CELLKAFC: KNOWLEDGE-AWARE SINGLE-CELL PERTURBATION PREDICTION WITH FLOW CORRECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Single-cell perturbation-response modeling is essential for decoding cellular mechanisms and accelerating drug discovery. However, predicting perturbation-induced gene expression distributions remains a formidable machine learning challenge: the data is inherently high-dimensional, sparse, and heterogeneous, and target conditions frequently exhibit substantial distribution shifts from cell distributions. While existing methods improve conditional prediction, they often underutilize gene-level biological priors and lack explicit mechanisms to correct condition-dependent distribution shifts. In this work, we introduce CellKAFC, a Knowledge-Aware Flow Correction framework for single-cell perturbation modeling. CellKAFC integrates biological priors into a conditional transport formulation and couples this process with an explicit correction for distributional mismatch across varying conditions. Through extensive evaluation on public benchmarks—including Norman, RPE1, and ComboSciPlex—CellKAFC consistently outperforms prior state-of-the-art methods. Notably, it reduces mean squared error by 12.5–40.5% relative to corresponding uncorrected baselines. Furthermore, an orthogonal decomposition analysis reveals that 21.7–42.4% of this net improvement stems from components orthogonal to the decoded dynamic space, validating the efficacy of our correction mechanism. These results demonstrate that knowledge-aware transport coupled with explicit shift correction provides a principled and highly transferable representation for single-cell perturbation modeling.

1 INTRODUCTION

Single-cell perturbation screens reveal how genetic and chemical interventions reshape cellular states, including coordinated transcriptional programs and genetic interactions at single-cell resolution (Norman et al., 2019; Replogle et al., 2022). Yet possible interventions, cell contexts and combinations far exceed what can be measured experimentally. Predicting unobserved responses can therefore reduce screening costs, prioritize experiments and support the discovery of cellular mechanisms and therapeutic opportunities (Lotfollahi et al., 2023; Roohani et al., 2024).

Single-cell RNA profiling measures each cell only once, so perturbation datasets contain separate control and perturbed populations rather than paired before-and-after observations (Bunne et al., 2023; Jiang et al., 2024). A model must therefore predict how the control population changes under an intervention without observing the outcome for each input cell. This is difficult because expression data are high-dimensional and sparse, perturbations may affect only a few genes, and cells can respond differently to the same intervention. Responses also vary with cell type, batch and dose. With limited examples for each condition, the model must recover gene-specific effects while representing variation across cells.

Recent methods can be grouped by the space in which they model responses. The first group operates primarily in a latent gene-expression space. CPA composes basal-cell and perturbation factors, while CellOT learns transport maps between unpaired populations (Lotfollahi et al., 2023; Bunne et al., 2023). Biolord disentangles cellular attributes, and SAMS-VAE models sparse additive perturbation effects (Piran et al., 2024; Bereket & Karaletsos, 2023). These methods reduce dimensionality and support compositional prediction and transfer across contexts. However, decoding can attenuate rare or condition-specific responses and leave mean bias in gene expression space even when

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.

latent distributions align. Enlarging the latent space increases the estimation burden under sparse observations.

The second group models responses in gene space or retains explicit gene-level outputs through gene-token objectives. GEARS uses gene–gene knowledge graphs to predict multigene responses, while scGPT learns transformer-based representations of gene expression and perturbations (Roohani et al., 2024; Cui et al., 2024). These methods preserve individual gene responses and can directly incorporate biological knowledge. However, sparse and imbalanced outputs increase learning demands and sensitivity to distribution shift, potentially favoring frequent genes or observed condition averages.

The third group couples latent representations with gene-space decoding or population-level distribution matching. PerturbNet maps perturbations to cell states, while CellFlow uses flow matching to generate responses (Yu et al., 2025; Klein et al., 2025). STATE uses distributional objectives across cell sets, and scPRAM combines variational alignment, optimal transport and gene-level attention (Adduri et al., 2026; Jiang et al., 2024). scDFM and scBIG incorporate gene interactions or gene modules into conditional flow matching (Yu et al., 2026; Ruan et al., 2026). CINEMA-OT separates confounding from perturbation-associated variation before transport, but is not a full conditional generative predictor (Dong et al., 2023). These hybrid designs model heterogeneous responses, but still couple cell-dependent dynamics with condition-level mean shifts. Restricted representations may leave systematic residuals in gene space, while expanding the modeled space increases estimation demands under sparse observations and distribution shift.

This motivates CellKAFC (Knowledge-Aware Flow Correction), which combines knowledge-conditioned transport with residual mean correction. Conditional flow matching updates control-cell effect coordinates defined by PCA–ICA, using fused biological priors. An expression-space kernel then transfers training-condition residual means to each query, correcting mean bias, including components outside the decoder space, without changing differences between cells. Our contributions are threefold. First, we integrate complementary biological priors into constrained transport for sparse perturbation-response prediction. Second, we correct residual mean bias separately and quantify the correction within and outside the dynamic space. Third, evaluations on Norman, RPE1 and ComboSciPlex demonstrate improved prediction of unseen and combinatorial perturbations, supported by ablations and response-level analyses.

2 Related work

Knowledge learning and injection. A line of work improves perturbation prediction by encoding relationships among genes, perturbations, and biological attributes. GEARS propagates perturbation signals through gene–gene knowledge graphs and functional relationships to predict novel multigene outcomes (Roohani et al., 2024). scGPT learns large-scale gene-token representations that can be adapted to expression and perturbation contexts, whereas the scBIG preprint groups genes into data-driven modules to provide module-inductive representations (Cui et al., 2024; Ruan et al., 2026). Related structured-latent approaches, including Biolord and SAMS-VAE, disentangle known attributes or impose sparse additive mechanism shifts to improve interpretability and factorized generalization (Piran et al., 2024; Bereket & Karaletsos, 2023). These studies establish the value of biological structure and pretrained representations, while our focus is to adaptively combine complementary prior sources according to their availability and relevance to each intervention.

Perturbation distribution learning. In addition to encoding biological information, perturbation models must capture how an intervention changes the distribution of cell states. CPA represents basal states and perturbation effects as compositional latent factors, CellOT learns a neural optimal-transport map between unpaired control and perturbed populations, and CINEMA-OT separates confounding from perturbation-associated variation before transport (Lotfollahi et al., 2023; Bunne et al., 2023; Dong et al., 2023). Recent generative and hybrid methods extend this idea with population-level objectives: PerturbNet models single-cell responses for unseen chemical and genetic interventions, scPRAM combines variational alignment and optimal transport with gene-level attention, and the CellFlow preprint and scDFM use conditional flow matching to model heterogeneous responses (Yu et al., 2025; Jiang et al., 2024; Klein et al., 2025; Yu et al., 2026). The scBIG preprint combines gene-module representations with conditional flow matching, whereas STATE predicts responses across sets of cells with distributional objectives (Ruan et al., 2026; Adduri et al., 2026). These approaches

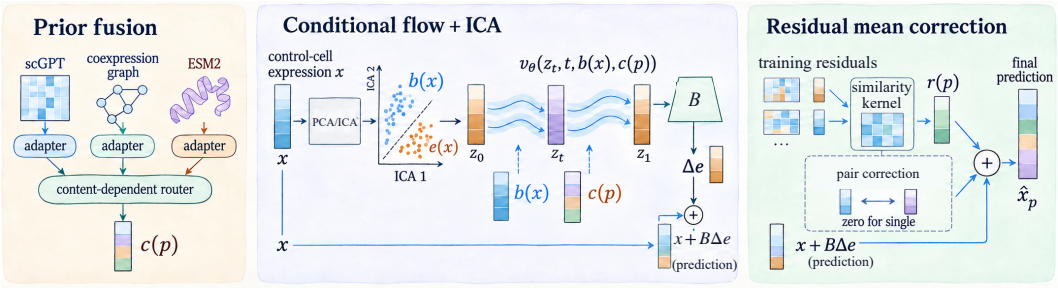


Figure 1: Overview of CellKAFC, illustrated with genetic perturbation priors. Left: adapters and a content-dependent router fuse available priors into $c(p)$. Middle: PCA–ICA separates background $b(x)$ and effect $e(x)$ coordinates; the conditional flow maps $z_0 = e(x)$ to z_1, and B decodes the displacement $\Delta e = z_1 - z_0$. Right: a similarity kernel transfers training-condition residual means to obtain $r(p)$. The pair-specific offset shown in the dashed box is zero for single interventions and is included in the uncorrected prediction before fitting the residual map.

model heterogeneous cellular responses; our work complements this direction by separately correcting condition-specific mean errors that remain in gene expression space after transport.

3 METHOD

Problem setup. We study perturbation prediction from unpaired single-cell expression data. Each cell has a preprocessed expression vector $x \in \mathbb{R}^{G_{\text{in}}}$, and condition p specifies a single genetic or chemical intervention or a pair of interventions. Control and perturbed populations are measured separately, without cell-level correspondence. Given a held-out condition, we predict its perturbed population by transforming control cells using intervention information, without using the query's measured response during fitting.

Overview. CellKAFC combines prior fusion, constrained flow matching and residual mean correction (Figure 1). Source-specific adapters and a learned router integrate biological priors into the condition representation $c(p)$. Conditioned on $c(p)$ and the cell background $b(x)$, the flow transports effect coordinates, and the fixed decoder B maps their displacement back to expression changes. A similarity kernel then transfers training-condition residual means to obtain a shared offset $r(p)$, which adjusts the predicted population mean while preserving differences between cells.

