BASIL: BATCH-AWARE SINGLE-CELL PERTURBATION RESPONSE MODELING WITH CONDITIONAL DISTRIBUTION ALIGNMENT

Anonymous authors
Paper under double-blind review

ABSTRACT

Predicting transcriptional responses to unseen genetic perturbations can reduce the experimental search space in functional genomics. However, single-cell perturbation screens are complicated by experimental batch variation and by destructive sequencing, which provides no one-to-one correspondence between control and perturbed cells. We introduce BASIL, a batch-aware framework centered on *batch-conditioned dual-resolution learning*. For each perturbation–batch condition, same-batch control cells define an explicit pseudo-bulk response target, while the same deterministic state-conditional response operator transforms individual control cells and is regularized against the corresponding cell-unpaired perturbed population. Conditional maximum mean discrepancy (MMD) provides this population-level regularization without assuming cell-level pairs. To support out-of-distribution prediction, BASIL combines GenePT features with multi-relational biological priors to construct perturbation representations that remain available for unseen genes and combinations. Across three genome-scale CRISPRi datasets and the Norman CRISPRa benchmark, BASIL achieves strong performance against representative perturbation-prediction baselines, including challenging unseen-condition and gene-holdout combinatorial settings. Ablations further support the roles of condition-level pseudo-bulk supervision, batch-aware referencing, GenePT-based perturbation features, and conditional population regularization. Overall, BASIL provides a practical direct predictor for robust genetic perturbation response modeling under batch variation and cell-unpaired observations. Code is available at `https://anonymous.4open.science/r/BASIL-main-624C`.

1 INTRODUCTION

Single-cell perturbation screens provide causal measurements of how genetic interventions reshape cellular states and have become essential for studying gene function, regulatory programs, and genetic interactions (Norman et al., 2019; Replogle et al., 2022; Nadig et al., 2025). However, the space of possible interventions grows far beyond experimental throughput when cell type, genetic background, dosage, and perturbation combinations are considered jointly (Lotfollahi et al., 2023; Wenkel et al., 2026). Accurate *in silico* perturbation prediction can therefore prioritize informative experiments, extend perturbation atlases to unexplored conditions, and support predictive modeling of cellular responses (Li et al., 2025; Yu et al., 2026; Adduri et al., 2026).

Recent approaches explore complementary directions: CPA models compositional intervention effects, GEARS incorporates gene relationships for unseen perturbations, scGPT and scFoundation provide pretrained representations, X-Pert studies perturbation–cell interactions, and scDFM aligns control and perturbed distributions without cell correspondence (Lotfollahi et al., 2023; Roohani et al., 2024; Cui et al., 2024; Hao et al., 2024; Li et al., 2025; Yu et al., 2026). Despite these advances, two challenges remain in experimental single-cell perturbation screens.

First, perturbation effects should be measured relative to the appropriate control state, as batch variation can be confounded with biological responses (Wenkel et al., 2026). Second, destructive sequencing prevents one-to-one control–perturbed cell correspondence (Bunne et al., 2023; Yu et al., 2026). Thus, populations can be matched by condition but remain cell-unpaired. Mean-response

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.

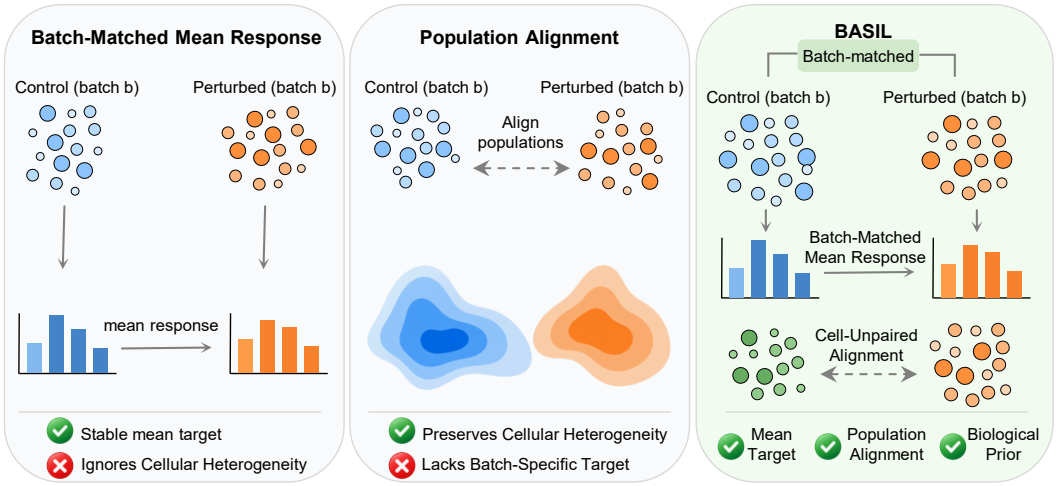


Figure 1: Comparison of supervision paradigms. Mean-response supervision provides a stable target but loses cellular heterogeneity, while population alignment preserves distributions but lacks a batch-specific response reference. BASIL combines batch-matched mean-response learning with cell-unpaired population alignment and incorporates biological priors for unseen perturbations.

regression provides a stable target but compresses heterogeneity, whereas population objectives preserve variation by directly comparing predicted and observed cell distributions without artificial cell pairing (Bunne et al., 2023; Yu et al., 2026), but they require a reliable reference state to determine the perturbation-induced change. In addition, predicting an unseen perturbation requires informative representations beyond perturbation identifiers, because a gene without observed perturbation-response examples cannot be represented solely through learned response statistics. These observations motivate a framework that combines reliable batch-conditioned response supervision, cell-unpaired population alignment, and biologically informed perturbation representations for unseen interventions.

We introduce BASIL, a batch-aware single-cell perturbation response framework built around *batch-conditioned dual-resolution learning*. As illustrated in Figure 1, each perturbation–batch condition supplies two complementary supervision signals: an explicit batch-matched pseudo-bulk target anchors the condition-level perturbation shift, while the corresponding cell-unpaired control and perturbed populations provide condition-local population regularization. Crucially, both resolutions use the same deterministic state-conditional response operator. This couples a stable mean-response target to single-cell population structure without constructing synthetic cell pairs or optimizing a separate generator.

To improve generalization beyond observed perturbations, BASIL initializes perturbations using GenePT functional representations (Chen & Zou, 2023) and enriches them through relation-aware message passing over complementary biological knowledge graphs. GenePT provides semantic information about gene functions, while biological relations provide regulatory, functional, and protein-interaction context (Li et al., 2026; Aleksander et al., 2023; Szklarczyk et al., 2023; Wenkel et al., 2026). This design enables perturbation representations that remain informative for genes absent from perturbation-response training and can support multi-gene intervention modeling.

Our contributions are summarized as follows:

- **Batch-conditioned dual-resolution learning under one shared response operator.** We couple an explicit batch-matched pseudo-bulk response target with condition-local, cell-unpaired population regularization, and apply the same deterministic state-conditional response operator at both resolutions. This jointly anchors the perturbation-induced mean shift and constrains single-cell responses without artificial cell pairing or a separate generator.
- **Biologically informed perturbation representation for generalization.** We integrate GenePT semantics with relation-aware biological graph encoding to construct perturbation representations that do not rely only on observed perturbation identities and can support unseen interventions.

