scPertWGF: Wasserstein Gradient Flow for Single-Cell Perturbation Prediction

Anonymous authors
Paper under double-blind review

Abstract

Predicting how an unseen genetic or chemical perturbation reshapes a cell population is a central problem in functional genomics. Due to the destructive nature of single-cell sequencing, control and perturbed populations are strictly unpaired. While recent deep generative models strive to capture complex cellular heterogeneity, they often fail to outperform simple linear baselines on mean expression shifts. We propose **scPertWGF**, a **W**asserstein **G**radient **F**low framework for **s**ingle-**c**ell perturbation prediction that explicitly models **pert**urbation effects through changes in mean expression and distributional shape. A closed-form ridge head predicts the population mean shift, while a conditional residual flow captures heterogeneity in cellular responses directly in gene expression space. To guide residual generation, a smoothed Wasserstein gradient flow objective aligns generated and observed populations, with flow matching providing complementary supervision of transport velocities. Among the evaluated baselines, scPertWGF achieves the lowest L_2 prediction error and highest Pearson delta correlation on four genetic perturbation screens and one drug screen, with the lowest energy distance on four, demonstrating its superior capability to capture both mean shifts and complex population shapes.

1 Introduction

Genetic and chemical perturbations are widely used to study gene function, cellular regulation, and drug action (Srivatsan et al., 2020; Norman et al., 2019). Single-cell transcriptomics enables systematic measurement of how perturbations alter gene expression across cell populations (Srivatsan et al., 2020; Replogle et al., 2022). However, the combinatorial explosion of the perturbation space vastly outpaces experimental capacity. Consequently, developing computational models capable of predicting population-level cellular responses to unseen perturbations has emerged as a critical frontier in machine learning for biology (Roohani et al., 2024).

A growing body of work has developed increasingly expressive models for this task. Early methods such as scGen and CPA represent perturbation effects as shifts in a latent space, predicting the expected transcriptional response but offering limited description of variation among cellular states (Lotfollahi et al., 2019; 2023). GEARS (Roohani et al., 2024) incorporates gene relationships and biological knowledge to predict responses to unseen perturbations, yet focuses primarily on the expected response. Foundation models such as Geneformer, scGPT, and scFoundation provide pretrained representations of genes and cells, offering broader biological context without specifying how a perturbation changes the population distribution (Theodoris et al., 2023; Cui et al., 2024; Hao et al., 2024). More recent approaches, including CellOT, CellFlow, and diffusion-based models such as PerturbDiff, move toward direct generation of perturbed populations and can capture variation among cellular states (Bunne et al., 2023; Klein et al., 2025; Yuan et al., 2026). Recent benchmarks, however, show that deep predictors across these families do not consistently outperform deliberately simple linear baselines on metrics based on changes in mean expression (Ahlmann-Eltze et al., 2025; Wei et al., 2026).

These findings highlight the challenge of predicting both population mean expression and distributional shape, with three key issues: (1) **Entangled Optimization of Mean and Cellular Variation:** Perturbations alter both the population mean expression and cellular heterogeneity. Learning them jointly within an unconstrained, high-dimensional space leads to severe optimization entanglement,

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.

making it exceptionally difficult for deep models to simultaneously preserve accurate mean shifts and rich, non-linear population variation. (2) **Difficulty in Aligning Unpaired Populations:** Destructive profiling provides separate samples of control and perturbed cells, leaving individual response trajectories unobserved. Learning must therefore align predicted and observed populations from unpaired samples while preserving the heterogeneity of cellular responses. (3) **Information Loss and High Dimensionality:** Latent space representations may lose subtle and coordinated gene expression patterns that contribute to specific perturbation responses, whereas direct modeling in gene expression space must contend with technical noise and high dimensionality.

To address these challenges, we propose **scPertWGF**, a Wasserstein Gradient Flow framework for single-cell perturbation prediction. The core idea is to separate the perturbation response into a predictable shift in population mean expression and residual cellular variation, and, during training, to constrain the generated population toward the observed perturbed distribution: (1) **Mean Prediction with a Closed-Form Ridge Head:** A regularized ridge head fitted on training population means predicts the robust global shift in mean expression. For unseen perturbations, it utilizes gene response coordinates for genetic interventions and molecular fingerprints with dose information for drugs. (2) **Residual Flow for Population Shape Modeling:** A conditional residual flow transports noise smoothed control cells toward the mean subtracted perturbed state directly in gene expression space, modeling coordinated variation across genes without a latent encoder or decoder. Two complementary objectives guide the residual flow. Flow matching provides local transport supervision, while a smoothed Wasserstein gradient flow objective enforces distributional alignment between the generated and observed populations through multi phase JKO updates. (3) **Mean Guided Inference:** A scalar guidance step calibrates the generated population toward the predicted mean while preserving its covariance structure. Across five perturbation screens, including four genetic and one drug setting, scPertWGF achieves the lowest L2 prediction error and the highest Pearson delta correlation on all five screens, while obtaining the lowest energy distance on four. These results indicate that scPertWGF can accurately predict transcriptional responses to unseen perturbations while preserving the distributional characteristics of the perturbed cell population, providing a basis for in silico prediction of cellular responses before experimental measurement.