3.1 BIOLOGICAL PRIOR INTEGRATION

For genetic interventions, available priors include scGPT gene embeddings, coexpression-graph features from training expression, and ESM2 protein-sequence embeddings (Cui et al., 2024; Lin et al., 2023). These sources provide information about gene expression relationships and protein sequences. For drugs, ChemBERTa representations and Morgan fingerprints encode molecular structure, while a biological branch combines drug-target features with external LINCS responses (Chithrananda et al., 2020; Rogers & Hahn, 2010; Subramanian et al., 2017). The response features exclude the evaluated A549 context; a bounded learned gate controls their contribution and is zero when response evidence is unavailable. Source construction and protocol-specific subsets are given in Appendix B.

We learn how to combine the available sources for each intervention. For component a, a trainable adapter maps each frozen prior feature $u_k(a)$ into a shared space, yielding $h_k(a)$ after scale matching. A router takes these projected features and their availability indicators as input and produces logits $\ell_k(a)$. With mask $m_k(a) \in \{0, 1\}$, the fused representation is

$$h(a) = \sum_k \alpha_k(a) h_k(a), \qquad \alpha_k(a) = \frac{m_k(a) e^{\ell_k(a)}}{\sum_j m_j(a) e^{\ell_j(a)}}. \tag{1}$$

The weights are normalized over available, enabled sources; $h(a) = 0$ if none is available. We sum the component-token representations to obtain $c(p)$, using a fixed control/padding slot for single

interventions. This pooling is invariant to component order. The feature tables remain frozen, while the adapters and router are trained with the flow model. The resulting $c(p)$ conditions transport, and its normalized form defines condition similarity for residual correction.

3.2 KNOWLEDGE-CONDITIONED FLOW MATCHING

We restrict learned expression changes to a selected effect space. Standardization, PCA and ICA are fitted to training expression. ICA components are ranked by their maximum Chatterjee dependence (Chatterjee, 2021) on training-condition indicators: the least dependent form background coordinates $b(x)$, and the remaining components form effect coordinates $e(x) \in \mathbb{R}^d$. The flow updates $e(x)$ while using the fixed $b(x)$ as cell context. A decoder $B \in \mathbb{R}^{G_{\text{in}} \times d}$ maps effect-coordinate changes back to expression space.

Following conditional flow matching (Lipman et al., 2023), we transport the control-cell effect state through

$$\frac{\mathrm{d}z_t}{\mathrm{d}t} = v_\theta(z_t, t, b(x), c(p)), \qquad z_0 = e(x), \qquad t \in [0, 1]. \tag{2}$$

The velocity network combines $b(x)$, a sinusoidal time embedding and $c(p)$ to condition residual blocks acting on the projected state z_t. Each block uses this conditioning vector to scale and shift its normalized hidden state. Integrating Equation 2 gives the final effect state $F_\theta(e(x), b(x), c(p))$, from which we predict

$$\widehat{x}_p^0 = x + B\{F_\theta(e(x), b(x), c(p)) - e(x) + q_\phi(p)\}. \tag{3}$$

Here $q_\phi(p)$ is a bounded pair offset fitted to training-pair residual means after the flow is frozen. Its inputs are the sum, absolute difference and elementwise product of the two component representations. It is zero for single interventions and, like the flow update, remains within the decoded effect space. Adding the decoded change to x preserves source expression outside that space.

Training objectives. For each training condition, balanced entropic optimal transport couples control and perturbed cells using normalized squared distance in the retained ICA coordinates. From each coupled pair, we obtain source and target effect states e_0 and e_1. At $t \sim \mathcal{U}(0, 1)$, the interpolation $z_t = (1 - t)e_0 + te_1$ and reference velocity $e_1 - e_0$ define

$$\mathcal{L}_{\text{FM}} = \mathbb{E}\left[\frac{1}{d}\left\|v_\theta(z_t, t, b(x), c(p)) - (e_1 - e_0)\right\|^2\right]. \tag{4}$$

The expectation is over conditions, coupled cells and transport times. These pairs provide training supervision, not observed cellular trajectories.

We also supervise predicted populations and their expression means. The one-step estimate $\widetilde{e}_1 = z_t + (1-t)v_\theta(z_t, t, b(x), c(p))$ approximates the terminal effect state. For each condition, we compare predicted samples $\{\widetilde{e}_i\}_{i=1}^m$ with target samples $\{e_i\}_{i=1}^m$ using maximum mean discrepancy (Gretton et al., 2012):

$$\mathcal{L}_{\text{MMD}}^{(p)} = \frac{1}{m^2}\sum_{i,j} k_\sigma(\widetilde{e}_i, \widetilde{e}_j) + \frac{1}{m^2}\sum_{i,j} k_\sigma(e_i, e_j) - \frac{2}{m^2}\sum_{i,j} k_\sigma(\widetilde{e}_i, e_j), \tag{5}$$

where $k_\sigma(u, v) = \exp[-\|u - v\|^2/(2\sigma^2)]$, with bandwidth estimated from pairwise distances in the combined batch. Averaging over conditions gives $\mathcal{L}_{\text{MMD}}$. For mean supervision, $\widehat{\mu}_{p,\text{flow}}$ denotes the average decoded prediction from a differentiable flow rollout, using Equation 3 with $q_\phi = 0$. We compare it with the observed training mean μ_p on the input gene panel:

$$\mathcal{L}_\mu = \mathbb{E}_p\left[\frac{1}{G_{\text{in}}}\left\|\widehat{\mu}_{p,\text{flow}} - \mu_p\right\|^2\right]. \tag{6}$$

The overall objective is

$$\mathcal{L} = \mathcal{L}_{\text{FM}} + 0.02\mathcal{L}_{\text{MMD}} + \lambda_\mu(s)\mathcal{L}_\mu, \tag{7}$$

where $\lambda_\mu(s)$ increases linearly during warmup. The three terms supervise velocity, the predicted effect-state population and decoded expression means, respectively. Architecture, sampling schedules, integration settings and pair-offset fitting are detailed in Appendix B.

3.3 RESIDUAL MEAN CORRECTION

After fitting the flow model in Section 3.2, we freeze the neural predictor and correct its remaining errors in mean gene expression. For training condition p_j, the residual $r_j = \mu_j - \widehat{\mu}_j^0$ is the observed mean minus the predicted mean on the G output genes. We stack these residuals into $R \in \mathbb{R}^{J \times G}$ for J training conditions. For a query p, we use normalized condition embeddings $\bar{c}(p)$ to measure similarity:

$$H_{ij} = \exp\left[\frac{\bar{c}(p_i)^\top \bar{c}(p_j) - 1}{\tau}\right], \qquad k_j(p) = \exp\left[\frac{\bar{c}(p_j)^\top \bar{c}(p) - 1}{\tau}\right]. \tag{8}$$

Here H compares training conditions with one another, and $k(p)$ compares the query with each training condition. Kernel ridge regression uses these similarities to estimate the query's residual. We add the result to each uncorrected cell prediction:

$$r(p) = \mathcal{G}\big(\beta k(p)^\top (H + \lambda I)^{-1} R\big)^\top, \qquad \widehat{x}_p = \widehat{x}_p^0 + r(p), \tag{9}$$

where τ, λ and β control kernel bandwidth, ridge regularization and correction strength.

We choose these parameters by leaving out each training condition in turn and predicting its residual from the remaining conditions. The gate $\mathcal{G}$ applies correction only if these predictions improve on zero correction and the query is sufficiently similar to a training condition. It also limits the correction norm. Query responses are not used for fitting or selection. Details are given in Appendix B.

With B restricted to the G output genes, let P denote the orthogonal projector onto its column space. Let I denote the $G \times G$ identity matrix. Then $I - P$ projects onto the orthogonal complement. For control mean μ_X and target mean μ_p, the uncorrected mean error has two parts, inside and outside this space:

$$\frac{\left\|\widehat{\mu}_p^0 - \mu_p\right\|^2}{G} = \underbrace{\frac{\left\|P(\widehat{\mu}_p^0 - \mu_p)\right\|^2}{G}}_{E_\parallel^0} + \underbrace{\frac{\|(I-P)(\mu_X - \mu_p)\|^2}{G}}_{E_\perp^0}. \tag{10}$$

Because the flow updates lie in the column space of B, $(I - P)B = 0$ and hence $(I - P)\widehat{\mu}_p^0 = (I - P)\mu_X$. Thus, outside the flow's update space, the predicted mean remains equal to the control mean, which explains the use of μ_X in the second term. Correction acts directly in gene expression space and can therefore adjust both terms without changing the flow's update space. All cells receive the same offset, so their mean changes while differences between cells and covariance remain unchanged. The derivation and projection control are given in Appendices A and D.2.

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

Datasets and prediction tasks. We evaluate genetic perturbation prediction on Norman K562 and RPE1, and drug response prediction on ComboSciPlex A549 (COM) (Norman et al., 2019; Replogle et al., 2022; Lotfollahi et al., 2023). Norman includes two protocols: Additive predicts held-out gene pairs, with incomplete training coverage of their constituent single responses; Holdout evaluates unseen single perturbations and their combinations. RPE1 tests single-gene perturbations, whereas COM includes both single drugs and drug pairs. Detailed data splits and preprocessing are provided in Appendix B.

Baselines. We compare against Control and six established perturbation prediction models. Control uses the unperturbed expression profile as the prediction. GEARS (Roohani et al., 2024) incorporates gene coexpression and functional relationships through graph neural networks. CPA (Lotfollahi et al., 2023) represents basal cell states, perturbations and covariates as compositional latent factors. scGPT (Cui et al., 2024) provides a transformer-based architecture for modeling gene expression and perturbation responses. STATE (Adduri et al., 2026) uses a transformer to model perturbation responses across sets of cells. The generative baselines include scDFM (Yu et al., 2026), which combines conditional flow matching with graph-informed attention and distributional regularization, and CellFlow (Klein et al., 2025), which learns conditional transformations between control and perturbed cell populations through flow matching.

Table 1: Norman Additive: held-out pair prediction, averaged across three folds. Bold/underlined values are best/second best by unrounded scores; N.A. denotes undefined control correlations.