- **Evaluation across single-gene and combinatorial OOD settings.** We evaluate BASIL on genome-scale CRISPR perturbation benchmarks and increasingly stringent combinatorial gene-holdout settings, with controlled ablations of dual-resolution supervision, batch matching, conditional MMD, and GenePT features.

2 Related Work

Single-cell perturbation response prediction. Single-cell perturbation prediction aims to infer transcriptomic responses from cellular states and perturbation specifications, with out-of-distribution generalization as a central challenge. CPA models compositional effects, GEARS uses gene relationships, scGPT/scFoundation provide pretrained representations, X-Pert studies perturbation–cell interactions, and TxPert incorporates batch-aware referencing and biological graphs (Lotfollahi et al., 2023; Roohani et al., 2024; Cui et al., 2024; Hao et al., 2024; Li et al., 2025; Wenkel et al., 2026). BASIL differs by coupling an explicit batch-matched condition-level response target with cell-unpaired population regularization through the same deterministic response operator.

Distribution-aware and transport-based perturbation modeling. Distribution-aware methods address unpaired populations using optimal transport, distribution matching, and flow models. CellOT and scDFM learn population transitions with transport objectives (Bunne et al., 2023; Yu et al., 2026), while MMD provides a non-parametric distribution criterion (Gretton et al., 2012). Rather than using population alignment as a standalone transport objective, BASIL ties condition-local MMD to an explicit batch-conditioned pseudo-bulk target through the same deterministic response operator.

Foundation representations and biological priors. Large single-cell models such as scGPT and scFoundation learn reusable representations from expression corpora (Cui et al., 2024; Hao et al., 2024). GenePT derives gene representations from textual functional descriptions and is naturally available for genes without perturbation-response training examples (Chen & Zou, 2023). Biological graphs provide complementary structure: GEARS uses gene relationships, while multi-relation encoders can preserve distinct regulatory and functional signals (Roohani et al., 2024; Wenkel et al., 2026). BASIL uses GenePT as semantic initialization and refines it with TF-regulatory, GO-similarity, and STRING-interaction messages (Li et al., 2026; Aleksander et al., 2023; Szklarczyk et al., 2023; Wenkel et al., 2026) before coupling the resulting perturbation representation to batch-conditioned mean and population objectives.

3 Method

3.1 Problem Setup

Let $x \in \mathbb{R}^G$ denote the normalized expression profile of a control cell, where G is the number of modeled genes. Let $y \in \mathbb{R}^G$ denote the expression profile of a perturbed cell, p denote a genetic perturbation (single-gene or multi-gene intervention), and b denote the experimental batch. For each perturbation–batch condition (p, b), the experiment provides two cell populations:

$$\mathcal{C}_b = \{x_i\}_{i=1}^{N_b}, \qquad \mathcal{Y}_{p,b} = \{y_j\}_{j=1}^{M_{p,b}}. \tag{1}$$

Here, $\mathcal{C}_b$ and $\mathcal{Y}_{p,b}$ are *batch-matched* because they are collected under the same experimental batch, but they are *cell-unpaired* because destructive sequencing does not provide a one-to-one correspondence between individual control and perturbed cells.

We summarize each population by its pseudo-bulk mean:

$$\bar{x}_b = \frac{1}{N_b} \sum_{i=1}^{N_b} x_i, \qquad \bar{y}_{p,b} = \frac{1}{M_{p,b}} \sum_{j=1}^{M_{p,b}} y_j. \tag{2}$$

Throughout this paper, we use "batch-matched" exclusively to describe consistency at the experimental-condition level, and use "cell-unpaired" to describe the absence of one-to-one correspondence between individual control and perturbed cells. These two concepts refer to different levels of the data structure and are not contradictory.

The observed batch-conditioned perturbation response is defined as:

$$\Delta_{p,b} = \bar{y}_{p,b} - \bar{x}_b. \tag{3}$$

Given a control state x and perturbation representation p, BASIL predicts an expression change vector $d_\theta(x, p)$ and obtains the final expression profile:

$$f_\theta(x, p) = \mathrm{clamp}(x + d_\theta(x, p), 0, \infty). \tag{4}$$

The mean-response branch supervises the condition-level response $d_\theta(\bar{x}_b, p)$, while the cell-unpaired population alignment objective evaluates the generated cells $f_\theta(x_i, p)$ against the observed perturbed cell population.

3.2 Model Overview

As illustrated in Figure 2, BASIL consists of three coupled modules.

(1) Batch-aware condition construction. For each perturbation–batch pair (p, b), BASIL identifies control cells from the same batch and constructs a pseudo-bulk reference $\bar{x}_b$. Meanwhile, the original control and perturbed single-cell populations are retained for distribution alignment. This design explicitly separates two supervision levels: the pseudo-bulk response provides a stable target for biological perturbation effects, while the cell-level populations preserve response heterogeneity without introducing artificial cell pairing.

(2) Biological perturbation prior encoder. Each perturbation is represented by GenePT features and enriched through message passing over multiple biological relation graphs. This module provides semantic and biological information for perturbations that may not have been experimentally observed during training.

(3) State-conditional response network. A shared response function receives the basal control state and perturbation representation, and predicts perturbation-induced responses for both batch-matched control pseudo-bulk states and individual batch-matched control cells. The shared network connects mean-response prediction and population-level alignment, avoiding a separate generator that is optimized independently from the supervised response target.

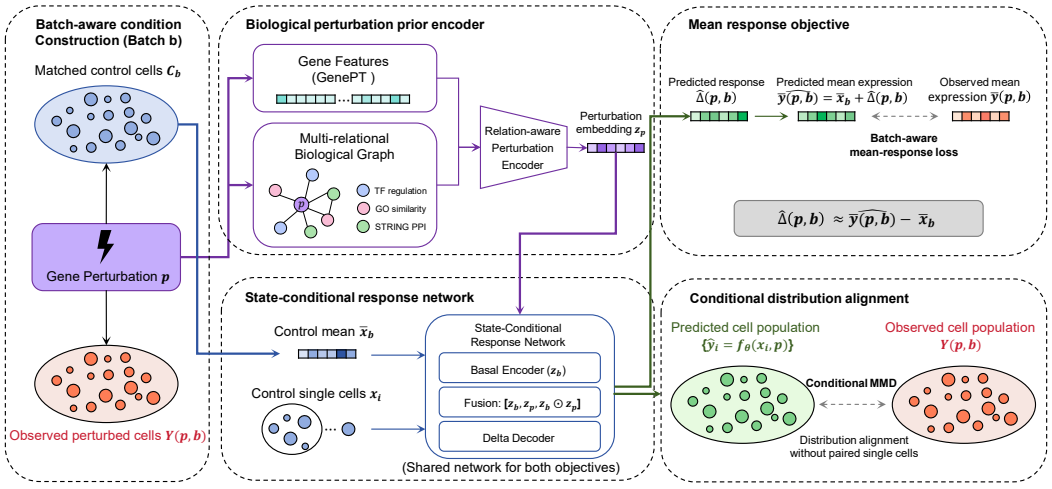


Figure 2: Overview of BASIL. For each perturbation–batch condition, batch-matched control cells (from the same experimental batch) define the pseudo-bulk reference and provide the basal cell population. GenePT features are refined with multi-relational biological priors to obtain the perturbation embedding. The same state-conditional response network acts on either the batch-matched control mean or individual batch-matched control cells, producing the condition-level response and the predicted perturbed-cell population, respectively. Conditional MMD aligns predicted and observed populations within the same perturbation–batch condition without assuming one-to-one cell correspondence.