Our contributions are summarized as follows:

- **Explicit Mean and Shape Decomposition for Prediction.** We decompose the population response into a predictable mean shift and residual cellular variation, modeled by a closed form ridge head and a conditional residual flow, respectively. A shared translation calibrates the generated population to the predicted mean while preserving its centered distribution.
- **Population-Level Learning from Unpaired Samples.** We combine flow matching with a smoothed Wasserstein gradient flow objective to learn conditional transport between unpaired control and perturbed populations, providing both local transport and population distribution supervision.
- **Direct Residual Modeling in Gene Expression Space.** We model cellular variation directly in raw gene expression space, avoiding information loss from latent compression. Annealed instance noise within the JKO proximal scheme enables the conditional velocity field to capture coordinated variation across genes while reducing overfitting.

2 Related Work

Perturbation Prediction. Perturbation prediction has evolved from latent response modeling toward direct generation of perturbed cell populations. Early methods such as scGen and CPA represent perturbation effects as structured shifts in a latent space, while SAMS-VAE and PerturbNet extend this paradigm with sparse perturbation effects and conditional flow-based generation (Lotfollahi et al., 2019; 2023; Bereket & Karaletsos, 2023; Yu et al., 2025). GEARS instead incorporates gene co-expression and ontology relationships to model genetic interactions and predict responses to unseen perturbations (Roohani et al., 2024). More recently, foundation models including scGPT, Geneformer, scFoundation, and State have introduced pretrained representations that provide broader cellular and biological context for perturbation modeling (Cui et al., 2024; Theodoris et al., 2023; Hao et al., 2024; Adduri et al., 2025). A parallel line of work treats perturbation prediction as distribution generation rather than latent state prediction. CellOT learns optimal transport maps between unpaired control and perturbed populations, while Squidiff and PerturbDiff use diffusion-based genera-

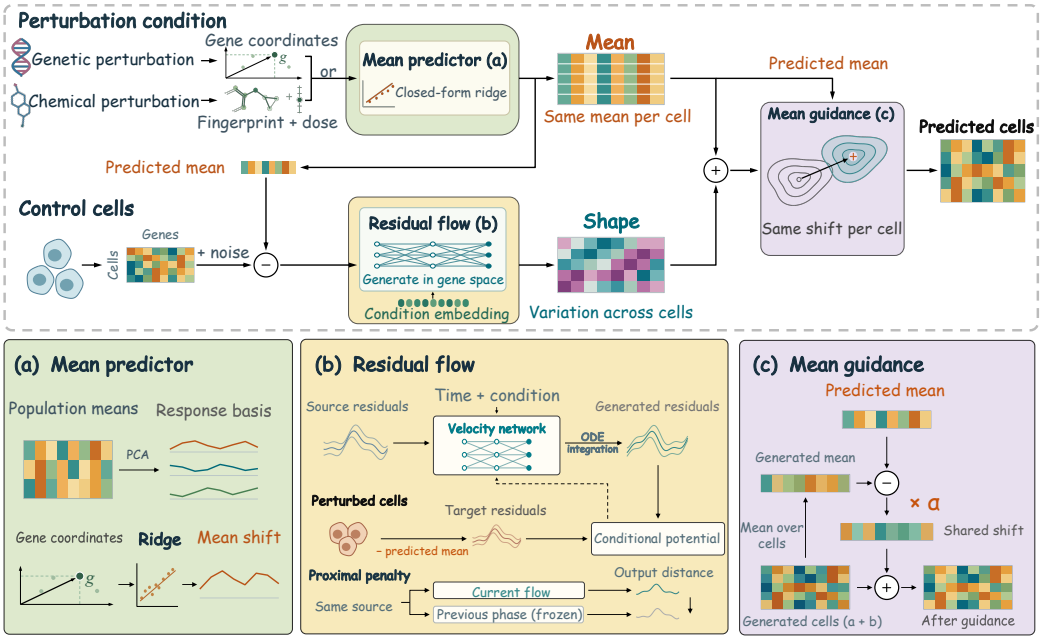

Figure 1: Framework Overview of the Proposed scPertWGF. (a) The closed-form head predicts the mean response; (b) the residual flow transports noisy control cells toward the mean-subtracted response; (c) mean guidance combines the two at inference.

tion to model conditional population changes (Bunne et al., 2023; He et al., 2026; Yuan et al., 2026). CellFlow and scDFM further use conditional flow matching for population generation, with scDFM performing transport directly in gene expression space (Klein et al., 2025; Yu et al., 2026). These approaches broaden the scope of perturbation prediction from expected responses to population-level generation. Recent benchmarks also indicate that performance on mean response metrics remains challenging for complex deep models (Ahlmann-Eltze et al., 2025; Wei et al., 2026).

Flow Matching and Wasserstein Gradient Flows. Flow matching learns a time-dependent velocity field that transports samples between probability distributions along prescribed probability paths (Lipman et al., 2022). Conditional formulations extend this framework to diverse biological conditions and have recently been applied to single-cell population prediction (Klein et al., 2025; Yu et al., 2026). Wasserstein gradient flows provide a distribution-level formulation by describing probability evolution as the gradient flow of an energy functional under the Wasserstein metric. The JKO scheme discretizes this evolution into a sequence of proximal optimization problems, each involving an optimal transport problem (Jordan et al., 1998). Recent work introduces neural parameterizations to improve the computational tractability of these optimal transport based proximal updates (Choi et al., 2024). These frameworks capture complementary aspects of distribution transport: flow matching focuses on sample transport dynamics, whereas Wasserstein gradient flows characterize distribution evolution through successive optimal transport based optimization.