Model	L2 ↓	MSE ↓	MAE ↓	Pearson DE ρ ↑	Pearson Δ ↑	DS L1 ↑
Control	4.4089	0.0215	0.0371	N.A.	N.A.	0.5333
GEARS	2.9383	0.0096	0.0461	0.8518	0.7734	0.8578
CPA	4.1791	0.0195	0.0667	0.6436	0.4792	0.6533
scGPT	1.8156	0.0044	0.0191	0.9005	0.8853	0.8933
STATE	<u>1.7054</u>	<u>0.0038</u>	<u>0.0172</u>	0.9074	<u>0.9128</u>	0.9111
scDFM	2.0847	0.0047	0.0276	0.8641	0.8740	<u>0.9200</u>
CellFlow	1.8923	0.0046	0.0181	<u>0.9150</u>	0.8817	0.8830
CellKAFC (ours)	**1.5105**	**0.0030**	**0.0156**	**0.9363**	**0.9257**	**0.9304**

Table 2: RPE1 single-gene perturbation prediction. The adopted CellKAFC model uses expression-graph and ESM2 priors; all six benchmark metrics are reported.

Model	L2 ↓	MSE ↓	MAE ↓	Pearson DE ρ ↑	Pearson Δ ↑	DS L1 ↑
Control	6.3567	0.0480	0.1351	N.A.	N.A.	0.5013
GEARS	5.3563	0.0316	0.1190	0.6776	0.6156	0.5747
CPA	6.3641	0.0478	0.1338	0.4478	0.2326	0.5013
scGPT	5.1757	0.0300	0.1087	0.7463	0.6547	0.4948
STATE	<u>4.9584</u>	<u>0.0265</u>	<u>0.1042</u>	<u>0.7524</u>	<u>0.6714</u>	0.5013
scDFM	5.9139	0.0375	0.1288	0.6332	0.5644	0.5191
CellFlow	5.3943	0.0337	0.1118	0.7078	0.5942	**0.5916**
CellKAFC (ours)	**4.5632**	**0.0233**	**0.0964**	**0.7589**	**0.6914**	<u>0.5750</u>

Metrics. We use six metrics to evaluate predicted condition-mean responses on the 1,000-gene evaluation panel. Euclidean distance (L2), mean squared error (MSE) and mean absolute error (MAE) measure discrepancies between predicted and observed mean expression. Pearson Δ measures correlation between their control-relative mean shifts. DE Spearman measures rank correlation between predicted and observed log fold changes over significantly differentially expressed genes. DS L1 evaluates perturbation discrimination by ranking observed conditions according to L1 distances between absolute predicted and observed mean shifts. Lower values indicate better performance for L2, MSE and MAE; higher values are better for the remaining metrics. Formal definitions are provided in Appendix C.

4.2 Prediction across genetic and chemical perturbations

Combinatorial perturbation prediction (Norman Additive). Our model achieved the best performance on all six metrics for held-out gene pairs, reducing MSE by 22.5% relative to STATE (Table 1). The improvements in DE Spearman and DS L1 help explain this advantage. Higher DE Spearman indicates better agreement with the observed fold-change rankings among differentially expressed genes. Higher DS L1 indicates more accurate matching between predicted responses and the correct perturbations. Together, these results show better recovery of both expression levels and the response patterns of unseen combinations. Our model retained this advantage even when some constituent single-gene responses were unavailable during training.

Unseen perturbation prediction (Norman Holdout). Norman Holdout tests interventions involving genes excluded from training perturbations. Our model achieved the lowest L2, MSE and MAE for both singles and doubles (Table 3). MSE decreased by 20.4% relative to STATE on singles and by 8.5% relative to CellFlow on doubles. These gains support the model's ability to predict expression levels for interventions beyond those observed during training. For doubles, our model also achieved the highest Pearson Δ and DS L1. The higher correlation shows better agreement in expression changes relative to control, while the higher discrimination score shows better matching to the target condition. Thus, the advantage on unseen combinations extends from expression accuracy to recovery of their specific perturbation responses.

Single-gene perturbation prediction (RPE1). RPE1 evaluates 386 held-out single-gene interventions in a second cellular context, with models trained separately on this dataset. Our model was best on five of the six metrics and reduced MSE by 12.1% relative to STATE (Table 2). Pearson Δ increased from 0.6714 to 0.6914. Because this metric compares changes relative to control, the increase shows better recovery of perturbation effects across genes. The highest DE Spearman further indicates better agreement in fold-change rankings among responding genes. These results address a key concern about latent-space prediction: retaining gene-level response information after decoding. Our framework used a constrained transport space while achieving accurate expression levels and differential-response patterns in its final predictions.

Drug perturbation prediction (COM). COM extends the evaluation to drug perturbations, using molecular and drug-associated features to describe the interventions. Across seven held-out conditions,

Table 3: Norman Holdout: single- and double-perturbation results, averaged across three folds and ranked separately within each stratum.

Single perturbations

Model	L2 ↓	MSE ↓	MAE ↓	DE ρ ↑	Pearson Δ ↑	DS L1 ↑
Control	3.2747	0.0119	0.0305	N.A.	N.A.	0.5754
GEARS	3.5795	0.0134	0.0565	**0.7856**	0.4994	0.6619
CPA	4.2687	0.0190	0.0710	0.2357	0.0651	0.5661
scGPT	3.5973	0.0140	0.0341	0.5634	0.3366	0.5482
STATE	3.1375	0.0102	0.0320	0.7120	0.4535	0.5915
scDFM	3.3034	0.0119	0.0394	0.7325	0.3906	0.5822
CellFlow	2.9374	0.0119	0.0279	0.6914	**0.6013**	**0.8288**
CellKAFC (ours)	**2.7376**	**0.0081**	**0.0255**	0.7517	0.5278	0.6410

Double perturbations

Model	L2 ↓	MSE ↓	MAE ↓	DE ρ ↑	Pearson Δ ↑	DS L1 ↑
Control	4.4249	0.0219	0.0409	N.A.	N.A.	0.5195
GEARS	3.3725	0.0124	0.0522	0.8323	0.6853	0.7831
CPA	5.1714	0.0289	0.0904	0.4874	0.2620	0.5717
scGPT	2.7255	0.0091	0.0274	0.7956	0.7387	0.7266
STATE	2.6209	0.0085	0.0269	0.7819	0.7246	0.8097
scDFM	2.8646	0.0093	0.0352	0.8245	0.7112	0.7694
CellFlow	2.3774	0.0067	0.0229	**0.8633**	0.8102	0.8196
CellKAFC (ours)	**2.2943**	**0.0062**	**0.0229**	0.8556	**0.8176**	**0.8246**

Table 4: Ablation results on RPE1, averaged over three training seeds. Each corrected variant fits its own endpoint. Bold/underlined values indicate the best/second-best result in each column.

Variant	L2 ↓	MSE ↓	MAE ↓	DE ρ ↑	Pearson Δ ↑	DS L1 ↑
w/o graph	4.6161	0.0242	0.0978	0.7551	0.6885	0.5789
w/o ESM2	4.6968	0.0244	0.0994	**0.7592**	0.6886	0.5240
w/o endpoint	5.1850	0.0303	0.1115	0.6098	0.6014	0.5109
w/o ICA effect basis	4.6091	0.0237	0.0975	0.7473	0.6780	0.5685
w/o background freezing	4.6910	0.0244	0.0993	0.7421	0.6737	0.5559
w/o flow dynamics	4.6542	0.0247	0.0983	0.7389	0.6701	**0.6116**
CellKAFC (ours)	**4.5632**	**0.0233**	**0.0964**	0.7589	**0.6914**	0.5750

our model achieved the lowest L2, MSE and MAE and the highest DE Spearman (Table 10 in Appendix C). Relative to CellFlow, MSE decreased by 2.6% and DE Spearman increased from 0.9248 to 0.9295. The lower errors show more accurate expression levels across the evaluation panel. The higher DE Spearman adds evidence that the model also recovered the relative ordering of drug-induced fold changes among responding genes. These gains show that the framework remains effective with a different type of perturbation information and delivers strong competitive performance on drug response prediction.

4.3 Ablation study

We examine the motivation for knowledge-conditioned transport and residual mean correction through controlled ablation experiments. All comparisons use the same RPE1 split, training budget and three training seeds (Table 4). For variants that retain residual correction, we refit it using their training-condition residuals. Detailed configurations and results across seeds are provided in Appendix D.

Effectiveness of biological priors. We removed the expression-graph prior and ESM2 embeddings separately to assess their contributions. These removals increased MSE by 3.61% and 4.48%, respectively, with both variants also showing higher L2 and MAE. The increase in MSE occurred across all three seeds, supporting a consistent benefit from combining the two sources. Neither coexpression information nor protein-sequence information alone matched the expression accuracy achieved with both. Removing ESM2 also reduced DS L1 from 0.5750 to 0.5240, indicating that the predicted responses were less accurately matched to their corresponding perturbations. Thus, protein-sequence information contributed to distinguishing perturbation responses as well as reducing expression errors.