The overall objective combines a batch-aware mean-response loss and a cell-unpaired conditional distribution alignment loss. The former anchors the predicted perturbation shift, while the latter constrains the predicted single-cell population under the corresponding perturbation–batch condition.

3.3 Batch-Aware Condition Construction

For each perturbation p, BASIL first partitions perturbed cells according to their experimental batch. A valid training condition is defined as a perturbation–batch pair (p, b) containing both control cells and perturbed cells from the same batch. The batch-matched control population provides the reference state:

$$(\bar{x}_b, \mathcal{C}_b, \mathcal{Y}_{p,b}). \tag{5}$$

Importantly, this construction only matches experimental conditions; it does not create correspondence between individual cells.

Because different perturbation–batch conditions may contain different numbers of cells, we define:

$$n_{p,b} = \min(N_b, M_{p,b}), \tag{6}$$

The corresponding condition weight is defined as:

$$w_{p,b} = \mathrm{clip}\left(\sqrt{\frac{n_{p,b}}{\mathrm{median}_{(p',b')\in\mathcal{D}}\, n_{p',b'}}}, 0.5, 2.0\right), \tag{7}$$

where $\mathcal{D}$ denotes the retained training conditions. The weight provides moderate preference to conditions with more reliable population estimates while preventing large batches from dominating optimization.

3.4 Biological Perturbation Prior

To support unseen perturbation prediction, BASIL does not represent perturbations only by trainable identifiers. Instead, each perturbation gene is initialized with a fixed GenePT embedding $e_p \in \mathbb{R}^D$ and refined with biological relation information.

Let P denote the total number of perturbation-gene nodes in the biological graph. Let $H^{(\ell)} \in \mathbb{R}^{P\times d}$ denote their node representations at graph layer ℓ, and let $A_r \in \mathbb{R}^{P\times P}$ represent the adjacency matrix of biological relation r, where $r \in \{\mathrm{TF}, \mathrm{GO}, \mathrm{STRING}\}$. The normalized node features are defined as:

$$U^{(\ell)} = \mathrm{LN}(H^{(\ell)}). \tag{8}$$

The relation-aware message aggregation is defined as:

$$M^{(\ell)} = U^{(\ell)} W_{\mathrm{self}} + \sum_{r\in\mathcal{R}} \sigma(\alpha_r) A_r U^{(\ell)} W_r, \tag{9}$$

where W_{self} and $W_r \in \mathbb{R}^{d\times d}$ are learnable transformations and α_r controls the contribution of each biological relation. TF regulatory edges are obtained from RegNetwork, while GO similarity and STRING interactions provide functional and protein-level biological relations.

After graph propagation, BASIL obtains contextualized embeddings for all P perturbation genes and selects only the nodes active in the current intervention. For perturbation slots $p = (p_1, \ldots, p_K)$, let h_{p_k} denote the final graph-encoded embedding for slot p_k and let m_k indicate whether that slot is active. The final perturbation representation is defined as:

$$z_p = \frac{\sum_{k=1}^{K} m_k h_{p_k}}{\max(1, \sum_{k=1}^{K} m_k)}. \tag{10}$$

For a single-gene perturbation, this selects the corresponding graph-encoded gene embedding; for a two-gene perturbation, it averages the two active embeddings. The formulation removes padding tokens and is invariant to perturbation order.

3.5 State-Conditional Response Network

Given a control expression profile x, BASIL first extracts the basal cellular state:

$$z_b = E_b(x). \tag{11}$$

The perturbation representation z_p is fused with the basal state as follows:

$$z_f = \Phi([z_b; z_p; z_b \odot z_p]), \tag{12}$$

where $[\cdot;\cdot]$ denotes concatenation and $\odot$ denotes element-wise multiplication.

The decoder predicts the perturbation-induced change:

$$d_\theta(x, p) = D_\Delta(z_f), \tag{13}$$

The final expression profile is given by:

$$f_\theta(x, p) = \mathrm{clamp}(x + d_\theta(x, p), 0, \infty). \tag{14}$$

For condition-level supervision, BASIL applies the same response function to the batch-matched pseudo-bulk:

$$\hat{\Delta}_{p,b} = d_\theta(\bar{x}_b, p). \tag{15}$$

For population alignment, the same network transforms individual control cells:

$$\hat{y}_i = f_\theta(x_i, p). \tag{16}$$

Parameter sharing ensures that the predicted mean response and predicted single-cell population are generated by the same learned response operator.

Because the response network is nonlinear, we generally have:

$$d_\theta(\bar{x}_b, p) \neq \frac{1}{N_b} \sum_i d_\theta(x_i, p). \tag{17}$$

Therefore, pseudo-bulk supervision is not treated as a substitute for averaging single-cell predictions; instead, it provides a direct condition-level response target, while the cell-unpaired population alignment objective regularizes cellular variation.

3.6 Batch-Aware Mean-Response Objective

For each perturbation–batch condition, BASIL minimizes the following batch-aware mean-response loss:

$$\mathcal{L}_{\mathrm{mean}} = \frac{\sum_{(p,b)} w_{p,b} \|\hat{\Delta}_{p,b} - \Delta_{p,b}\|_2^2 / G}{\sum_{(p,b)} w_{p,b}}. \tag{18}$$

This objective learns the perturbation-induced change relative to the correct same-batch control reference.

3.7 Cell-Unpaired Conditional Distribution Alignment

The mean objective alone cannot constrain the predicted cellular distribution. Therefore, BASIL introduces a population-level regularizer without assuming cell-level correspondence.

For each condition (p, b), we independently sample S control cells from $\mathcal{C}_b$ and S perturbed cells from $\mathcal{Y}_{p,b}$. The sampled control cells are transformed into predicted perturbed cells:

$$\hat{y}_i = f_\theta(x_i, p). \tag{19}$$

We align the predicted and observed populations using conditional MMD:

$$\widehat{\mathrm{MMD}}^2_{p,b} = \frac{1}{S^2} \sum_{i,i'} k(\hat{y}_i, \hat{y}_{i'}) + \frac{1}{S^2} \sum_{j,j'} k(y_j, y_{j'}) - \frac{2}{S^2} \sum_{i,j} k(\hat{y}_i, y_j). \tag{20}$$

Here, $k(\cdot,\cdot)$ is a multi-scale Gaussian kernel. MMD is calculated independently for every perturbation–batch condition, ensuring that unrelated perturbations are not mixed during distribution comparison.

The final objective is:

$$\mathcal{L} = \mathcal{L}_{\mathrm{mean}} + \lambda(t)\mathcal{L}_{\mathrm{MMD}}, \tag{21}$$

where $\lambda(t)$ is gradually increased during early training. The implementation details of sampling and kernel estimation are provided in Appendix B.

4 EXPERIMENTS

4.1 DATASETS AND EVALUATION PROTOCOL

Genome-scale single-gene perturbations. We evaluate unseen single-gene prediction on three CRISPRi cell-line datasets: RPE1 from Replogle et al. (Replogle et al., 2022), and HepG2 and Jurkat from Nadig et al. (Nadig et al., 2025). Test perturbation identities are disjoint from those used for model fitting. Within each dataset, control cells remain available as basal references and batch-matched control means are used whenever the experimental batch label is available. The single-gene experiments use the benchmark preprocessing with 1,000 highly variable genes, consistent with established unseen-perturbation evaluations (Wenkel et al., 2026; Li et al., 2025).

