from 0.495 to 0.071, PDS from 0.833 to 0.504, and REACH from 0.420 to 0.048. These results confirm that the model is sensitive to the correspondence between demonstrations and the target condition.

Borrowed context replacement. Borrowing the complete context of another target also lowers all three metrics for every tested overlap pattern. Sharing both perturbation and donor best preserves all three metrics among these replacements, reaching PDCorr 0.424, PDS 0.807, and REACH 0.238, but still does not recover matched-context performance (Figure 4c). Together with random replacement, these results show that RELAY retains useful predictive signal under partially noisy or mismatched context, while accurate target-specific demonstration selection remains important.

Attention to context. We examine which context populations receive more attention from target control cells, with each context group containing 5 demonstrations (Figure 5). In PBMC-2D, attention favors control-origin tokens in demonstrations sharing the target cell type, but perturbed-origin tokens in those sharing the target perturbation. PBMC-3D shows a similar pattern: demonstrations sharing both cell type and donor receive more attention on control-origin tokens, whereas those sharing perturbation alone or perturbation and donor receive more attention on perturbed-origin tokens. Thus, the relative attention to the two populations within a pair varies with the variables shared between the demonstration and the target. In PBMC-3D, two-variable matches also receive more attention than single-variable matches despite equal group sizes, consistent with the larger performance drops observed when two-variable matches are removed. These observations describe how attention is distributed, rather than establish which demonstrations cause better predictions. Moreover, control- and perturbed-origin tokens are not independent representations of the two states, as pair self-attention allows them to exchange information.

5 Discussion and Conclusion

RELAY predicts cellular perturbation responses from measured control–perturbed demonstrations, using their shared-variable relations to the target rather than explicit representations of variable values. This formulation supports prediction of unseen combinations and adaptation to unseen values through inference-time context. Pair self-attention and the energy-distance MMD both operate on cell populations, and the loss compares full cell sets rather than means, discouraging collapse to an averaged response. Broader evaluation is also constrained by the scarcity of high-quality perturbation datasets with sufficient cells per condition and dense coverage of multidimensional spaces that include dose and time. Such datasets would enable stronger tests of relational generalization beyond the variables studied here and support extensions from exact-match relations to graded relations over continuous experimental variables.

Reproducibility Statement

The paper specifies the model architecture, training objective, data preprocessing, evaluation protocol, and the two generalization settings. The public source datasets are identified in the Experimental Setup (Zhang et al., 2025; Replogle et al., 2022; Nadig et al., 2024; Oesinghaus et al., 2025). The appendix describes the condition-space partitions, base/context/test subsets, entropy-gap statistics, normalization and HVG selection, principal hyperparameters, ablations, and evaluation metrics. All reported aggregate results use the stated split definitions, and the same preprocessing and metric conventions across methods. The final release will provide the source code, fixed data split manifests, model checkpoints, training configurations, and evaluation scripts in a public repository.

Ethics Statement

This study relies solely on publicly accessible single-cell gene expression datasets, with all data used in compliance with the applicable data use agreements. No directly identifiable personal information was used in the analyses. The methods presented here are designed to support computational modeling and hypothesis generation in biological research. Although these methods may eventually contribute to drug discovery or therapeutic design, they are not intended for direct clinical application. We advocate the responsible and transparent use of generative models in biological research.

AI Use Statement

In this work, we used generative AI tools to assist with software implementation, literature search and summarization, figure preparation, and manuscript editing. They were also used to provide feedback on experimental design, refine hypotheses, and assist with the interpretation and presentation of results. The core research questions, methodological ideas, model design, experiments, and scientific conclusions were developed by the authors. All AI-assisted outputs were reviewed and verified by the authors, who take full responsibility for the final content of this work.

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*, 2026.

Constantin Ahlmann-Eltze, Wolfgang Huber, and Simon Anders. Deep-learning-based gene perturbation effect prediction does not yet outperform simple linear baselines. *Nature Methods*, 22(8): 1657–1661, 2025.

Guy W Bemis and Mark A Murcko. The properties of known drugs. 1. molecular frameworks. *Journal of medicinal chemistry*, 39(15):2887–2893, 1996.

Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)*, 57(1):289–300, 1995.

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.

Shuizhou Chen, Lang Yu, Kedu Jin, Songming Zhang, Hao Wu, Wenxuan Huang, Sheng Xu, Quan Qian, Qin Chen, Lei Bai, et al. Scale: Scalable conditional atlas-level endpoint transport for virtual cell perturbation prediction. *bioRxiv*, pp. 2026–03, 2026.

Changxi Chi, Yufei Huang, Jun Xia, Jiangbin Zheng, Yunfan Liu, Zelin Zang, and Stan Z Li. Departures: Distributional transport for single-cell perturbation prediction with neural schrödinger bridges. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 20544–20552, 2026.

Panpan Cui, Yiqi Liu, and Wenhao Sun. Perturbmap: Cross-context transfer of single-cell perturbation responses. *arXiv preprint arXiv:2607.28090*, 2026.

Payam Dibaeinia, Sudarshan Babu, Mei Knudson, Ali ElSheikh, Yibo Wen, Han Liu, Jason Perera, and Aly A Khan. Virtual cells need context, not just scale. *bioRxiv*, 2026.

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.

Jinghao Feng, Ziheng Zhao, Xiaoman Zhang, Mingfei Liu, Jingyi Chen, Xingran Quan, Boyang Fu, Jian Zhang, Yanfeng Wang, Ya Zhang, et al. A knowledge-driven framework for predicting single-cell responses for unprofiled drugs. *Nature Machine Intelligence*, pp. 1–14, 2026.