Effectiveness of constrained flow matching. Replacing the selected ICA effect coordinates with equal-rank PCA increased MSE by 1.65%. Allowing updates over all 128 ICA coordinates increased MSE by 4.74%, despite covering more observed response energy (82.5% versus 76.3%). Thus, selecting effect directions and freezing background coordinates produced lower error under the same training budget. Within this space, replacing flow dynamics with one-step displacement prediction increased MSE by 5.96% and reduced both DE Spearman and Pearson Δ, indicating less accurate

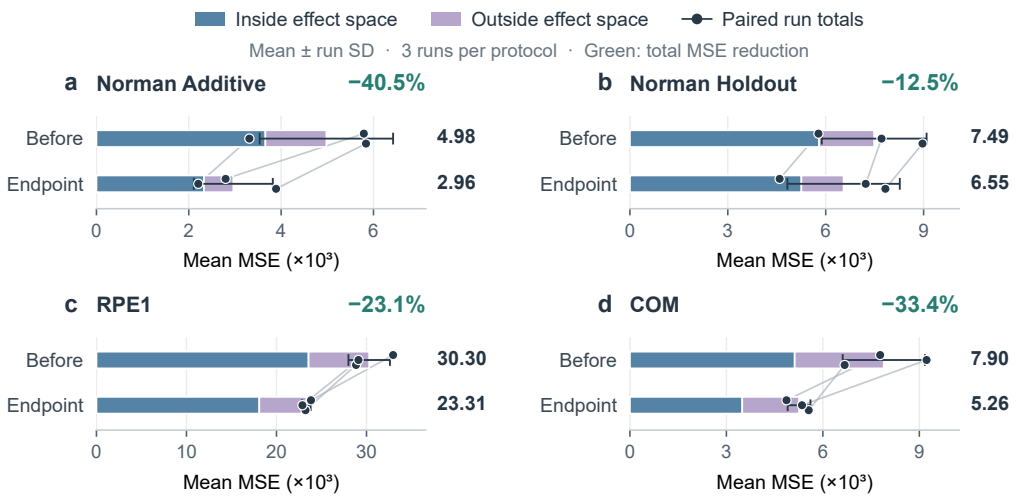


Figure 2: Endpoint mean-error decomposition for the adopted CellKAFC configurations. Stacks show three-run mean error within and outside the effect-decoder span; whiskers are sample SD of total MSE. Dots are run totals and faint lines connect matching runs. Right labels give total means and green labels give their percentage reduction. Norman uses three folds; RPE1/COM use three fixed-split seeds. Values are scaled by 1,000; axis ranges differ. Holdout combines singles and doubles for this decomposition.

expression levels and response patterns. Even with residual mean correction applied to both models, the flow model still achieved lower expression errors than the one-step predictor.

Effectiveness of residual mean correction. We compared predictions before and after correction while keeping the transport model fixed. Correction reduced RPE1 MSE by 23.1% and improved all six metrics. DE Spearman increased from 0.6098 to 0.7589, showing better recovery of fold-change rankings among differentially expressed genes. Pearson Δ increased from 0.6014 to 0.6914, indicating closer agreement with the observed expression changes relative to control. The gains therefore extended from reducing expression errors to recovering gene-level response patterns. Since the transport model was unchanged, these results show that a separate correction learned from training-condition residuals could reduce errors remaining after transport. Figure 2 extends this comparison to all four protocols, where correction reduced aggregate MSE by 12.5–40.5%. Applying the same condition-specific offset to every predicted cell improved the population mean while preserving differences between cells.

4.4 Case studies of genetic and drug combinations

Figure 3 examines the recovery of genetic and drug combination responses using cases selected near the median gain over CellFlow among eligible conditions.

Genetic combination: DUSP9+SNAI1. CellKAFC more closely captured the pattern of increased and decreased expression induced by this combination (Figure 3c). Our model was closer to the observed response than CellFlow for 19 of the 20 displayed genes, reducing their MAE by 30.0% (from 0.7106 to 0.4977). The improvement included both the induction of GYPB and ALAS2 and the reduction of CKS1B and UBE2S, showing that the gain extended across genes responding in opposite directions. Although the strongest induced responses remained underestimated, the agreement across individual genes supports more accurate reconstruction of this held-out combination response.

Drug combination: Panobinostat+Crizotinib. Both methods overlapped the dominant observed cell cluster in the shared UMAP, while the gene-response profiles distinguished their predictive accuracy (Figure 3b,d). CellKAFC was closer for 15 of the 20 displayed gene responses and reduced the average MAE across seeds by 18.0%. The improvement included increased PEG10 expression

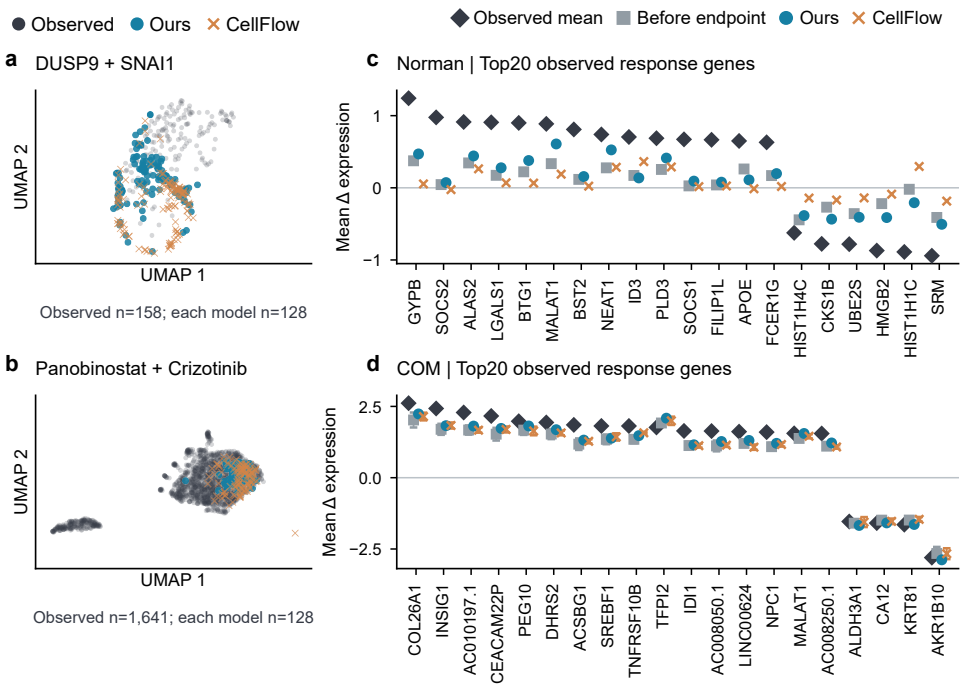


Figure 3: Combination response cases for CellKAFC and CellFlow. (a,b) Shared-reference UMAP with all 158/1,641 observed cells and 128 predictions per model; controls define the reference but are not plotted. (c,d) Top20 observed-response genes, including predictions before endpoint. Norman: one fold. COM: representative-seed UMAP and multi-seed response mean ± sample SD. Separate legends identify each series. Case selection and gene-level aggregation are detailed in Appendix F.

and decreased KRT81 and CA12 expression, indicating more accurate recovery of both induced and repressed transcripts. These differences in response amplitude were less apparent from the broad population overlap in the two-dimensional projection. The gene-level comparison thus provides quantitative support for CellKAFC's more faithful reconstruction of this drug-combination response.

Together, these cases link the benchmark gains to improved recovery of individual responses under genetic and drug perturbations. UMAPs provide qualitative population context. Case selection and interpretation details are given in Appendix F.

5 Conclusion

We presented CellKAFC, a framework for predicting single-cell responses to genetic and drug perturbations. By combining biological priors with constrained flow matching and residual mean correction, it models cell-dependent responses while addressing mean-expression errors left after transport. Across Norman, RPE1 and COM, CellKAFC achieved strong expression accuracy, supporting the effectiveness of this design in the evaluated settings.

The study has two main limitations. First, model configurations were selected using the reported benchmarks, so their performance on untouched datasets remains to be tested. Second, the drug evaluation covers few held-out conditions, and residual correction does not improve every query. Future work will evaluate fixed configurations on larger independent screens and examine when correction transfers reliably.

AI USE STATEMENT

OpenAI Codex assisted with figure preparation, literature checking, and manuscript drafting and editing. The authors remain responsible for the accuracy of the reported results, citations, and conclusions.

REPRODUCIBILITY STATEMENT

Appendix B documents the dataset partitions, preprocessing, model and training settings, and residual correction procedure; Appendix C defines the evaluation metrics and aggregation. We will provide the source code with the submission. Reproducing the reported results also requires the ordered condition and gene manifests and frozen prior assets.

ETHICS STATEMENT

This study uses publicly available single-cell perturbation datasets. No new human or animal samples were collected, and no personally identifiable or patient-level information was analyzed. CellKAFC is intended for computational research on perturbation responses. Its predictions should not be used for clinical decisions or interpreted as evidence of drug efficacy or synergy without independent experimental validation.

REFERENCES

Abhinav K. Adduri, Dhruv Gautam, Beatrice Bevilacqua, Mohsen Naghipourfar, Alishba Imran, Rohan Shah, Noam Teyssier, Rishi Verma, Christopher Carpenter, et al. Predicting cellular responses to perturbation across diverse contexts with State. *Cell*, 189(19):5914–5931.e20, 2026. doi: 10.1016/j.cell.2026.07.052. URL https://doi.org/10.1016/j.cell.2026.07.052.

Michael Bereket and Theofanis Karaletsos. Modelling Cellular Perturbations with the Sparse Additive Mechanism Shift Variational Autoencoder. In *Advances in Neural Information Processing Systems 36*, pp. 1–12, 2023. doi: 10.52202/075280-0001. URL https://doi.org/10.52202/075280-0001.

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. doi: 10.1038/s41592-023-01969-x. URL https://doi.org/10.1038/s41592-023-01969-x.

Sourav Chatterjee. A new coefficient of correlation. *Journal of the American Statistical Association*, 116(536):2009–2022, 2021. doi: 10.1080/01621459.2020.1758115. URL https://doi.org/10.1080/01621459.2020.1758115.

Seyone Chithrananda, Gabriel Grand, and Bharath Ramsundar. ChemBERTa: Large-Scale Self-Supervised Pretraining for Molecular Property Prediction. *arXiv preprint arXiv:2010.09885*, 2020. URL https://arxiv.org/abs/2010.09885.

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. doi: 10.1038/s41592-024-02201-0. URL https://doi.org/10.1038/s41592-024-02201-0.