Norman combinatorial perturbations. We further evaluate on the Norman CRISPRa dataset (Norman et al., 2019) using two complementary out-of-distribution split constructions for condition-level and gene-level generalization (Li et al., 2025). In the *condition-holdout* setting, eligible perturbation conditions are treated as complete interventions and partitioned into disjoint training, validation, and test subsets. A test condition is therefore unseen as a complete intervention, although its constituent gene identities may still occur in other training conditions. We then consider a stricter *gene-holdout* setting: 30% of single-gene identities are withheld from training, double perturbations composed only of seen genes remain in training, and held-out double perturbations are separated into *one-unseen* and *two-unseen* combinations according to whether one or both constituent genes are absent from training. The Norman experiments use 5,000 genes and report scores over all evaluated genes and DEGs.

Metric. Our primary metric is **Pearson** Δ, which compares predicted and observed perturbation-induced expression changes after centering by the appropriate control reference. Correlation-based evaluation of perturbation-induced changes is widely used in perturbation-response prediction, including GEARS and recent OOD benchmarks such as TxPert and X-Pert (Roohani et al., 2024; Wenkel et al., 2026; Li et al., 2025). This choice is particularly appropriate because the target of perturbation prediction is the induced transcriptional response, not reconstruction of the large shared basal-expression component: control centering makes Pearson Δ directly assess the direction and relative gene-wise pattern of the perturbation effect. Higher values indicate better agreement. We also include AverageKnown as a sanity check for perturbation-specific predictive value. For Norman, Pearson Δ is reported over all evaluated genes (All) and benchmark-defined differentially expressed genes (DEG). Appendix C gives the exact definition.

4.2 IMPLEMENTATION DETAILS

The default hidden dimension is 512 with three relation-aware graph layers and dropout 0.1. Gene perturbations use 3,072-dimensional GenePT features. We use AdamW with learning rate 10^{-4}, weight decay 10^{-4}, cosine annealing to 10^{-6}, and gradient clipping at 1.0. Conditional MMD uses 16 control and 16 target cells per condition, a maximum weight of 0.10, and a 2,000-step linear warmup. Baseline scores come from our own reruns under matched benchmark protocols. Complete settings are given in Appendices D and E.4.

4.3 UNSEEN SINGLE-GENE PERTURBATION PREDICTION

Table 1 compares BASIL with statistical, graph-based, foundation-model, perturbation-specific, mean-response, and distribution-aware baselines; references and brief descriptions for all methods are provided in Appendix E.5. BASIL gives the highest mean Pearson Δ on all three cell lines. The ranking of competing methods is strongly context dependent: TxPert (Wenkel et al., 2026) is particularly competitive on RPE1, while AverageKnown (Li et al., 2025) is a strong baseline on HepG2 and Jurkat. The latter shows that a dataset-wide average perturbation response can explain a nontrivial component of correlation-based performance even after control centering. BASIL nevertheless remains strongest across RPE1, HepG2, and Jurkat, indicating that perturbation-specific modeling contributes beyond this shared response structure.

Table 1: Pearson Δ on unseen single-gene perturbations. The best mean in each column is bold.

Method	RPE1	HepG2	Jurkat
CPA	0.085	-0.049	0.106
Linear	0.211	0.231	0.161
Linear-scGPT	0.325	0.270	0.190
CellOracle	0.041	0.038	0.030
GEARS	0.561	0.212	0.131
scFoundation	0.488	0.340	0.230
scGPT	0.585	0.294	0.171
scLAMBDA	0.631	0.405	0.434
AverageKnown	0.543	0.619	0.508
X-Pert	0.634	0.527	0.481
TxPert	0.672	0.500	0.520
scDFM	0.570	0.428	0.213
BASIL	**0.688**	**0.671**	**0.539**

4.4 Norman Condition-Holdout and Combinatorial Generalization

The Norman benchmark separates two forms of out-of-distribution generalization. In *condition holdout*, the complete test intervention is absent from training, but its constituent gene identities may still occur in other training conditions. The *gene-holdout* setting is stricter: selected perturbation genes are removed from training and double perturbations are evaluated according to whether one or both constituent genes are unseen. The one-unseen and two-unseen settings therefore test progressively stronger compositional generalization.

Among evaluated methods, BASIL gives the highest mean in all six columns of Table 2; references for the compared methods are summarized in Appendix E.5. The advantage is present under condition holdout and becomes more pronounced as the evaluation moves to gene-level holdout, with the largest separation in the two-unseen setting. This pattern is consistent with the intended role of external perturbation priors: when one or both perturbation genes lack training response profiles, the model must rely more heavily on information attached to gene identity and biological context rather than memorized perturbation effects.

Table 2: Norman benchmark under condition-holdout and gene-holdout generalization. "One-unseen" and "two-unseen" denote double perturbations with one or two constituent genes withheld from training. The best mean in each column is bold.

	Unseen condition		One-unseen combination		Two-unseen combination	
Method	All	DEG	All	DEG	All	DEG
Linear	0.480	0.561	0.519	0.588	0.474	0.518
Linear-scGPT	0.528	0.618	0.519	0.596	0.481	0.533
CellOracle	0.114	0.171	0.130	0.180	0.090	0.120
GEARS	0.626	0.772	0.558	0.681	0.425	0.515
scFoundation	0.683	0.805	0.512	0.596	0.264	0.288
scGPT	0.650	0.764	0.350	0.450	0.319	0.394
AverageKnown	0.561	0.634	0.527	0.596	0.487	0.536
X-Pert	0.732	0.813	0.727	0.796	0.474	0.549
scDFM	0.621	0.687	0.544	0.612	0.497	0.571
BASIL	**0.759**	**0.836**	**0.778**	**0.848**	**0.631**	**0.730**

4.5 Ablation Study

We evaluate five controlled variants on RPE1 and HepG2, yielding six directly comparable configurations including the full model. The dataset split, model capacity, optimization budget, and test protocol are fixed within each dataset. *Cell-wise supervision* replaces the pseudo-bulk mean target with the expression difference of a randomly sampled control cell and target cell from the same perturbation–batch condition, while retaining conditional MMD; this is a synthetic training pair rather than a true biological correspondence. *w/o GenePT features* replaces GenePT with fixed nonsemantic

perturbation features. *w/o batch matching* keeps the perturbation–batch target groups but uses a global control pool during training. *w/o conditional MMD* removes the population regularizer, whereas *MMD-only* removes explicit mean-response supervision. Appendix F gives the exact construction of each variant.

Table 3: Ablation study on RPE1 and HepG2 using Pearson Δ ($\uparrow$). Scores are rounded to three decimals.