3 Method

Problem Setup. The fundamental task is to predict the transcriptomic distribution of a cell population in response to an unseen experimental intervention. Let $\mathcal{G} = \{g_1, \ldots, g_G\}$ denote the common set of G genes profiled across all cells. In this shared expression space, a vector $c_x = (c_x(g_1), \ldots, c_x(g_G)) \in \mathbb{R}^G$ describes the state of an unperturbed baseline cell, whereas a vector $x = (x(g_1), \ldots, x(g_G)) \in \mathbb{R}^G$ describes a cell measured under a specific perturbation condition p. This condition p defines the biological intervention, such as targeting one or two CRISPR genes or applying a drug at a defined dose. Training data comprise unperturbed and perturbed populations, $X_0 \sim \mathbb{P}_0$ and $X_p \sim \mathbb{P}_p$, for observed conditions $p \in \mathcal{P}_{\text{train}}$. Destructive single-cell profiling ren-

ders these observations strictly unpaired. The objective is to learn a conditional generative model $p_\theta(x \mid c_x, p)$ that predicts the target distribution $\mathbb{P}_{p^*}$ for unseen conditions $p^* \notin \mathcal{P}_{\text{train}}$.

Model Overview. Figure 1 provides an overview of scPertWGF, which decomposes perturbation responses into a population mean shift and residual distributional structure. Given control cells and a perturbation condition, the model first predicts the perturbation-induced mean shift with a closed-form ridge head (§ 3.1), and then models the residual response distribution using a conditional flow (§ 3.2). The flow is trained with flow matching for local transport supervision and a Wasserstein gradient flow objective for distribution-level alignment. During inference, the predicted mean is combined with the generated residuals and calibrated through mean guidance (§ 3.3) to obtain the final response population.

3.1 Mean prediction with a closed-form ridge estimator

For genetic perturbations, we adopt the bilinear ridge estimator of Ahlmann-Eltze et al. (2025) to predict the perturbation-induced shift in population mean, capturing coordinated transcriptional changes induced by the perturbed gene. The estimator is fitted in closed form independently of the conditional residual flow.

We apply PCA to the control mean and the response means from the training perturbations to obtain a rank-r principal component basis $U \in \mathbb{R}^{G\times r}$. For each gene g, let $u_g \in \mathbb{R}^r$ denote the transpose of the corresponding row of U. For S training single-gene perturbations, let $Y_c \in \mathbb{R}^{G\times S}$ contain their mean expression shifts after centering across perturbations, and let $B \in \mathbb{R}^{S\times r}$ contain the corresponding coordinates of the perturbed genes. Let $\bar{y}, \bar{c} \in \mathbb{R}^G$ denote the average mean shift and the control mean, respectively. The coefficient matrix $K \in \mathbb{R}^{r\times r}$ is estimated by ridge regression:

$$\hat{\delta}_p = UKu_g + \bar{y}, \qquad K = (U^\top U + \lambda_r I)^{-1} U^\top Y_c B (B^\top B + \lambda_r I)^{-1}, \tag{1}$$

where $\lambda_r > 0$ is the ridge penalty and $I = I_r$. The predicted population mean is $\hat{\mu}_p = \bar{c} + \hat{\delta}_p$.

The PCA basis summarizes coordinated transcriptional variation observed across training perturbations. Thus, although a gene may not have been used as a perturbation target during training, its expression profile is represented in the basis and can contribute to predicting the response to an unseen target gene.

For a double perturbation involving genes A and B, we assume additive mean responses:

$$\hat{\mu}_{A+B} = m_A + m_B - \bar{c}, \tag{2}$$

where m_A and m_B are the observed population means of the corresponding single perturbations in the training set. If a constituent single perturbation is unavailable, its mean is replaced by the average population mean over all training single perturbations. This additive estimate provides the mean reference for the combined response.

For a drug d administered at dose D, we predict the mean expression shift using ridge regression on the molecular fingerprint of d and a log-transformed dose feature. The regression model is fitted using training perturbation conditions only.

3.2 Residual flow for population shape modeling

The population mean captures the overall transcriptional response to a perturbation but does not characterize variation across individual cells. By decoupling the predicted mean shift, the resulting residuals isolate intrinsic cellular heterogeneity and any remaining mean-estimation error. scPertWGF captures this complex residual distribution via a conditional flow T, transporting a noise-smoothed control population toward the perturbed target state.

The transport endpoints are defined by shifting both the noisy control and observed response cells by the predicted mean: $z_0 = c_x + \sigma\Sigma\xi - \hat{\mu}_p$ and $z_1 = x - \hat{\mu}_p$. Here, $\xi \sim \mathcal{N}(0, I_G)$, Σ is a diagonal matrix of per-gene standard deviations from the control population, and σ scales the source noise, effectively relaxing the finite empirical control sample into a continuous source measure. The transport map T is governed by the neural ordinary differential equation (ODE) $\dot{z}_t = v_\theta(z_t, t, e_p)$, integrated from $t = 0$ to $t = 1$, where e_p denotes the perturbation condition embedding. An MLP

parameterizes the velocity field v_θ across all G genes jointly, conditioned on the current state, time, and e_p (implementation details in Appendix A.1).

The flow matching objective optimizes v_θ to regress the constant velocity $z_1 - z_0$ along the linear probability path $z_t = (1-t)z_0 + tz_1$, where $t \sim \mathcal{U}[0,1]$ and the endpoints are sampled independently (Lipman et al., 2022). While this objective ideally recovers the marginal velocity field and the target distribution, finite-sample regimes often lack tight distributional guarantees. To promote population-level alignment, the flow matching objective is augmented with a Wasserstein gradient flow (WGF) formulation.