Mingze Dong, Bao Wang, Jessica Wei, Antonio H. de O. Fonseca, Curtis J. Perry, Alexander Frey, Feriel Ouerghi, Ellen F. Foxman, et al. Causal identification of single-cell experimental perturbation effects with CINEMA-OT. *Nature Methods*, 20(11):1769–1779, 2023. doi: 10.1038/s41592-023-02040-5. URL https://doi.org/10.1038/s41592-023-02040-5.

Arthur Gretton, Karsten M. Borgwardt, Malte J. Rasch, Bernhard Sch"olkopf, and Alexander Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13(25):723–773, 2012. URL https://www.jmlr.org/papers/v13/gretton12a.html.

Qun Jiang, Shengquan Chen, Xiaoyang Chen, and Rui Jiang. scPRAM accurately predicts single-cell gene expression perturbation response based on attention mechanism. *Bioinformatics*, 40(5): btae265, 2024. doi: 10.1093/bioinformatics/btae265. URL https://doi.org/10.1093/bioinformatics/btae265.

Dominik Klein, Jonas Simon Fleck, Daniil Bobrovskiy, Lea Zimmermann, Sören Becker, Alessandro Palma, Leander Dony, Alejandro Tejada-Lapuerta, et al. CellFlow enables generative single-cell phenotype modeling with flow matching. *bioRxiv*, 2025. doi: 10.1101/2025.04.11.648220. URL https://doi.org/10.1101/2025.04.11.648220.

Zeming Lin, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Nikita Smetanin, Robert Verkuil, et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379(6637):1123–1130, 2023. doi: 10.1126/science.ade2574. URL https://doi.org/10.1126/science.ade2574.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow Matching for Generative Modeling. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/pdf?id=PqvMRDCJT9t.

Mohammad Lotfollahi, Anna Klimovskaia Susmelj, Carlo De Donno, Leon Hetzel, Yuge Ji, Ignacio L Ibarra, Sanjay R Srivatsan, Mohsen Naghipourfar, et al. Predicting cellular responses to complex perturbations in high-throughput screens. *Molecular Systems Biology*, 19(6):e11517, 2023. doi: 10.15252/msb.202211517. URL https://doi.org/10.15252/msb.202211517.

Leland McInnes, John Healy, Nathaniel Saul, and Lukas Großberger. UMAP: Uniform manifold approximation and projection. *Journal of Open Source Software*, 3(29):861, 2018. doi: 10.21105/joss.00861. URL https://doi.org/10.21105/joss.00861.

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. doi: 10.1126/science.aax4438. URL https://doi.org/10.1126/science.aax4438.

Zoe Piran, Niv Cohen, Yedid Hoshen, and Mor Nitzan. Disentanglement of single-cell data with biolord. *Nature Biotechnology*, 42(11):1678–1683, 2024. doi: 10.1038/s41587-023-02079-x. URL https://doi.org/10.1038/s41587-023-02079-x.

Reactome Consortium. Reactome pathway gene sets, 2026. URL https://reactome.org/download-data. GMT snapshot accessed 15 September 2026.

Joseph M. Replogle, Reuben A. Saunders, Angela N. Pogson, Jeffrey A. Hussmann, Alexander Lenail, Alina Guna, Lauren Mascibroda, Eric J. Wagner, et al. Mapping information-rich genotype-phenotype landscapes with genome-scale Perturb-seq. *Cell*, 185(14):2559–2575.e28, 2022. doi: 10.1016/j.cell.2022.05.013. URL https://doi.org/10.1016/j.cell.2022.05.013.

David Rogers and Mathew Hahn. Extended-Connectivity Fingerprints. *Journal of Chemical Information and Modeling*, 50(5):742–754, 2010. doi: 10.1021/ci100050t. URL https://doi.org/10.1021/ci100050t.

Yusuf Roohani, Kexin Huang, and Jure Leskovec. Predicting transcriptional outcomes of novel multigene perturbations with GEARS. *Nature Biotechnology*, 42(6):927–935, 2024. doi: 10.1038/s41587-023-01905-6. URL https://doi.org/10.1038/s41587-023-01905-6.

Jiafa Ruan, Ruijie Quan, Liyang Xu, Zongxin Yang, and Yi Yang. Beyond Independent Genes: Learning Module-Inductive Representations for Single-Cell Gene Perturbation Prediction. *arXiv preprint arXiv:2602.04901v2*, 2026. doi: 10.48550/arXiv.2602.04901. URL https://arxiv.org/abs/2602.04901v2.

Aravind Subramanian, Rajiv Narayan, Steven M. Corsello, David D. Peck, Ted E. Natoli, Xiaodong Lu, Joshua Gould, John F. Davis, et al. A Next Generation Connectivity Map: L1000 Platform and the First 1,000,000 Profiles. *Cell*, 171(6):1437–1452.e17, 2017. doi: 10.1016/j.cell.2017.10.049. URL https://doi.org/10.1016/j.cell.2017.10.049.

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. URL https://arxiv.org/abs/2602.07103v1.

Hengshi Yu, Weizhou Qian, Yuxuan Song, and Joshua D. Welch. PerturbNet predicts single-cell responses to unseen chemical and genetic perturbations. *Molecular Systems Biology*, 21(8): 960–982, 2025. doi: 10.1038/s44320-025-00131-3. URL https://doi.org/10.1038/s44320-025-00131-3.

A GEOMETRY OF THE CONSTRAINED PREDICTOR

Fixed-space mean error. All statements in this section condition on the fitted decoder and the actual source-control population. For a specified gene panel, write $\widehat{x}^0_{i,p} = x_i + Ba_{i,p}$ and $S = \mathrm{col}(B)$. If Q has orthonormal columns spanning S, then $P = QQ^\top$, $P^2 = P$, and $(I - P)B = 0$. Averaging cells gives $(I - P)\widehat{\mu}^0_p = (I - P)\mu_X$. Orthogonality of P and $I - P$ then proves Equation 10. For any predictor in this fixed update class, MSE is at least $G^{-1} \left\|(I - P)(\mu_X - \mu_p)\right\|^2$. Attainability of this lower bound requires sufficient within-space estimation capacity and is not assumed. A different source distribution, decoder or gene panel changes the bound.

Expression geometry rather than ICA-coordinate energy. The evaluation decoder is obtained by selecting the 1,000 ordered evaluation genes before SVD. Singular directions above 10^{-10} times the largest singular value define Q. This is an orthonormal basis in the unstandardized preprocessed expression metric used for MSE. The outside-space component can include directions corresponding to frozen ICA coordinates and directions excluded by dimensionality reduction. It should not be identified with a particular background coordinate or biological process. In contrast, distribution diagnostics divide each gene by its training standard deviation before computing distances.

Mean supervision and residual support. On any training panel with a fixed decoder, the outside-space term has zero gradient with respect to the dynamic predictor's coefficients. A generated-mean loss can change within-space fitting without expanding the update support. Endpoint fitting adds a different function class. Before its scalar gate and norm cap, its output is a linear combination of observed training residual vectors. Consequently, $r(p) \in \mathrm{span}\{r_j\}$, including after a whole-vector scalar rescaling or zeroing. The correction can extend beyond S but need not span arbitrary expression directions.

Translation invariance and beneficial correction. With a fixed condition offset $r(p)$, subtracting the corresponding predicted mean cancels that offset, proving the centered-translation identity in Section 3. Every centered empirical moment and internal pairwise distance is therefore unchanged for the same cells. In the outside component, write $\epsilon_\perp = (I - P)(\widehat{\mu}^0_p - \mu_p)$ and $r_\perp = (I - P)r(p)$. The decrease in outside squared error is

$$G(E^0_\perp - E^{\mathrm{end}}_\perp) = -2\epsilon_\perp^\top r_\perp - \|r_\perp\|^2 . \tag{11}$$

It is positive only when correction direction and magnitude are sufficiently appropriate. Norm clipping alone cannot guarantee improvement. An outside-space improvement also does not guarantee that the inside-space error will improve.

Projected-output diagnostic. Replace $r(p)$ by $Pr(p)$ after fitting the complete endpoint, while leaving the predictor and every fitted parameter unchanged. Its inside error is identical to the full endpoint, whereas its outside error equals the pre-endpoint error. Its total MSE is therefore $E^{\mathrm{end}}_\parallel + E^0_\perp$. Table 13 is calculated from this identity using frozen condition-level errors. No constrained calibration model is trained or tuned in this comparison.

B IMPLEMENTATION AND ADOPTED CONFIGURATIONS

B.1 DATASETS AND SPLITS

We use condition-level partitions of Norman K562, RPE1 and ComboSciPlex A549 (COM). Each dataset is fitted separately under its recorded protocol. Table 5 summarizes the perturbed-condition counts; control cells are not counted as perturbation conditions. The Norman results average multiple evaluation folds, using a different training seed for each fold. RPE1 and COM each use one fixed split with multiple training seeds.

Table 5: Dataset partitions used in the reported experiments. Counts refer to the archived condition labels. Separate validation counts are shown where defined by the fixed split; RPE1 validation conditions are excluded from model fitting.

Protocol	Run	Train	Validation	Test singles	Test pairs
Norman Additive	A	62	–	0	15
	B	62	–	0	15
	C	62	–	0	15
Norman Holdout	A	46	–	6	25
	B	43	–	7	27
	C	45	–	7	25
RPE1	Fixed	1,041	116	386	0
COM	Fixed	24	–	2	5

Norman Additive. The Additive protocol withholds gene-pair conditions from training. It does not require both constituent single-perturbation responses to be available in the fitted training set. Across the 45 test condition–fold records, both single references are present for 14 records. The evaluation therefore concerns prediction of held-out combinations under incomplete single-reference coverage.

Norman Holdout. The Holdout protocol excludes the held-out genes from training perturbation conditions and evaluates their single perturbations and combinations. Metrics are computed separately for the single- and double-perturbation groups. We retain the archived condition labels, including the two gene/control orderings for single perturbations, to preserve the original evaluation protocol.