Variant	RPE1	HepG2
Cell-wise supervision	0.661	0.623
w/o GenePT features	0.637	0.599
w/o batch matching	0.679	0.654
w/o conditional MMD	0.684	0.668
MMD-only (w/o mean loss)	0.618	0.568
Full BASIL	**0.688**	**0.671**

The ablations show that the two training resolutions play unequal but complementary roles. Replacing pseudo-bulk supervision with random same-condition cell pairs degrades performance on both datasets, supporting condition-level averaging as a more stable target when true cell correspondence is unavailable. Replacing GenePT with nonsemantic features also causes a substantial drop, supporting the value of external perturbation information for unseen-gene prediction. Batch matching gives a smaller but consistent benefit, indicating that the reference state used to define the perturbation response remains important. Removing conditional MMD lowers the mean Pearson Δ, whereas MMD-only training performs substantially worse than the full model. Together, these results support the central dual-resolution design: explicit condition-level supervision identifies the perturbation shift, while cell-unpaired population regularization improves the shared response operator without replacing the mean-response objective.

5 Conclusion

We presented BASIL, a batch-aware framework for unseen single-cell perturbation prediction without one-to-one control–perturbed cell correspondence. BASIL combines batch-matched pseudo-bulk supervision, GenePT-based perturbation representations, a shared state-conditional response map, and conditional MMD population alignment.

Across CRISPRi cell lines and the Norman CRISPRa benchmark, BASIL achieves strong Pearson Δ performance, including gene-holdout combinatorial settings. Ablations support the batch-conditioned dual-resolution design, GenePT-based perturbation representation, batch-aware referencing, and condition-local population regularization. Future work will extend the framework to broader cell types, perturbation modalities, and more expressive models of combinatorial interactions.

AI use statement

Generative AI tools were used for manuscript drafting assistance, language refinement, and consistency checking. They were not used to generate experimental results or determine the model architecture, experimental design, or scientific conclusions. The authors reviewed and verified all AI-assisted content and take full responsibility for the final manuscript.

Reproducibility statement

The experiment configuration records the dataset version, preprocessing parameters, evaluated gene set, perturbation split and fold, enabled biological relation graphs, model hyperparameters, optimizer schedule, checkpoint-selection criterion, and control-matching mode. These details are provided in Sections 3–4 and Appendices A–F to facilitate reproduction.

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.

Suzi A Aleksander, James Balhoff, Seth Carbon, J Michael Cherry, Harold J Drabkin, Dustin Ebert, Marc Feuermann, Pascale Gaudet, Nomi L Harris, et al. The gene ontology knowledgebase in 2023. *Genetics*, 224(1):iyad031, 2023.

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.

Yiqun T. Chen and James Zou. GenePT: a simple but hard-to-beat foundation model for genes and cells built from ChatGPT. *bioRxiv*, 2023.

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.

Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *The journal of machine learning research*, 13:723–773, 2012.

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.

Kenji Kamimoto, Blerta Stringa, Christy M Hoffmann, Kunal Jindal, Lilianna Solnica-Krezel, and Samantha A Morris. Dissecting cell identity via network inference and in silico gene perturbation. *Nature*, 614(7949):742–751, 2023.

Bowen Li, Chuanyuan Wang, Yijuan Wang, Pengpai Li, and Zhi-Ping Liu. Regnetwork 2025: an integrative data repository for gene regulatory networks in human and mouse. *Nucleic Acids Research*, 54(D1):D1234–D1241, 2026.

Chen Li, Lei Wei, and Xuegong Zhang. Unified multimodal learning enables generalized cellular response prediction to diverse perturbations. *bioRxiv*, pp. 2025–11, 2025.

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.

Thomas M Norman, Max A Horlbeck, Joseph M Replogle, Alex Y Ge, Albert Xu, Marco Jost, Luke A Gilbert, and Jonathan S Weissman. Exploring genetic interaction manifolds constructed from rich single-cell phenotypes. *Science*, 365(6455):786–793, 2019.

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.

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

Damian Szklarczyk, Rebecca Kirsch, Mikaela Koutrouli, Katerina Nastou, Farrokh Mehryary, Radja Hachilif, Annika L Gable, Tao Fang, Nadezhda T Doncheva, Sampo Pyysalo, et al. The string database in 2023: protein–protein association networks and functional enrichment analyses for any sequenced genome of interest. *Nucleic acids research*, 51(D1):D638–D646, 2023.

Gefei Wang, Tianyu Liu, Jia Zhao, Youshu Cheng, and Hongyu Zhao. Modeling and predicting single-cell multi-gene perturbation responses with scLAMBDA. *bioRxiv*, 2024.

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

F Alexander Wolf, Philipp Angerer, and Fabian J Theis. Scanpy: large-scale single-cell gene expression data analysis. *Genome biology*, 19(1):15, 2018.

Chenglei Yu, Chuanrui Wang, Bangyan Liao, and Tailin Wu. scDFM: Distributional flow matching model for robust single-cell perturbation prediction. In *International Conference on Learning Representations*, 2026.

A Detailed Model Specification

A.1 Batch-Matched Condition Construction

The implementation first builds a cache that connects each perturbation to the experimental batches in which it is observed. When the dataset contains a `gemgroup` column, `gemgroup` is used as the batch key. If no such label is available, the code falls back to a single global batch. Let I_b^{ctrl} and $I_{p,b}^{\mathrm{pert}}$ denote the control and perturbed-cell index sets for batch b. A valid training condition contains (i) a control mean computed from I_b^{ctrl}, (ii) a target mean computed from $I_{p,b}^{\mathrm{pert}}$, and (iii) the underlying cell indices used later by the conditional MMD objective.

With the default configuration, a perturbation–batch condition is retained only when a batch-matched control set is available. The pseudo-bulk cache also records the number of target and control cells for each retained condition. The effective count is defined as:

$$n_{p,b} = \min\left(|I_b^{\mathrm{ctrl}}|, |I_{p,b}^{\mathrm{pert}}|\right), \tag{22}$$

The default sample weight is defined as:

$$w_{p,b} = \mathrm{clip}\left(\sqrt{n_{p,b}/\operatorname{median}_{(p',b')\in\mathcal{D}} n_{p',b'}}, 0.5, 2.0\right). \tag{23}$$

Here $\mathcal{D}$ denotes all retained perturbation–batch conditions used to construct the pseudo-bulk cache. This weighting is applied to both the mean-response loss and, after an eligibility mask, the condition-level MMD loss.

Balanced condition sampling. A training mini-batch is sampled in two stages. First, perturbation identities are drawn uniformly with replacement from the set of perturbations represented in the pseudo-bulk cache. Second, one available batch row is sampled for each selected perturbation. This prevents perturbations observed in many batches from being selected simply because they contribute more pseudo-bulk rows. For each selected row, S control cells and S target cells are sampled independently from the corresponding batch-specific pools; sampling uses replacement when a pool contains fewer than S cells.

Fallback behavior. The code contains a global-control fallback for datasets or configurations without a valid batch-specific control pool. However, the default manuscript configuration sets `pseudobulk_require_batch_control=True`, so unmatched perturbation–batch conditions are dropped during pseudo-bulk construction. This is the setting assumed in the method description.

A.2 Perturbation Feature Projection and Relation-Aware Graph Encoder

For genetic perturbations, each gene is associated with a fixed 3,072-dimensional GenePT vector. The feature projector used in our implementation is defined as:

$$e_p \xrightarrow{\text{Linear}} \mathrm{GELU} \xrightarrow{\text{LayerNorm}} \mathrm{Linear} \to h_p^{(0)} \in \mathbb{R}^d, \tag{24}$$

followed by an additional input LayerNorm before graph propagation. The default model dimension is $d = 512$.