Let ρ_{src} and ρ_p denote the marginal distributions of z_0 and z_1 for a fixed condition p. Denoting T_k as the transport map at step k and $\rho_k = (T_k)_{\#}\rho_{\text{src}}$ as its pushforward distribution, the ideal Jordan-Kinderlehrer-Otto (JKO) update with step size $h > 0$ is defined as:

$$\rho_{k+1} = \arg\min_{\rho} \ \text{KL}(\rho \,\|\, \rho_p) + \frac{1}{2h} W_2^2(\rho, \rho_k). \tag{3}$$

This represents an implicit Euler discretization of the gradient flow of the KL energy (Jordan et al., 1998). The KL divergence drives the generated distribution toward the target response ρ_p, while the 2-Wasserstein metric W_2^2 acts as a proximal penalty to restrict the displacement from the previous iterate ρ_k.

Building upon the semi-dual JKO framework (Choi et al., 2024), the training alternates between updating the velocity field and a neural potential within each phase, utilizing the current flow to warm-start subsequent phases. To construct a tractable and regularized approximation of Eq. (3), the final objective integrates a variational KL divergence, an inherited-coupling upper bound for the Wasserstein cost, a mean-anchor constraint, and the base transport supervision (Appendix A.2). Crucially, this flow operates directly within the raw gene expression space, stabilized by annealed instance noise to smooth both empirical measures during matching.

Training Objective. The velocity field v_θ and the neural potential v_ϕ are optimized in an alternating manner. Let $\hat{z} = T(z_0)$ denote a generated residual capturing perturbation-induced cellular heterogeneity, and $z_1 \sim \rho_p$ an experimentally measured target residual. To mitigate the intrinsic sparsity of single-cell transcriptomes, the empirical distributions are smoothed during phase k. Specifically, the potential evaluates noise-corrupted states $\tilde{\hat{z}} = \hat{z} + \sigma_k \Sigma \xi_g$ and $\tilde{z}_1 = z_1 + \sigma_k \Sigma \xi_r$, where $\xi_g, \xi_r \sim \mathcal{N}(0, I_G)$ are independent standard normal variables, and σ_k is an annealed instance-noise scale. Conditioned on the perturbation p, which is omitted from the notation v_ϕ for brevity, the optimization employs a gradient-penalty weight β and a mean-anchor weight λ_m to formulate the dual objectives:

$$\mathcal{L}_\phi = \mathbb{E}\big[v_\phi(\tilde{\hat{z}})\big] + \mathbb{E}\big[e^{-v_\phi(\tilde{z}_1)} - 1\big] + \beta\, \mathbb{E}\big\|\nabla_{\tilde{z}_1} v_\phi(\tilde{z}_1)\big\|^2, \tag{4}$$

$$\mathcal{L}_\theta = \mathcal{L}_{\text{TS}} + \mathcal{L}_{\text{WGF}} + \lambda_m\, \mathcal{L}_{\text{mean}}. \tag{5}$$

Here, $\mathcal{L}_{\text{TS}}$ enforces local transport supervision via flow matching regression on linear trajectories, serving as the sole velocity objective during the adversarial warm-up. To promote global distributional alignment, the WGF loss integrates a variational matching objective with a proximal regularization term governed by λ_p, approximating the implicit JKO update in Eq. (3):

$$\mathcal{L}_{\text{WGF}} = -\mathbb{E}[v_\phi(\tilde{\hat{z}})] + \frac{\lambda_p}{2hG}\, \mathbb{E}\|\hat{z} - \hat{z}_{\text{old}}\|^2. \tag{6}$$

The first term leverages the learned potential to align the generated manifold with the target response distribution. In the proximal term, $\hat{z}_{\text{old}} = T_{\text{old}}(z_0)$ represents the output from the frozen transport map of the preceding phase. This shared source establishes a coupling where the expected squared displacement per gene upper-bounds the normalized Wasserstein cost, constraining structural divergence between phases, as proven in Appendix A.2.

To stabilize the macro-transcriptomic response, the mean anchor $\mathcal{L}_{\text{mean}} = \frac{1}{G}\|\overline{\tilde{\hat{z}}} - \bar{z}_1\|^2$ penalizes deviations between the generated and target minibatch means, indicated by overlines. Notably, both the proximal constraint in $\mathcal{L}_{\text{WGF}}$ and the $\mathcal{L}_{\text{mean}}$ anchor are evaluated strictly on uncorrupted transcriptomic states prior to noise injection. The overall optimization procedure is stabilized via a bounded potential architecture, gradient clipping, and an adversarial warm-up, utilizing an exponential moving average of the velocity parameters for final inference. Comprehensive hyperparameter configurations are provided in Appendix A.1.

3.3 Mean-guided inference

For an unseen condition p^*, single-cell profiles are reconstructed by adding predicted means to generated residuals, $\hat{x}_i = \hat{z}_i + \hat{\mu}_{p^*}$. Because target residuals are defined relative to the head prediction, their mean reflects the offset between observed and predicted means. For N generated cells, $\overline{\hat{z}} = N^{-1} \sum_{i=1}^{N} \hat{z}_i$ denotes the empirical residual mean. The final mean $\overline{\hat{x}} = \hat{\mu}_{p^*} + \overline{\hat{z}}$ combines the head prediction with the residual mean estimated by the flow.