RPE1. The fixed simulation split uses an initial training fraction of 0.75 before the train/validation partition. The final disjoint sets contain 1,041 training, 116 validation and 386 test single-gene conditions. The validation conditions are not included in the fitted training set, and the same test conditions are used for multiple training seeds.

COM. The fixed chemical-perturbation split contains 24 training and seven test conditions. The held-out set comprises two single-drug conditions and five drug pairs. Multiple training seeds use this same partition, and the reported benchmark metrics aggregate the seven test conditions.

B.2 Architecture and preprocessing

Preprocessing and inference. Expression panels use approximately 5,000 highly variable genes, with target-gene inclusion where required by the data loader. The Additive panels contain 5,014, 5,015 and 5,014 genes across folds. Standardization, PCA, ICA and expression-graph assets use their training expression. The generated-mean loss acts on the training expression panel; endpoint fitting and reported evaluation concern the ordered 1,000-gene output panel. Full reproduction requires the stored ordered genes, conditions and source-control indices, beyond these dimensional summaries.

The numerical flow uses updates $z_{k+1} = z_k + K^{-1} v_\theta(z_k, (k + 1/2)/K, b, c)$ for $k = 0, \ldots, K - 1$. This is an Euler state update with midpoint time evaluation. The differentiable training rollout follows the same discrete update. Its mean-loss weight reaches the maximum at the final selected RPE1 step. There is no trained Pair contribution in RPE1: its padded single-condition input yields exactly zero output.

Genetic priors. The graph contains top-30 coexpression neighborhoods, 32 spectral coordinates and 232 hashed neighborhood features. ESM2 features come from full-protein representations of the 3B-parameter model. The adapter projects to 256 dimensions; an implementation compatibility layout with inactive channels does not constitute additional biological sources. Original router proportions in the scGPT/graph/ESM2 order are $(0.65, 0.30, 0.05)$ for Norman and $(0.55, 0.30, 0.15)$ for RPE1. The adopted Additive mask gives effective scGPT/ESM2 initialization $(13/14, 1/14)$; the RPE1 mask gives graph/ESM2 $(2/3, 1/3)$. Disabled branches are removed from the condition, routing and pair-feature information paths. Inactive parameter slots may remain for checkpoint compatibility.

Table 6: Shared architecture and optimization settings. Frozen prior features are distinct from trainable adapters and routing.

Item	Setting
Velocity network	6 conditional residual blocks, width 512
Condition / time embedding	256 / 64 dimensions
Normalization / activation	Layer normalization / SiLU
Conditional modulation	Feature-wise scale and shift
Flow dropout	0.05
Optimizer	AdamW; learning rate 2×10^{-4}
Weight decay / gradient clipping	10^{-5} / 1.0
Batch size	256
MMD weight	0.02; single RBF scale
Generated-mean sampling	Every 5 steps; 4 conditions; 32 controls each
Balanced OT regularization	0.1 after condition cost normalization
Pair bottleneck / gate initial / cap	8 / 0.05 / 0.20
Prediction cells per condition	128 source controls

Table 7: Protocol-specific settings. Additive and RPE1 use their adopted dual-source configurations; holdout and COM retain three sources. Router proportions are effective initial values when all retained sources are available, not trained weights.

Setting	Additive	Holdout	RPE1	COM
PCA / ICA dimensions	128 / 128	128 / 128	128 / 128	64 / 64
Background / effect	64 / 64	64 / 64	64 / 64	24 / 40
Flow training steps	5,000	5,000	500	6,000
ODE / mean rollout steps	24	24	24	36
Mean-loss maximum weight	10	10	10	3
Mean warmup steps	500	500	500	250
Prior projection dropout	0.05	0.05	0.05	0.10
Expert dropout setting	0.05	0.05	0.05	0
Pair fitting steps	500	500	0	250
Pair learning rate	10^{-3}	10^{-3}	–	10^{-3}
Pair weight decay	10^{-3}	10^{-3}	–	10^{-4}
Pair dropout	0.05	0.05	–	0
Pair gene / cosine / L1 weights	1/.05/.001	1/.05/.001	–	.5/0/0

Availability masking also affects vector scale references. The RPE1 no-scGPT encoder uses a retained source as its scale reference rather than hidden scGPT content. ALG1L has no remaining available source and receives an exactly zero prior mixture. It is not silently assigned the removed source. This defines the behavior of Equation 1 when its nominal denominator vanishes.

Drug priors. ChemBERTa features are 768-dimensional mean-pooled representations. Morgan fingerprints have radius 3 and 512 bits. The biological branch combines independently projected target and external response features as $h_{\text{bio}} = (1 - g)h_{\text{target}} + gh_{\text{response}}$, where $g = 0.3\sigma(\eta)$ starts at 0.15 and is zero when response evidence is missing. Targets are represented using frozen scGPT embeddings; target evidence includes curated mechanisms and high-confidence human single-protein binding records. The response prior is a spectral representation of a LINCS-derived similarity kernel, using MCF7, PC3 and VCAP at 24 hours and excluding A549. Target and response coverage are 17/17 and 15/17 compounds. This is an adapted external-response feature, internally named BiRD-LINCS, rather than a claim to reproduce an entire external model. Detailed source accession and target-selection manifests remain part of the research assets needed for release. The initial ChemBERTa/Morgan/biology proportions are $(4/7, 8/35, 1/5)$; no trainable drug-identity embedding is added.

Table 8: Endpoint fits associated with the adopted predictions. Values are read from the current configuration and completion records.

Protocol	Run	τ	Ridge	β
Additive	Run A	0.8	0.1	0.981539
	Run B	0.8	0.1	1.003365
	Run C	0.8	0.1	1.064485
Holdout	Run A	0.8	0.1	1.019290
	Run B	0.4	0.001	1.039256
	Run C	0.8	0.1	0.997031
RPE1	Run A	0.8	10	1.057563
	Run B	0.8	10	1.072683
	Run C	0.8	10	1.051126
COM	Run A	0.4	0.1	1.017498
	Run B	0.4	0.1	1.065166
	Run C	0.8	1	1.143324

Pair offset. For two active components, symmetric features include the component sum, absolute difference and elementwise product. A rank-8 head with $\tanh$ output and empirical effect scaling is multiplied by a learned gate capped at 0.2. The gate starts at 0.05. The head is fitted to training double-condition residual means after flow freezing. It adds a condition offset within S and does not identify a unique biological interaction mechanism.

B.3 Endpoint fitting and safeguards

For candidate bandwidth and ridge, let $A = (H + \lambda I)^{-1} R$ and $C = (H + \lambda I)^{-1}$. The reference leave-one-out prediction is $\widehat{r}_{j,-j} = r_j - A_j / C_{jj}$. Its fitted scale is

$$\beta = \mathrm{clip}_{[0,1.5]} \left(\frac{\sum_j \widehat{r}_{j,-j}^{\top} r_j}{\sum_j \left\| \widehat{r}_{j,-j} \right\|^2} \right), \tag{12}$$

with a numerical guard for a zero denominator. The bandwidth grid is $\{0.05, 0.1, 0.2, 0.4, 0.8\}$ and the ridge grid is $\{0.001, 0.01, 0.1, 1, 10\}$. Candidates are ranked lexicographically by residual MSE, MAE and negative row-wise Pearson correlation. The similarity threshold is the 10th percentile of training references' maximum cosine similarity to another reference. Correction requires sufficient query similarity and a positive leave-one-reference-out improvement over zero correction. Its norm is capped at 1.5 times the 95th percentile of training residual norms. These safeguards do not establish statistical coverage or per-condition improvement.

C Metrics, data checks, and analysis coverage

Six benchmark metrics. For predicted and observed means on the 1,000 evaluation genes, L2 is $\left\| \widehat{\mu}_p - \mu_p \right\|_2$, MSE is $G^{-1} \left\| \widehat{\mu}_p - \mu_p \right\|_2^2$, and MAE is $G^{-1} \left\| \widehat{\mu}_p - \mu_p \right\|_1$. Each is averaged separately, so the reported average L2 need not equal the square root of G times average MSE. Pearson Δ correlates predicted and observed control-relative mean shifts. DE Spearman correlates predicted and observed log fold changes on genes significant in the observed data, using the evaluator's 0.05 FDR threshold. These observed reference statistics are scoring targets, not training responses for query conditions.

The inspected DS L1 implementation ranks real perturbations by L1 distance between the absolute predicted and real mean shifts. For zero-based rank r_p among M evaluated conditions, the score is $1 - r_p / M$. Target-gene exclusion follows the evaluator's label-matching rules. DS L1 is a mean-response retrieval score, not signed direction accuracy or distribution fidelity. The historical evaluation environment records cell-eval 0.5.42, NumPy 2.2.6, SciPy 1.16.1, Scanpy 1.11.4 and AnnData 0.12.1; exact environment records must accompany the corresponding runs.

Additional diagnostics. Response cosine uses the predicted and real mean shifts relative to the same real control mean. It is distinct from Pearson Δ. Top20 MAE selects the 20 genes with the largest

Table 9: Coverage of paired frozen-prediction analyses. A condition-run record is not an independent biological replicate. External models here are a subset of those in the complete benchmark tables.

Protocol	Records	Distinct conditions	Paired external models
Additive	45	35	CellFlow, scDFM
Holdout	97	54	CellFlow, scDFM
RPE1	1,158	386	scDFM
COM	21	7	CellFlow, CPA

absolute observed mean shifts within the 1,000-gene panel; those genes are not necessarily the evaluator's significant DE set. Response energy retention is calculated per condition as $\|P\Delta_p\|^2 / \|\Delta_p\|^2$, then averaged within runs and across runs. It is a descriptive reachability measure, not a quality objective. Condition plots average repeated appearances of the same condition over available runs, whereas benchmark tables average conditions within each run first.