Let $H^{(\ell)} \in \mathbb{R}^{P\times d}$ collect perturbation tokens for P perturbation nodes, and define $U^{(\ell)} = \mathrm{LN}(H^{(\ell)})$. For relation r, $A_r \in \mathbb{R}^{P\times P}$ is the adjacency or affinity matrix and $W_r \in \mathbb{R}^{d\times d}$ is a relation-specific feature transformation. One graph block computes:

$$M^{(\ell)} = U^{(\ell)}W_{\mathrm{self}} + \sum_{r\in\mathcal{R}} g_r A_r U^{(\ell)} W_r, \qquad g_r = \sigma(\alpha_r), \tag{25}$$

The message-passing update is followed by:

$$H' = H^{(\ell)} + \mathrm{Dropout}(M^{(\ell)}), \qquad H^{(\ell+1)} = H' + \mathrm{MLP}(\mathrm{LN}(H')). \tag{26}$$

This ordering matches the implementation: feature transformations act on the d-dimensional channel axis, while A_r aggregates across the P node axis. The MLP expands the hidden width by a factor of four, uses GELU and dropout, and projects back to d. Relation gates are initialized with negative logits, so each relation begins with a conservative contribution and can be increased during training.

Supported relations. The default perturbation graph contains three biological relation types. TF regulatory edges are derived from RegNetwork 2025 (Li et al., 2026); GO functional similarity is constructed from Gene Ontology annotations (Aleksander et al., 2023); and protein interactions are taken from STRING (Szklarczyk et al., 2023). Each relation has its own adjacency or affinity matrix, feature transformation, and learnable gate. The relation matrices are scaled independently in the current implementation (default: no additional normalization for TF, max normalization for GO similarity, and log1p-max normalization for STRING) before entering the gated message-passing blocks. Keeping the relations separate allows the encoder to combine directed regulatory evidence, functional similarity, and physical interaction evidence without collapsing them into a single graph.

A.3 Combinatorial Perturbation Encoding

The model accepts one or more perturbation IDs. Let $m_k \in \{0, 1\}$ indicate whether slot k is an active perturbation rather than the control token. The combined perturbation representation is defined as:

$$z_p = \frac{\sum_k m_k h_{p_k}}{\max(1, \sum_k m_k)}. \tag{27}$$

This is slightly different from averaging over all fixed slots: an input such as `A+control` uses the full embedding of `A` rather than shrinking it by a factor of two. For a two-gene intervention, `A+B` and `B+A` therefore have the same representation.

A.4 Basal Encoder, Fusion, and Delta Decoder

The basal encoder takes the full selected gene-expression vector and applies LayerNorm, a linear expansion to $2d$, GELU, dropout, a linear projection to d, and a final LayerNorm. Let z_b denote the result. Fusion uses three blocks of information:

$$[z_b; z_p; z_b \odot z_p] \in \mathbb{R}^{3d}, \tag{28}$$

which are normalized and processed by a two-layer MLP. The delta decoder then maps the fused representation through a $d \to 2d \to G$ MLP. The final expression prediction is given by:

$$d_\theta(x, p) = D_\Delta(z_f), \qquad f_\theta(x, p) = \mathrm{clamp}(x + d_\theta(x, p), 0, \infty)\,. \tag{29}$$

The non-negativity clamp is applied only to the expression prediction; the unconstrained delta $d_\theta(x, p)$ is retained for the mean-response objective.

B Conditional MMD Details

B.1 Multi-Scale Gaussian Kernel

For one perturbation–batch condition, let $\hat{Y} = \{\hat{y}_i\}_{i=1}^S$ be predicted cells and $Y = \{y_j\}_{j=1}^S$ observed perturbed cells. The implementation first concatenates the two sets, computes pairwise squared Euclidean distances, and uses the median off-diagonal distance as a condition-specific scale. If m denotes this median squared distance, the bandwidths are defined as:

$$\Sigma = \left\{\sqrt{s\,m} : s \in \{0.5, 1, 2, 4\}\right\}. \tag{30}$$

The kernel averages four Gaussian RBF kernels:

$$k(a, b) = \frac{1}{|\Sigma|} \sum_{\sigma \in \Sigma} \exp\left(-\frac{\|a - b\|_2^2}{2\sigma^2}\right). \tag{31}$$

The median scale is detached from the gradient graph. The adaptive bandwidth rule is useful because expression dispersion can differ substantially across perturbations and batches; one global kernel width would impose the same scale on heterogeneous conditions.

B.2 Empirical MMD Estimator

For conditional alignment, our implementation uses the non-negative biased empirical MMD estimator (Gretton et al., 2012):

$$\widehat{\mathrm{MMD}}^2(\hat{Y}, Y) = \frac{1}{S^2}\sum_{i,i'} k(\hat{y}_i, \hat{y}_{i'}) + \frac{1}{S^2}\sum_{j,j'} k(y_j, y_{j'}) - \frac{2}{S^2}\sum_{i,j} k(\hat{y}_i, y_j), \tag{32}$$

followed by a numerical clamp at zero. The diagonal terms are retained. This biased estimator is used as a training objective in the small-S regime; it is not presented as an unbiased estimate of test-set distribution distance. When a condition contains fewer than S cells, sampling with replacement can introduce duplicate observations, which is one reason we interpret conditional MMD as a regularizer rather than a standalone population-fidelity metric.

B.3 Condition Weighting and Warmup

Conditional MMD is evaluated only for pseudo-bulk rows containing at least `conditional_mmd_min_target_cells` target cells; the default threshold is three. Eligible rows retain the pseudo-bulk sample weight $w_{p,b}$. The per-condition values are aggregated as:

$$\mathcal{L}_{\mathrm{MMD}} = \frac{\sum_{(p,b)} \tilde{w}_{p,b}\widehat{\mathrm{MMD}}^2_{p,b}}{\sum_{(p,b)} \tilde{w}_{p,b}}, \tag{33}$$

where $\tilde{w}_{p,b} = w_{p,b}$ for eligible conditions and zero otherwise.

The effective MMD coefficient follows a linear warmup:

$$\lambda(t) = \lambda_{\max} \min\left(\frac{t}{T_{\mathrm{warm}}}, 1\right), \tag{34}$$

with default $\lambda_{\max} = 0.10$ and $T_{\mathrm{warm}} = 2000$ optimization steps. The intent is to first establish a stable mean-response mapping and then progressively impose the population-level constraint.

B.4 Why MMD Is Conditioned on Perturbation and Batch

A global MMD between all predicted and all observed perturbed cells would admit a degenerate solution in which cells generated for one perturbation are matched against cells from another perturbation. BASIL instead computes one MMD value per selected perturbation–batch condition. Thus the distribution objective is local to the same condition used to define the mean response. The mean loss provides explicit first-moment supervision, while MMD supplies a complementary kernel two-sample constraint on the predicted and observed cell populations.

C Evaluation Metric

C.1 Pearson Δ

For a perturbation p, let $\hat{\delta}_p$ and δ_p denote the predicted and observed perturbation-induced mean-expression changes after centering by the corresponding control reference. Pearson Δ is defined as:

$$\text{Pearson } \Delta(p) = \mathrm{corr}(\hat{\delta}_p, \delta_p). \tag{35}$$

The dataset-level score averages the per-perturbation correlations according to the benchmark protocol. Pearson Δ is the primary metric in this work because perturbation-response prediction is fundamentally concerned with the induced change relative to a basal state. Correlation-based evaluation of perturbation-induced changes is widely used in the field, including GEARS and recent OOD evaluations such as TxPert and X-Pert (Roohani et al., 2024; Wenkel et al., 2026; Li et al., 2025). Centering by the appropriate control suppresses agreement arising from the large shared basal-expression component and focuses the score on whether the model recovers the direction and relative gene-wise structure of the perturbation response. Shared response structure can still yield nontrivial correlation, which motivates reporting AverageKnown alongside learned models as a non-parametric sanity-check baseline.