Yue Guo, Hao Zhang, Haitao Hu, Jialu Wu, Ji Cao, Chang-Yu Hsieh, and Bo Yang. Modelling drug-induced cellular perturbation responses with a biologically informed dual-branch transformer. *Nature Machine Intelligence*, 8(1):96–112, 2026.

Siyu He, Yuefei Zhu, Daniel Naveed Tavakol, Haotian Ye, Yeh-Hsing Lao, Zixian Zhu, Cong Xu, Shradha Chauhan, Guy Garty, Raju Tomer, et al. Squidiff: predicting cellular development and responses to perturbations using a diffusion model. *Nature methods*, 23(1):65–77, 2026.

Noah Hollmann, Samuel Müller, Lennart Purucker, Arjun Krishnakumar, Max Körfer, Shi Bin Hoo, Robin Tibor Schirrmeister, and Frank Hutter. Accurate predictions on small data with a tabular foundation model. *Nature*, 637(8045):319–326, 2025.

Yuge Ji, Mohammad Lotfollahi, F Alexander Wolf, and Fabian J Theis. Machine learning for perturbational single-cell omics. *Cell Systems*, 12(6):522–537, 2021.

Shiyu Jiang, Zhaoyu Fang, Yujie Zhang, Xuting Zhang, Jérémie Kalfon, Weixu Wang, Xi Fu, Aakash Patel, Syed Rizvi, Yuancheng Ryan Lu, et al. The landscape of single-cell foundation models: Design principles, applications, and open challenges. 2026.

Kasia Z Kedzierska, Lorin Crawford, Ava P Amini, and Alex X Lu. Zero-shot evaluation reveals limitations of single-cell foundation models. *Genome Biology*, 26(1):101, 2025.

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.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pp. 19730–19742. PmLR, 2023.

Jiawen Liu, Xuechenxiao Cao, Yutong Li, Bing Liu, Jiaming Liang, Tinghe Zhang, Xiaoqi Sheng, and Hongmin Cai. Demixpert: Decomposed response modeling with gaussian mixtures for ood single-cell perturbation prediction. *arXiv preprint arXiv:2608.23114*, 2026.

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, Steven A McCarroll, Jonathan S Weissman, Elise B Robinson, and Luke J O'Connor. Transcriptome-wide characterization of genetic perturbations. *BioRxiv*, 2024.

Lukas Oesinghaus, Sören Becker, Larsen Vornholz, Efthymia Papalexi, Joey Pangallo, Amir Ali Moinfar, Jenni Liu, Alyssa La Fleur, Maiia Shulman, Simone Marrujo, et al. A single-cell cytokine dictionary of human peripheral blood. *BioRXiv*, 2025.

Giovanni Palla, Alexander Hillsley, Yang-Joon Kim, and Loic A Royer. Tabular foundation models are competitive cellular perturbation predictors across biological scales. *bioRxiv*, pp. 2026–06, 2026.

Shi-ang Qi and Paidamoyo Chapfuwa. Mechanisms matter: Transportability of cellular perturbation effects. *bioRxiv*, pp. 2026–05, 2026.

Jingang Qu, David Holzmüller, Gaël Varoquaux, and Marine Le Morvan. TabICL: A tabular foundation model for in-context learning on large data. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 50817–50847. PMLR, 2025. URL https://proceedings.mlr.press/v267/qu25d.html.

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.

Marvin Sextro, Weronika Kłos, and Gabriel Dernbach. Mappfn: Learning causal perturbation maps in context. *arXiv preprint arXiv:2601.21092*, 2026.

Peiting Shi, Ningfeng Que, Xianzhe Huang, Xiaofei Wang, and Jianzhong Jeff Xi. Statexdiff: Cell state-contextualized multimodal diffusion for single-cell perturbation prediction. *arXiv preprint arXiv:2605.16104*, 2026.

Handong Wang, Jiaxin Qi, Haochen Feng, and Baisheng Lai. Poppert: Population-level joint-distribution modeling for single-cell perturbation prediction. *arXiv preprint arXiv:2609.01357*, 2026.

Bernard L Welch. The generalization of 'student's'problem when several different population varlances are involved. *Biometrika*, 34(1-2):28–35, 1947.

Yan Wu, Esther Wershof, Sebastian Schmon, Marcel Nassar, Błażej Osiński, Ridvan Eksi, Zichao Yan, Rory Stark, Kun Zhang, and Thore Graepel. Perturbench: Benchmarking machine learning models for cellular perturbation analysis. *Advances in Neural Information Processing Systems*, 38, 2026.

Chenglei Yu, Chuanrui Wang, Bangyan Liao, and Tailin Wu. scdfm: Distributional flow matching model for robust single-cell perturbation prediction. *arXiv preprint arXiv:2602.07103*, 2026.

Xinyu Yuan, Xixian Liu, Ya Shi Zhang, Zuobai Zhang, Hongyu Guo, and Jian Tang. Perturbdiff: Functional diffusion for single-cell perturbation modeling. *arXiv preprint arXiv:2602.19685*, 2026.

Jesse Zhang, Airol A Ubas, Richard De Borja, Valentine Svensson, Nicole Thomas, Neha Thakar, Ian Lai, Aidan Winters, Umair Khan, Matthew G Jones, et al. Tahoe-100m: A giga-scale single-cell perturbation atlas for context-dependent gene function and cellular modeling. *BioRxiv*, pp. 2025–02, 2025.

Hongxu Zhu, Amir Asiaee, Leila Azinfar, Jun Li, Han Liang, Ehsan Irajizad, Kim-Anh Do, and James P Long. Auc-pr is a more informative metric for assessing the biological relevance of in silico cellular perturbation prediction models. *bioRxiv*, 2025.