Checks and precision. Prediction genes and condition identities were aligned with common real reference populations before analysis. An independent calculation checked 19,978 mean-MSE rows, with maximum discrepancy 9.97×10^{-17}; orthogonal decomposition errors summed to MSE within 1.13×10^{-16}. The maximum departure from a uniform endpoint shift was 4.81×10^{-7}. These checks establish numerical consistency of the archived outputs, not an independent training reproduction. Current adopted metrics and completed RPE1 GEARS/scDFM results use full-precision records; other benchmark values retain the available reported precision. COM MAE 0.02815476867 is rounded to 0.02815. No missing seed-level baseline statistics are reconstructed from rounded means.

Aggregation and reporting. Metrics are averaged over test conditions within each run and then across runs. The adopted models use multiple Norman evaluation folds with a different training seed for each fold; RPE1 and COM use multiple training seeds on fixed splits. Thus, Norman has one training seed per fold, whereas RPE1 and COM share their test conditions across seeds. Run standard deviations describe fold or training variation; cells and repeated subsamples are not treated as independent experimental replicates. Bold and underlined entries identify the best and second-best available point estimates within each reporting stratum. Each benchmark table includes the models with available results for that protocol. Configuration development used these benchmarks, so the reported results characterize the selected configurations rather than an untouched test evaluation. STATE is the lowest-MSE external comparator for Additive and RPE1 in the complete benchmark tables, although their available paired prediction analyses use other models.

Table 10: COM drug perturbation prediction over seven held-out conditions. Results include all methods with available matched local evaluations.

Model	L2 ↓	MSE ↓	MAE ↓	DE ρ ↑	Pearson Δ ↑	DS L1 ↑
Control	8.7747	0.0890	0.1114	N.A.	N.A.	0.5714
scGPT	2.4909	0.0071	0.0338	0.9032	0.9139	**0.8707**
CPA	8.2034	0.0776	0.1051	0.5883	0.4191	0.5714
STATE	2.5308	0.0073	0.0334	0.9166	<u>0.9178</u>	<u>0.8639</u>
CellFlow	<u>2.2487</u>	<u>0.0054</u>	<u>0.0285</u>	<u>0.9248</u>	**0.9371**	0.8571
CellKAFC (ours)	**2.1943**	**0.0053**	**0.0282**	**0.9295**	0.9086	0.8503

D Additional ablation and correction analyses

D.1 Component configurations and additional comparisons

Table 4 evaluates interventions on the adopted graph+ESM2 configuration. Each of the five modified predictors is trained for 500 steps with multiple training seeds on the same split of 1,041 training and 386 test conditions. The original control-source pairing, training draws, background coordinates

and 128 prediction cells per condition are retained. The full-model implementation reproduces the adopted weights and training histories for all training seeds. The no-endpoint row uses that same frozen model before correction, without a separate training run. Pair outputs are identically zero for RPE1 single-gene conditions and do not define an additional ablation.

Prior and structural interventions. All new variants keep scGPT disabled. Removing graph leaves only ESM2; removing ESM2 leaves only graph. Deletion masks block the corresponding prior throughout feature encoding, availability flags, routing inputs and scale normalization. Unused parameter slots are retained to preserve initialization and random draws. Input-invariance checks verify that changing a disabled prior cannot affect condition embeddings or residual features. Since the condition representation also defines endpoint similarities, these interventions measure the joint effect on dynamic prediction and residual transfer.

The *w/o ICA effect basis* variant replaces the 64 effect coordinates with equal-rank PCA while retaining the original ICA64 background. The *w/o background freezing* variant expands the effect coordinates to all 128 retained ICA components, including previously frozen directions; some network dimensions change. Neither replacement is a full-gene unconstrained model. The *w/o flow dynamics* variant retains graph+ESM2, the same control source and background, but predicts a displacement at $t = 0$ in one step. It is trained against the coupled displacement with the corresponding population losses, so it changes both prediction and learning rather than only the numerical solver. This one-step formulation is termed Direct in the historical comparisons. These labels describe specific replacements or relaxations, rather than removal of the entire representation or predictor.

Each modified predictor fits its own endpoint from its 1,041 training-condition residuals using the original training-only selection procedure. Test responses are attached for scoring after predictions are frozen. Table 12 reports the paired pre-endpoint and final MSE means and sample SD; the accompanying `dual_ablation_per_seed.csv` retains all six metrics for both stages. Seed variation is descriptive and is not a confidence interval or a significance test.

MSE increased in all three seeds for either prior deletion, *w/o background freezing* and *w/o flow dynamics*, and in two of three seeds for *w/o ICA effect basis*. Metric-specific rankings also differed: the graph-only variant had the highest DE Spearman (0.7592), while *w/o flow dynamics* had the highest DS L1 (0.6116). The main-text interpretation concerns the mean-expression accuracy supported by the matched comparisons.

Historical configuration comparisons. The earlier three-prior study is retained separately in Table 11 and the archived `ablation_six_metrics.csv`. Its reference uses scGPT, graph and ESM2; its graph and ESM2 deletions therefore differ from the current two-source ablations. Relative to that historical reference, the adopted graph+ESM2 configuration reduced aggregate MSE by approximately 0.42% and improved four other metrics. MSE improved in two of three seeds, while aggregate DS L1 changed from 0.5922 to 0.5750. The historical three-prior reference had MSE 0.023412 ± 0.000220, compared with 0.023625 ± 0.000188 for its Direct variant. These comparisons document configuration selection and are not reinterpreted as removals from the current model.

Metric and protocol dependence. In the historical three-prior RPE1 comparison, Direct had the highest DS L1, whereas Flow had lower final MSE. Across the archived protocol comparisons, PCA had lower MSE on Additive and Holdout singles, and Direct on Holdout doubles. In COM, prior deletions improved pre-endpoint MSE, whereas the complete prior set had the lowest final MSE. These comparisons do not identify a universally optimal basis, dimension or prediction rule. RPE1 was selected for the main component display after inspecting the existing benchmarks, where its adopted configuration led five of six metrics. Both that choice and configuration development used the benchmarks; the new matched ablations are not an untouched evaluation. Centered-distribution comparisons of the historical models are reported separately in Appendix E.

Table 11: Historical RPE1 prior-source comparison against scGPT+graph+ESM2, before and after endpoint. Errors are $\times 10^3$. In-space values use the common restricted decoder; pre-endpoint outside-space error is shared up to numerical precision. These source combinations are distinct from the current two-source deletions in Table 4. The CellKAFC row uses graph+ESM2. Bold/underlined values are lowest/second lowest in each column.

Variant	Pre MSE	Final MSE	Pre in	Final in
Three-prior reference	30.7110	<u>23.4120</u>	23.9730	**18.0820**
CellKAFC (ours)	**30.3010**	**23.3140**	**23.5630**	<u>18.0910</u>
scGPT+ESM2	31.1130	23.6760	24.3750	18.3350
scGPT+graph	<u>30.6670</u>	23.5460	<u>23.9290</u>	18.2220

Table 12: Dual-prior RPE1 variants before and after their own endpoint correction. MSE is $\times 10^3$; $\pm$ denotes sample SD over three training seeds on the same split. Reduction compares the two three-seed means.

Variant	Before endpoint	After endpoint	Reduction
CellKAFC (ours)	30.3010 ± 2.3050	23.3140 ± 0.4820	23.1000%
w/o graph	30.7400 ± 3.7680	24.1550 ± 0.9070	21.4000%
w/o ESM2	31.0790 ± 1.9230	24.3580 ± 0.0570	21.6000%
w/o ICA effect basis	32.1550 ± 5.6600	23.6990 ± 0.3210	26.3000%
w/o background freezing	37.4710 ± 4.5960	24.4190 ± 0.4190	34.8000%
w/o flow dynamics	30.6250 ± 2.5420	24.7030 ± 0.4930	19.3000%

D.2 In-space and outside-space correction

Figure 2 separates mean error within and outside the decoded dynamic space for each adopted configuration. Total and outside-space error decreased in all 12 run summaries. Most of the net improvement occurred within the dynamic space, while outside-space reductions accounted for 21.7–42.4% of the gain. Thus, correction both refined already representable mean responses and supplied useful offsets beyond the neural update directions.

Projected-output control. To test whether these additional directions contributed to accuracy, we replaced the fitted correction $r(p)$ with its projection $Pr(p)$ while keeping every fitted parameter unchanged. This preserves exactly the full correction's within-space component. Full correction nevertheless reduced MSE by a further 5.7–18.8% relative to this projected output (Table 13). Because the only removed component is orthogonal to the decoder span, the difference directly measures the realized contribution of expression-space mean offsets unavailable to the fixed dynamic branch. It does not compare separately optimized full-space and restricted calibrators.

Table 13: Projected-output control: MSE $\times 10^3$. Projected keeps only $Pr(p)$ from the fitted full correction; no calibrator is refitted. Outside share is its contribution to net squared-error reduction. Bold/underlined errors are lowest/second lowest within each row.

Protocol	Before	Projected	Full	Outside share
Norman additive	4.9800	<u>3.6500</u>	**2.9650**	34.0000%
Norman holdout	7.4850	<u>6.9450</u>	**6.5480**	42.4000%
RPE1	30.3010	<u>24.8290</u>	**23.3140**	21.7000%
COM	7.8950	<u>6.2640</u>	**5.2600**	38.1000%

D.3 Condition-level effects of correction

After averaging each condition over its available runs, correction improved 34/35 Additive conditions, 40/54 Holdout conditions, 272/386 RPE1 conditions and all seven COM conditions. The Holdout-

single aggregate nevertheless worsened slightly. Condition-level gains therefore do not imply that correction improves every query.