For the single-gene CRISPRi benchmarks, Pearson Δ is computed over the fixed evaluated gene space. For Norman, "All" uses the full evaluated gene set, whereas "DEG" restricts the same correlation to benchmark-defined differentially expressed genes. The DEG view places greater weight on genes with a measurable perturbation response and therefore complements the broader all-gene score. Pearson Δ directly evaluates the principal benchmark target—the perturbation-induced response vector—while conditional MMD provides complementary cell-unpaired population regularization during training.

D Training and Inference Details

D.1 Training Procedure

One training iteration follows the procedure below.

1. Sample a mini-batch of perturbation identities and one batch-specific pseudo-bulk row for each identity.
2. Load the batch-matched control mean $\bar{x}_b$, target mean $\bar{y}_{p,b}$, perturbation ID sequence, sample weight, and independently sampled same-batch control and target single cells.
3. Encode all perturbation nodes through the shared GenePT projector and relation-aware graph blocks, select the active perturbation tokens, and average active tokens for combinations.
4. Predict the pseudo-bulk response $\hat{\Delta}_{p,b}$ and compute weighted mean-delta MSE.
5. Repeat the complete perturbation ID sequence for every sampled control cell, apply the same response network, and obtain S predicted single cells per condition.
6. Compute condition-specific adaptive-bandwidth MMD values, apply the target-cell eligibility mask, and aggregate them with pseudo-bulk weights.
7. Form $\mathcal{L} = \mathcal{L}_{\text{mean}} + \lambda(t)\mathcal{L}_{\text{MMD}}$, backpropagate, clip the gradient norm if enabled, and update the model with AdamW.
8. Step the cosine learning-rate scheduler. At evaluation intervals, compute the configured validation metric and save the best checkpoint.

The default configuration evaluates every 2,000 steps and uses an early-stopping patience of five evaluation rounds. Training is capped at 20,000 optimization steps. The code logs total loss, mean-delta loss, conditional MMD, effective MMD weight, train/validation scores, and pseudo-bulk cache statistics for reproducibility.

D.2 Inference and Batch-Aware Aggregation

At test time, prediction is performed per perturbation. The reported experiments use batch-matched pseudo-bulk control means for evaluation: the implementation builds one control mean for each target batch (`gemgroup` when available), predicts from each batch-specific mean, and aggregates the resulting predictions with weights proportional to the observed number of target cells in the corresponding batch. This mirrors the composition of the benchmark pseudo-bulk and avoids evaluating a multi-batch perturbation against one arbitrary control batch. Under this reported evaluation path, no fixed-size control-cell sample is used. The configuration parameter `eval_control_cells=128` applies only to alternative sampled-control evaluation paths when pseudo-bulk control-mean evaluation is disabled. The target-cell counts used for retrospective aggregation are an evaluation convenience rather than information assumed to be available prospectively; deployment on a new perturbation would require a prespecified aggregation rule such as uniform batch weighting or weighting by available control counts.

For single-cell prediction, control cells can be sampled from the same experimental batch as the target perturbation. The response network is deterministic given a control state and perturbation representation; population diversity therefore comes from the empirical control-state distribution rather than an additional noise variable.

D.3 Computational Considerations

The perturbation graph encoder operates on the perturbation-node graph and is shared across all conditions in a forward pass. For the distributional branch, each condition expands to S control cells, so response-network compute grows linearly in S. MMD additionally requires pairwise distances within each predicted and observed set, giving $O(S^2G)$ distance computation per condition in the direct implementation. With the default $S = 16$, this cost is small relative to full-transcriptome response-network evaluation. The pseudo-bulk cache is precomputed once before training, avoiding repeated aggregation of batch-matched control and target means.

E Experimental Details

E.1 Single-Gene CRISPRi Datasets

We use three genome-scale CRISPRi cell-line datasets for the main single-gene benchmark: RPE1 from Replogle et al. (Replogle et al., 2022), and HepG2 and Jurkat from Nadig et al. (Nadig et al., 2025). The task is OOD prediction of perturbation identities not observed during training. Control cells remain available as basal references, and non-single-gene perturbation conditions are excluded from this benchmark. The processed single-gene expression space contains 1,000 highly variable genes and is kept fixed within each direct comparison.

The data loader uses an `unseen` split mode that partitions perturbation identities into disjoint train, validation, and test sets within each fold. The same split definition is used consistently for model training and evaluation.

E.2 Norman CRISPRa Dataset

Norman et al. (Norman et al., 2019) profiled single- and double-gene CRISPRa interventions in K562. Our data pipeline canonicalizes each condition into two slots, places active genes before control padding, and preserves both genes for double perturbations. We use two distinct split constructions. For condition holdout, all eligible single- and double-gene perturbation conditions are shuffled and assigned to disjoint training, validation, and test subsets; the current implementation uses 70%/10%/20%, respectively. For gene holdout, 70% of single-gene identities define the seen set and the remaining 30% are held out. Training contains seen single perturbations and double perturbations composed entirely of seen genes, whereas the test set contains only double perturbations with exactly one or two held-out genes, yielding the one-unseen and two-unseen subsets.

E.3 Preprocessing and Expression Space

The data-processing code uses Scanpy-based highly variable gene selection (Wolf et al., 2018). The single-gene RPE1, HepG2, and Jurkat benchmarks use 1,000 highly variable genes, whereas the Norman results in Table 2 use 5,000 genes. Within each direct comparison, the processed expression space, perturbation labels, and split definition are fixed.

The model predicts normalized expression values in the selected gene space. Batch matching is performed after preprocessing, using the experimental batch label carried in the processed `AnnData` object. The matched-control mean and target mean are therefore computed in exactly the same expression representation used by the model.

E.4 Default Hyperparameters

Table 4: Default configuration used in our implementation.

Hyperparameter	Default value
Model dimension d	512
Relation-aware graph layers	3
Dropout	0.1
GenePT dimension	3072
Condition batch size	24
Optimizer	AdamW
Learning rate	10^{-4}
Minimum learning rate	10^{-6}
Weight decay	10^{-4}
Maximum optimization steps	20,000
Gradient-norm clipping	1.0
Mean-response loss weight	1.0
Conditional MMD cells per condition	16
Conditional MMD minimum target cells	3
Maximum conditional MMD weight	0.10
Conditional MMD warmup	2,000 steps
Evaluation interval	2,000 steps
Early-stopping patience	5 evaluation rounds
Random seeds	0 / 1 / 2 / 3
Evaluation control input	batch-matched pseudo-bulk means
TF / GO / STRING graphs	enabled / enabled / enabled

All reported experimental scores, including BASIL, baselines, and ablations, are arithmetic means over these runs. The 16 cells per condition are used only for the training-time conditional MMD term. Reported benchmark predictions use batch-matched pseudo-bulk control means, as described in Appendix D, rather than a fixed-size sample of control cells.