To calibrate the population mean while preserving its covariance structure, mean guidance introduces a scalar calibration factor α:

$$\hat{x}'_i = \hat{x}_i + \alpha\big(\hat{\mu}_{p^*} - \overline{\hat{x}}\big) \;=\; \hat{\mu}_{p^*} + \hat{z}_i - \alpha\,\overline{\hat{z}}, \qquad \alpha \in [0, 1]. \tag{7}$$

The calibrated mean is $(1-\alpha)\overline{\hat{x}} + \alpha\hat{\mu}_{p^*}$, combining the flow and head estimates. Because this calibration constitutes a global translational shift across the response manifold, it strictly preserves the covariance structure and all higher-order central moments that encode intrinsic cellular heterogeneity. Setting $\alpha = 1$ rigidly anchors the generated population center to the closed-form deterministic prediction, whereas $\alpha = 0$ retains the raw generative output. The guidance scalar is set per experimental screen, with selection criteria detailed in Appendix A.7.

4 Experiments

4.1 Experimental Setup

Detailed explanations of extended data preprocessing, formal metric definitions, baseline evaluation conventions, and network hyperparameters are provided in Appendices A.1 and A.3.

Datasets. We evaluate on five public single-cell perturbation screens, comprising four CRISPR and one chemical dataset from the scPerturBench (Wei et al., 2026) and scPerturb (Peidli et al., 2024) collections, utilizing the standard scPerturBench data splits. Norman (Norman et al., 2019) pairs gene activations with rich double perturbations, Adamson (Adamson et al., 2016) targets single genes, Replogle (Replogle et al., 2022) provides the genome-scale K562 essential-gene screen, and Schmidt (Schmidt et al., 2022) provides CRISPRa Perturb-seq data in primary human T cells, with single- and two-gene conditions in the benchmark. For the drug screen, we use the MCF7 subset of sci-Plex 3 (Srivatsan et al., 2020), hereafter SciPlex, with nine unseen drugs at four doses as test conditions. Test perturbations are held out entirely, so every screen measures generalization to unseen conditions; a held-out combination may contain genes seen individually during training. All expression profiles are log-normalized.

Metrics. The evaluation employs five established metrics to assess perturbation responses. Pearson Δ measures the cross-gene agreement between predicted and observed expression changes relative to control. Mean absolute error, denoted MAE, and mean squared error, MSE, quantify per-gene deviations in the predicted mean profile; these three metrics follow the cell-eval implementations (Adduri et al., 2025). The L_2 metric computes the Euclidean distance between predicted and observed mean profiles (Ahlmann-Eltze et al., 2025). The energy distance, referred to as E-dist, evaluates the overall discrepancy between predicted and observed cell populations (Wei et al., 2026).

Baselines. All evaluated baselines are retrained on identical data splits and processed through a unified evaluation pipeline. The heuristic model baseControl duplicates the unperturbed control profile, whereas trainMean predicts the average of the training responses. GenePert applies regularized regression over perturbation-gene embeddings, and baseMLP operates as a multilayer perceptron trained independently for each screen. For genetic screens, the deep baselines comprise GEARS (Roohani et al., 2024), scGPT (Cui et al., 2024), CellFlow (Klein et al., 2025), STATE (Adduri et al., 2025), and scDFM (Yu et al., 2026). For the chemical screen, chemCPA is additionally implemented as a from-scratch reproduction.

Implementation details. scPertWGF is trained in PyTorch on a single NVIDIA A100 GPU. The optimization proceeds over five alternating phases, each comprising $4{,}000$ iterations with a batch

size of 256. To ensure balanced representation, each training step hierarchically samples a perturbation condition followed by its associated single-cell profiles. Final inference exclusively employs the exponential moving average (EMA) of the velocity network parameters to promote generative stability.

4.2 Results on the genetic screens

As detailed in Table 1, scPertWGF consistently improves prediction of the average transcriptional response to unseen genetic perturbations, achieving the lowest L2 and highest Pearson Δ on all four screens. The strongest L2 baseline differs across screens: GenePert leads on Norman and trainMean on Adamson, whereas scGPT leads on Replogle and Schmidt. Relative to these references, the L2 reduction ranges from 2.7% to 9.5%, showing that the gains extend across both simple mean predictors and deep models. Together, lower L2 and higher Pearson Δ indicate smaller errors in the average response and better recovery of the pattern of expression changes.

Table 1: Quantitative evaluation on genetic screens under the official splits. Best and second-best values are **bolded** and underlined, respectively.