E CENTERED DISTRIBUTION COMPARISONS

Mean accuracy and within-condition distribution shape. The endpoint applies one shared translation per condition, so it cannot change the centered predicted distribution. We therefore compared Flow and Direct after centering each predicted and observed population by its own mean, using the original three-prior reference models. Flow had lower centered empirical energy distance on Norman Additive, Norman Holdout and COM. In RPE1, the preferred model changed between the V statistic and its U-form sensitivity analysis, preventing a stable ranking. Moreover, the COM predicted-to-observed variance ratios were only 0.412 for Flow and 0.397 for Direct, indicating substantial underdispersion in both. These diagnostics separate evidence for cell-dependent transport from improvements attributable to mean translation: Flow provided task-dependent distributional benefits, while the endpoint addressed a different, condition-level error.

Evaluation protocol. Each gene is divided by its training standard deviation, protected below at 10^{-6}. For each condition, equal-size samples of predicted and real cells are drawn without replacement, with $n \leq 128$ and three fixed subsampling repeats. Each sample is centered by its own mean. For centered samples $\{u_i\}_{i=1}^n$ and $\{v_j\}_{j=1}^n$, the primary statistic is

$$V_n = \frac{2}{n^2}\sum_{i,j}\|u_i - v_j\| - \frac{1}{n^2}\sum_{i,j}\|u_i - u_j\| - \frac{1}{n^2}\sum_{i,j}\|v_i - v_j\|\,. \tag{13}$$

The U-form sensitivity replaces the two within-sample terms with sums over unequal indices divided by $n(n-1)$. Because the data have been centered by estimated sample means, this form is not asserted to be an unbiased estimator of a population centered distance. It can be negative, including for real-versus-real samples. RPE1 preferences differ between these two forms, precluding a uniform shape-superiority conclusion.

Sample-size sensitivity. The $n = 32, 64, 128$ sensitivity study compares original three-prior Flow and Direct models and includes disjoint real-versus-real samples. Eligibility requires at least $2n$ real cells. At each n, model comparisons share the eligible conditions and sampled indices. The eligible condition set changes with n, so the curve does not isolate sample size while holding the target population fixed. For example, RPE1 contributes 648, then fewer, and finally 51 eligible condition-run records at $n = 128$. Three subsampling repeats remain nested within the same condition and run.

Variance diagnostics. Variance ratios summarize total predicted-to-real marginal variance, not covariance recovery or subgroup coverage. Large amplitude differences in external reproduced predictions are descriptive findings whose causes require model-version, scale and sampler checks; they are not uniquely attributable to the proposed structure.

F CASE-SELECTION AND BIOLOGICAL INTERPRETATION

Selection protocol. Eligible combination cases have at least 64 observed perturbed cells and response energy at or above the scene median. The displayed cases were selected by gain nearest the eligible median and retained after visualization review. Additive selection is restricted to one evaluation fold (eight eligible candidates); COM to double perturbations (three candidates).

Coordinates and genes. UMAP (McInnes et al., 2018) is fitted on real control and real perturbed cells for the case, then predictions are transformed into that reference. It is a descriptive projection using observed test geometry, not a held-out representation learned solely from training data. All available real perturbed case cells and 128 predictions per model are shown. In Figure 3, 128 controls contribute to the reference embedding but are omitted from the display. Top20 genes are chosen by observed absolute mean response, then ordered by signed response. Norman uses one fold; COM gene summaries use means and sample SD across three seeds with a shared real-control reference. The displayed COM coordinate realization uses one representative training seed.

Displayed-gene error summaries. The main-text case errors use exactly the 20 genes displayed in each panel. For COM, these genes are fixed from the shared observed response used for the plot. We compute each model's absolute error separately for each seed on this fixed set, average over genes, and then average over the three seeds. Counts of genes closer to the observed response instead compare the displayed three-seed mean points. These two operations differ because absolute error is nonlinear. The fixed-panel COM MAEs are 0.305273 (CellKAFC) and 0.372360 (CellFlow); they should not be interchanged with historical per-run Top20 summaries that use run-specific response references or with the error of an averaged prediction. The Norman panel uses only the selected evaluation fold.

Combination scope. Across Additive's 45 condition-fold records, 14 have both observed training single references, nine have one, and 22 have none. The 14 records cover eight distinct combinations. A simple sum of observed single mean responses is available only on that subset; it has average MSE approximately 0.00777 versus 0.00349 for the adopted predictions over those records. The model has lower MSE in 11 of 14 records, but ETS2+MAPK1 favors the observed-single sum in all three folds. This is exploratory and does not represent all Additive conditions. DUSP9 and SNAI1 are both absent from the selected fold's training perturbation vocabulary, and the COM case lacks an observed training Crizotinib single response. Neither displayed combination can establish synergy against a complete observed-single reference.

F.1 Pathway-level response analysis and observed-gene coverage

Figure 4 summarizes the measured gene-set responses for the same genetic and drug combinations examined in Section 4.4. Each point is the mean control-relative expression change over the measured genes in a selected Reactome set. These summaries complement the individual-gene responses in Figure 3.

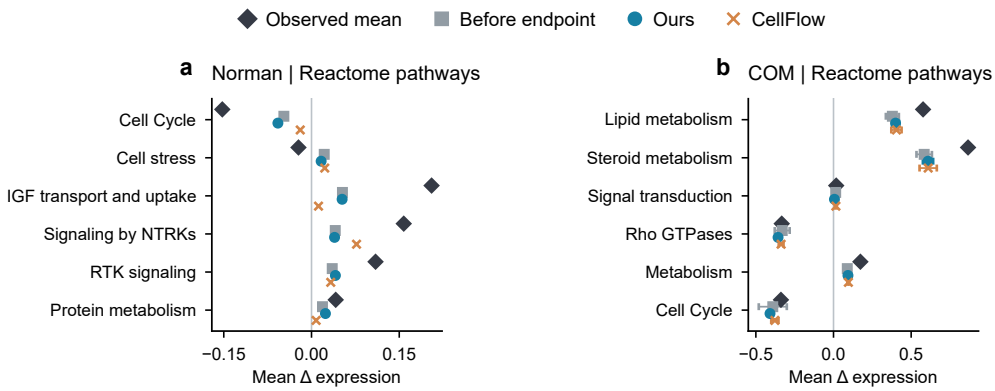


Figure 4: Pathway-level responses for (a) DUSP9+SNAI1 in Norman and (b) Panobinostat+Crizotinib in COM. Diamonds show observed means; squares, circles and crosses show predictions before endpoint, CellKAFC and CellFlow, respectively. Means use all measured members of each selected Reactome set. Norman uses the selected evaluation fold; COM shows multi-seed means with sample SD. Horizontal scales differ between panels. Selection, enrichment q values and measured-gene coverage are detailed below and in Table 14.

Genetic combination. For DUSP9+SNAI1, CellKAFC was closer to the observed mean on five of the six displayed gene sets, including cell cycle and IGF transport and uptake (Figure 4a). CellFlow was closer on NTRK signaling. The improvement across several measured gene sets complements the lower individual-gene error in the main text, although the magnitudes of the strongest positive mean responses remained underestimated.

Drug combination. For Panobinostat+Crizotinib, both methods recovered the directions of all six displayed gene-set means, with CellFlow closer to the observed means in each case (Figure 4b).

This comparison summarizes all measured members of each selected set, whereas the main-text errors assess the 20 genes with the largest observed responses. The two summaries therefore capture different aspects of response recovery: CellKAFC's lower Top20-gene error does not imply uniformly smaller errors after averaging over these gene sets.

Reactome summaries. Human pathway gene sets use the Reactome GMT snapshot downloaded on 15 September 2026 (Reactome Consortium, 2026). The tested universe is the actual 1,000-gene panel. Sets with 10–250 visible genes are eligible. The 100 largest absolute observed responses are separated by sign; hypergeometric enrichment is tested in each direction, followed by Benjamini–Hochberg correction over all tested pathway–direction pairs. Six sets are displayed by observed-response enrichment and a greedy redundancy filter with Jaccard threshold 0.5. Displayed pathway response averages all visible members. Its q value tests enrichment of the observed response, not the accuracy of any model or the significance of the displayed pathway mean. Panel coverage can be low, particularly for Norman and COM, so these summaries do not measure complete biochemical pathway activity. Pathway comparison counts use the displayed mean over all visible genes in each selected set, including genes outside the Top20.

Selection and coverage. Table 14 records every set shown in Figure 4. Enrichment direction refers to the sign of the selected Top100 observed responses used in the enrichment test; it need not equal the sign of the mean over all visible pathway members. For example, COM signal transduction is selected from downregulated genes while its all-visible-gene mean is slightly positive. The display includes four Norman sets with $q = 0.0655$ and does not treat all displayed sets as significant at 0.05. Full Reactome names, membership counts and per-run scores are retained in the accompanying source data.

Table 14: All pathways displayed in Figure 4. Direction and BH-adjusted q refer to observed-response enrichment; coverage is the visible fraction of the full Reactome set. These are not tests of model prediction accuracy.

Case	Display name	Direction	q	Visible/full	Coverage
Norman	Cell cycle	down	0.0364	12/657	1.8000%
Norman	Cell stress	down	0.0364	36/791	4.6000%
Norman	IGF transport	up	0.0655	10/124	8.1000%
Norman	NTRK signaling	up	0.0655	10/134	7.5000%
Norman	RTK signaling	up	0.0655	31/536	5.8000%
Norman	Protein metabolism	up	0.0655	57/2133	2.7000%
COM	Lipid metabolism	up	0.0001	28/759	3.7000%
COM	Steroid metabolism	up	0.0003	12/155	7.7000%
COM	Signal transduction	down	0.0032	88/2622	3.4000%
COM	Rho GTPases	down	0.0032	18/673	2.7000%
COM	Metabolism	up	0.0042	70/2197	3.2000%
COM	Cell cycle	down	0.0090	14/657	2.1000%