E.5 Baseline Descriptions

The main tables compare BASIL with representative statistical, naive mean-response, graph-based, foundation-model, perturbation-specific, and distribution-aware baselines. Every numerical baseline entry in Tables 1 and 2 is produced by our own reruns under the corresponding dataset preprocessing, split, expression space, and evaluation protocol; no table value is copied from a published result. We summarize the compared methods below.

Linear and Linear-scGPT. Linear baselines test whether the perturbation response can be explained by a simple linear mapping. Linear-scGPT augments a linear predictor with pretrained scGPT-derived representations. We rerun both methods on the same processed data, perturbation splits, expression spaces, and evaluation protocol used for the corresponding BASIL comparison.

CellOracle. CellOracle is a regulatory-network-oriented perturbation method (Kamimoto et al., 2023). We rerun CellOracle on the same processed benchmark splits and evaluation protocol used for the corresponding BASIL comparison.

GEARS. GEARS is a graph-based perturbation model that uses gene relationships to generalize to unseen single and combinatorial perturbations (Roohani et al., 2024). It provides a graph-based reference for OOD genetic perturbation prediction.

scGPT and scFoundation. These methods represent large pretrained single-cell models (Cui et al., 2024; Hao et al., 2024). Their inclusion tests whether generic single-cell pretraining alone is sufficient for perturbation-specific OOD prediction.

CPA. CPA learns compositional latent perturbation representations and provides a reference for combination generalization without the explicit relation-aware biological prior used by BASIL (Lotfollahi et al., 2023).

scLAMBDA. scLAMBDA uses language-model-derived gene embeddings to model single- and multi-gene perturbation responses (Wang et al., 2024), providing a relevant comparison for perturbation representations that incorporate external gene semantics.

AverageKnown. AverageKnown is a non-parametric mean-response baseline used in perturbation benchmarks such as X-Pert (Li et al., 2025). It predicts held-out perturbations from an average response estimated from perturbations observed in training and therefore does not condition on the identity of the unseen intervention. Its inclusion tests whether learned models capture perturbation-specific structure beyond a dataset-wide average response.

X-Pert. X-Pert uses a unified multimodal formulation for perturbations and cellular states and is included as a strong baseline in the Norman and single-gene comparisons (Li et al., 2025).

TxPert. TxPert combines perturbation representations from multiple biological graphs with batch-aware control referencing and is included in the single-gene benchmark (Wenkel et al., 2026).

scDFM. scDFM is a distribution-aware perturbation model based on conditional flow matching with an MMD objective (Yu et al., 2026). It provides a direct reference for population-level transport rather than only pseudo-bulk response prediction.

F Ablation Configuration Details

All ablation runs use the same train/validation/test split, optimization budget, model width, and evaluation procedure as the corresponding full-model run. The five variants are defined as follows.

Cell-wise supervision. The training loop first selects the same perturbation–batch condition as in the full model. Instead of supervising the mean-response branch with the batch-specific pseudo-bulk difference, it randomly samples one control cell from that batch and one target cell from the same perturbation and batch, and uses their expression difference as the point target for the mean-delta MSE. Conditional MMD is unchanged and still uses independently sampled control and target cells from the same perturbation–batch condition. This variant tests whether condition-level pseudo-bulk supervision is preferable to arbitrary single-cell pairing when no true cell correspondence exists.

w/o GenePT features. GenePT vectors are replaced by fixed unit-norm random perturbation features of the same input dimensionality, sampled once and held fixed, while the feature projector and downstream response architecture are retained. This tests the contribution of external gene-semantic information under an otherwise matched predictor.

w/o batch matching. Target cells remain grouped by perturbation and their original batch, so the target pseudo-bulk construction, sample weights, and model architecture are unchanged. During training, however, the control mean and the control-cell pool used by conditional MMD are replaced with a global pool containing control cells from all batches. Test-time evaluation remains batch matched for all variants. The comparison therefore isolates the effect of using same-batch controls during training rather than changing the evaluation target.

w/o conditional MMD. The conditional MMD term is disabled while the batch-aware pseudo-bulk mean-response objective is retained. This variant measures whether population-level regularization provides additional benefit beyond condition-level mean supervision.

MMD-only (w/o mean loss). The explicit pseudo-bulk mean-response loss is removed and training is driven by conditional MMD. Conditional MMD remains subject to the minimum target-cell eligibility rule, so this setting is best interpreted as a diagnostic of distribution-only training rather than a perfectly symmetric one-factor counterpart to the full objective. Its large performance degradation

shows that population alignment alone is insufficient to identify the condition-level perturbation effect measured by Pearson Δ.

G Limitations and Future Directions

Metric focus and complementary objectives. Pearson Δ is the primary endpoint because the task is to predict the perturbation-induced transcriptional response relative to an appropriate control state, and the metric directly evaluates the direction and relative gene-wise pattern of that response. Conditional MMD has a complementary role: it regularizes cell-level predictions within the same perturbation–batch condition while the explicit pseudo-bulk objective anchors the response vector evaluated by Pearson Δ. Future evaluations can broaden the metric suite to characterize additional biological and distribution-sensitive properties while retaining Pearson Δ as the principal response metric.

Finite-sample and high-dimensional MMD effects. Conditional MMD samples a fixed number of cells and uses replacement when a condition contains fewer cells than requested. This keeps the objective defined for sparse conditions but can duplicate observations and increase the bias of the empirical estimator. In addition, the adaptive bandwidth is estimated from the combined predicted and observed samples in the selected gene-expression space; large distribution shifts or high-dimensional distance concentration can affect its sensitivity. Adaptive sample sizes, stricter eligibility thresholds, within-population bandwidth estimates, or latent-space kernels are natural alternatives.

Dependence on batch-matched controls. The framework assumes that experimental batches contain adequate control cells. When controls are absent or extremely sparse, the batch-specific mean response is poorly determined. The implementation contains a global fallback, but this changes the interpretation of the target and should be reported explicitly whenever used.

Distribution alignment is not cell-level causal identification. MMD constrains empirical distributions through a two-sample discrepancy (Gretton et al., 2012) and is appropriate for cell-unpaired population comparison, but it cannot identify a unique mapping from an individual control cell to its counterfactual perturbed state. The deterministic response function used here should therefore be interpreted as a transport-like predictive mapping, not as recovery of a true longitudinal cell trajectory.

External priors can be incomplete or context-independent. GenePT features and public biological graphs improve coverage of unseen genes, but they can contain annotation bias and relationships that do not hold uniformly across cell types. Relation gates can down-weight a source globally within a layer, yet the current architecture does not infer a separate graph topology for each cellular context. In addition, the relation-specific graph scalings used here are not a substitute for a systematic study of degree-normalized message passing, especially for hub-heavy biological networks.

Simple composition for multiple perturbations. The current combination encoder averages active perturbation embeddings. This provides order invariance and supports unseen combinations, but it does not explicitly model higher-order interactions before fusion with the basal state. A future extension could use set attention or interaction-specific graph reasoning while preserving zero-shot support for unseen combinations.

Scalability to broader contexts. The current experiments focus on a small number of well-established cell lines and the Norman combination benchmark. Evaluating the same batch-conditioned dual-resolution objective across additional cell types, genetic perturbation settings, doses, and disease contexts will test how broadly the observed gains transfer beyond the present benchmarks.