	Norman					Adamson					Replogle					Schmidt				
	Pearson Δ	MSE	MAE	L2	E-dist	Pearson Δ	MSE	MAE	L2	E-dist	Pearson Δ	MSE	MAE	L2	E-dist	Pearson Δ	MSE	MAE	L2	E-dist
baseControl	-0.018	0.0065	0.032	5.27	1.40	0.009	0.0041	0.027	4.06	0.79	-0.002	0.0055	0.050	5.33	1.11	-0.017	0.0149	0.065	8.12	3.22
baseMLP	0.383	0.0250	0.058	9.94	4.87	0.149	0.0258	0.065	9.67	4.26	0.187	0.0056	0.052	5.53	1.12	0.241	0.0199	0.079	9.50	3.82
trainMean	0.650	0.0036	0.026	3.97	0.94	0.519	0.0029	0.023	3.47	0.58	0.174	0.0049	0.048	4.73	**0.96**	0.228	0.0127	0.063	7.53	2.99
GenePert	0.620	0.0034	0.026	3.92	0.91	0.468	0.0031	0.024	3.59	0.62	0.184	0.0056	0.053	5.53	1.12	0.206	0.0127	0.063	7.58	3.00
GEARS	0.528	0.0055	0.041	4.95	1.32	0.320	0.0035	0.025	3.73	0.69	0.171	0.0050	0.049	5.15	1.05	0.154	0.0169	0.077	8.82	3.45
scGPT	0.583	0.0039	0.026	4.16	0.99	0.538	0.0029	0.024	3.52	0.62	0.305	0.0040	0.043	4.52	1.06	0.334	0.0074	**0.046**	5.82	3.07
CellFlow	0.296	0.0050	0.030	4.85	1.68	0.307	0.0035	0.025	3.88	1.09	0.031	0.0054	0.048	5.31	1.84	0.075	0.0152	0.062	8.58	4.15
STATE	0.379	0.0054	0.030	4.84	1.17	0.053	0.0087	0.051	6.48	1.81	0.125	0.0054	0.051	5.34	1.13	0.096	0.0148	0.067	8.19	3.21
scDFM	0.466	0.0051	0.035	4.93	1.79	0.253	0.0040	0.033	4.17	0.80	0.184	0.0058	0.053	5.30	1.69	0.180	0.0122	0.064	7.53	2.98
scPertWGF	**0.664**	**0.0031**	**0.024**	**3.64**	**0.72**	**0.581**	**0.0024**	**0.020**	**3.14**	**0.48**	**0.318**	**0.0037**	**0.040**	**4.31**	1.73	**0.373**	**0.0069**	**0.046**	**5.66**	**2.24**

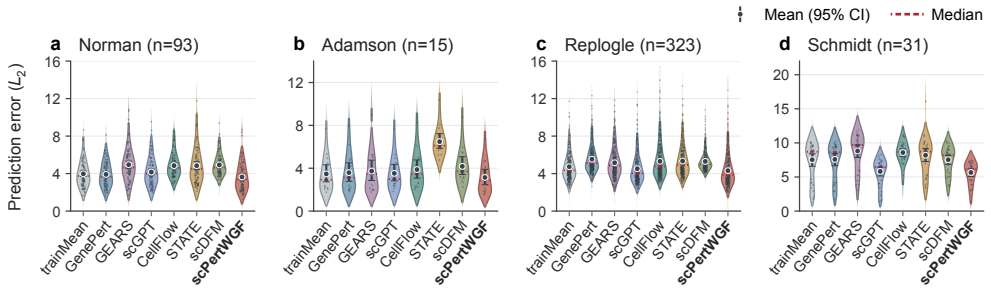

Figure 2: Distribution of L_2 prediction errors across official test splits. Dots plot errors per perturbation, where n is the condition count. Violin plots approximate error densities; dark red dashed lines and circles indicate medians and means, respectively. Vertical bars represent 95% bootstrap confidence intervals derived from 10,000 resamples. Y-axis scales are independent across panels.

The gains in average L2 are accompanied by lower median errors on all four screens (Figure 2). This agreement indicates that the advantage is also reflected in the typical prediction error across unseen conditions.

Accurate mean expression does not by itself establish agreement between response populations. scPertWGF achieves the lowest E-dist on Norman, Adamson, and Schmidt; on Norman, E-dist falls from 0.91 for GenePert to 0.72, a 20.9% improvement over the strongest baseline. For CEBPA+CEBPE, trainMean and GenePert concentrate predictions in narrow regions, while scPertWGF generates a broader spread across the observed embedding (Figure 3). Replogle illustrates the limits of mean accuracy: improved mean predictions do not yield the lowest E-dist, with a detailed analysis in Appendix A.6.

4.3 Results on the drug screen

The proposed formulation seamlessly extends to unseen chemical interventions excluded during training. As detailed in Table 2 for the SciPlex dataset, scPertWGF decreases the L_2 prediction error to 6.17 compared to 7.20 for the chemCPA baseline, yielding a 14.3% relative re-

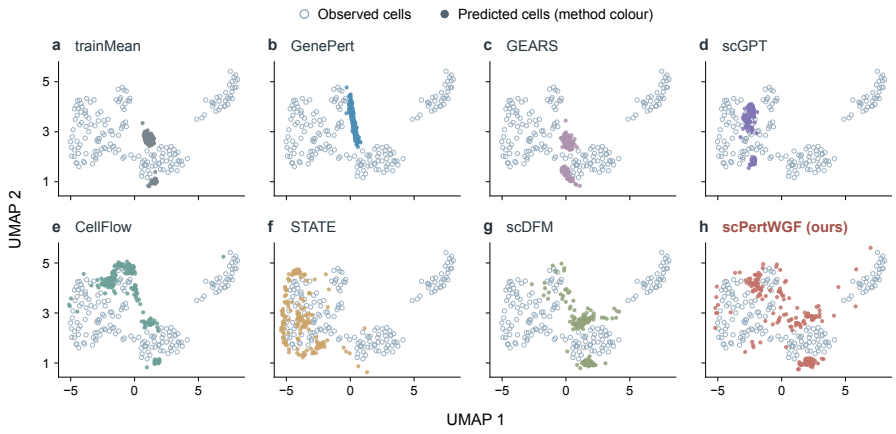

Figure 3: UMAP embeddings for the unseen CEBPA+CEBPE perturbation in the Norman dataset. Each panel overlays 180 generated cells depicted as colored markers alongside 180 observed cells shown as blue-grey open circles. Following the preprocessing protocol in Appendix A.3, predicted profiles are projected onto a UMAP manifold fitted exclusively to observed data.

duction. Concurrently, the Pearson Δ correlation reaches 0.467, substantially outperforming the 0.209 achieved by the trainMean heuristic. This dual improvement confirms that highly accurate average response estimations align directly with faithful expression pattern recovery. Furthermore, achieving the lowest energy distance indicates that these predictive advantages translate into precise distributional alignment, thereby validating the strong generalization capacity of the framework across diverse genetic and chemical perturbations.

Table 2: Results on SciPlex (MCF7) for 36 test conditions. Best in **bold**; second best underlined.

	Pearson $\Delta \uparrow$	MSE $\downarrow$	MAE $\downarrow$	L2 $\downarrow$	E-dist $\downarrow$
baseControl	-0.003	0.0146	0.031	7.77	3.42
baseMLP	0.186	0.0184	0.048	9.24	4.04
trainMean	0.209	0.0138	0.032	7.46	3.24
CellFlow	0.032	0.0158	0.033	8.76	11.67
chemCPA	0.197	0.0122	0.030	7.20	4.64
scPertWGF	**0.467**	**0.0088**	**0.025**	**6.17**	**2.44**

4.4 Ablation Study

The ablation studies examine the effects of the proposed decomposition on mean prediction and population-level agreement, as shown in Figure 4. Variant definitions are provided in Appendix A.4.

Mean Prediction and Calibration. Residual generation alone does not recover the most accurate population mean. Removing inference-time guidance lowers Pearson Δ from 0.664 to 0.600 and increases L2 from 3.64 to 3.94, showing that calibration corrects errors remaining in the generated average response. The variant trained without the mean head and evaluated without guidance performs notably worse across both metrics. The specific comparison controlling for this guidance is discussed separately in Appendix A.4. Furthermore, the mean anchor operates during the training phase and yields a highly selective effect. Without this anchor, the Pearson Δ falls to 0.644 while the L_2 error remains approximately 3.64. Its benefit is therefore more pronounced in the pattern of expression changes rather than in the magnitude of the mean profile error, complementing the final calibration step.

Population-level Training. The WGF objective contributes a benefit that mean-profile accuracy alone would largely miss. Removing this objective increases the energy distance from 0.722 to 0.881, representing a 22.0% deterioration, while the L_2 error remains nearly unchanged at 3.63 compared to 3.64 for the full model. This represents the largest E-dist increase among all tested component removals, supporting the role of population-level training in matching the response distribution. Removing transport supervision produces a comparatively smaller deterioration in E-dist, confirming its complementary role in learning the underlying transport velocity field. The training phase analy-

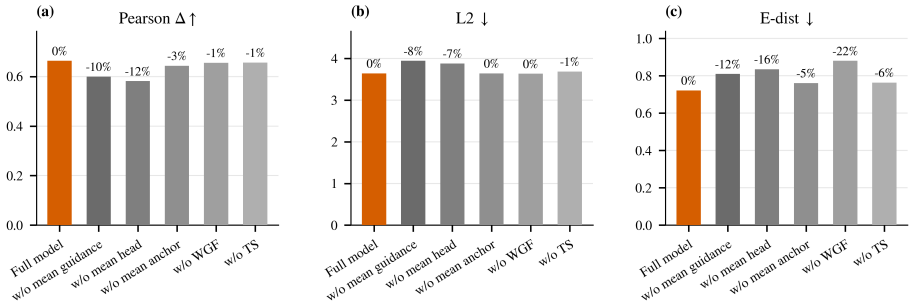

Figure 4: Component ablations on Norman. Labels show percentage changes from the full model, with negative values indicating worse performance.

sis and additional controls built around the identical mean head are reported extensively in Appendices A.4 and A.5.

4.5 Parameter sensitivity

The best balance between the head and flow means depends on the response population. With training fixed, Norman benefits from retaining both estimates. Specifically, the E-dist improves to 0.72 at $\alpha = 0.5$ compared to 0.80 when relying exclusively on the analytic mean, whereas the Pearson Δ correlation peaks at $\alpha = 0.7$, as reported in Table 3. By contrast, Schmidt and SciPlex favor full calibration to the head mean on both metrics. This variation explains why a single guidance strength is not optimal across all screens and criteria. Guidance applies the same translation to every generated cell, so these changes occur without altering the centered population structure. It can correct the location of the generated population but cannot replace learning its variation through the residual flow. The validation protocol used to set guidance and the subsequent evaluation across data splits are described in Appendices A.7 and A.8.

Table 3: Sensitivity to the guidance scalar α across five screens under the primary training settings.

	Norman		Adamson		Replogle		Schmidt		SciPlex	
α	Pear. Δ	Edist	Pear. Δ	Edist	Pear. Δ	Edist	Pear. Δ	Edist	Pear. Δ	Edist
0.0	0.600	0.81	0.462	0.60	0.208	1.84	0.216	2.97	0.306	2.97
0.3	0.649	0.74	0.528	0.54	0.256	1.77	0.265	2.56	0.389	2.72
0.5	0.664	**0.72**	0.561	0.51	0.291	1.74	0.306	2.38	0.429	2.60
0.7	**0.666**	0.73	**0.581**	0.48	0.318	**1.73**	0.349	2.27	0.455	2.51
1.0	0.656	0.80	0.577	**0.46**	**0.320**	1.74	**0.373**	**2.24**	**0.467**	**2.44**

5 Conclusion

We presented scPertWGF, which combines a separately fitted closed-form mean estimator with a conditional residual flow for single-cell perturbation prediction. Flow matching and a Wasserstein-inspired population objective train the flow directly in gene space, and scalar mean guidance calibrates the generated population at inference. Experiments on genetic and chemical perturbations demonstrate improved prediction of unseen responses, supporting the value of combining analytic mean estimation with residual population modeling.

The current evaluation focuses on unseen perturbations within individual screens, leaving generalization across cellular contexts untested. Future work can extend this decomposition across cellular contexts and other conditional generation tasks.

AI use statement

In this paper, we used generative AI tools as general-purpose assist tools for text refinement and language polishing. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

Ethics statement

This study uses only publicly available single-cell perturbation datasets, in accordance with their respective data-use agreements. No personally identifiable or patient-specific information is involved. The method is intended for in silico modeling and hypothesis generation in biological research. Predictions are not intended for direct clinical use, and we caution against interpreting them without experimental validation.

Reproducibility statement

All experiments use the public screens with their official train and test splits (seed 1), described with the preprocessing in Appendix A.1. Core hyperparameters of the head and the flow are listed there, and Algorithm 1 summarizes the training procedure, including the alternation, the warm-up, and the snapshot updates. All methods, ours and baselines alike, are evaluated through the single pipeline of Appendix A.3, whose metric definitions are given with their formulas. Robustness across three official data splits is reported in Appendix A.8. Source code, configuration files, and the result files in the official format will be released upon publication; the released pipeline covers training, generation, and the inference-time mean-guidance step.

References

Britt Adamson, Thomas M Norman, Marco Jost, Min Y Cho, James K Nuñez, Yuwen Chen, Jacqueline E Villalta, Luke A Gilbert, Max A Horlbeck, Marco Y Hein, et al. A multiplexed single-cell crispr screening platform enables systematic dissection of the unfolded protein response. *Cell*, 167(7):1867–1882, 2016.

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

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

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

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.

Jaemoo Choi, Jaewoong Choi, and Myungjoo Kang. Scalable wasserstein gradient flow for generative modeling through unbalanced optimal transport. *arXiv preprint arXiv:2402.05443*, 2024.

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

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.

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

Richard Jordan, David Kinderlehrer, and Felix Otto. The variational formulation of the fokker–planck equation. *SIAM journal on mathematical analysis*, 29(1):1–17, 1998.

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

Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*, 2022.

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

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

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.

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

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.

Ralf Schmidt, Zachary Steinhart, Madeline Layeghi, Jacob W Freimer, Raymund Bueno, Vinh Q Nguyen, Franziska Blaeschke, Chun Jimmie Ye, and Alexander Marson. Crispr activation and interference screens decode stimulation responses in primary human t cells. *Science*, 375(6580): eabj4008, 2022.

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

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

Zhiting Wei, Yiheng Wang, Yicheng Gao, Shuguang Wang, Ping Li, Duanmiao Si, Yuli Gao, Siqi Wu, Danlu Li, Kejing Dong, et al. Benchmarking algorithms for generalizable single-cell perturbation response prediction. *Nature Methods*, 23(2):451–464, 2026.

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

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.

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

A Appendix

Contents

A.1 Implementation details 12
A.2 The adversarial component as a variational bound 13
A.3 Evaluation protocol and metrics 14
A.4 Additional ablation analyses 16
A.5 Population controls using the same mean head 17
A.6 Population-objective analysis on Replogle 17
A.7 Guidance selection 18
A.8 Robustness across data splits 18

A.1 Implementation details

Common setup.

- **Data.** We use the preprocessed genetic datasets released by scPerturBench and retain their supplied gene sets: $G = 5{,}025$ (Norman), $5{,}043$ (Adamson), $5{,}822$ (Replogle), and $5{,}005$ (Schmidt), under the official train and test splits (seed 1). Genetic conditions use scGPT gene embeddings (512-dim); the chemical mean head uses an RDKit Morgan fingerprint (2048 bits) concatenated with $\frac{1}{4}\log_{10} D$.
- **Head.** Response basis rank $r = 10$, ridge penalty $\lambda_r = 0.1$; chemical ridge penalty 1.0. Double-perturbation means add the observed means of the constituent training singles and subtract the control mean; each unavailable constituent is replaced by the average population mean over all training singles.
- **Source and conditioning.** Source noise $\sigma = 0.1$ scaled by per-gene control standard deviations; perturbation embedding $e_p \in \mathbb{R}^{256}$; input concatenation and dose encoding are specified below.
- **Smoothed JKO.** $K_{\text{phase}} = 5$ phases $\times$ $4{,}000$ iterations, step size $h = 0.1$; mean-anchor weight $\lambda_m = 100$; proximal weight $\lambda_p = 1$; R1 weight $\beta = 1$; potential clamped to $[-6, 6]$ so the exponential moment of the variational bound stays finite; instance-noise scale $\sigma_k = 0.05/(k+1)$ for zero-indexed phases $k = 0, \ldots, K_{\text{phase}} - 1$, scaled by the per-gene control standard deviations.
- **Optimization.** Adam with betas $(0.5, 0.9)$, learning rates 2×10^{-4} (velocity) and 1×10^{-4} (potential), batch size 256, gradient clipping at norm 1.0, adversarial warm-up of $1{,}000$ iterations, EMA decay 0.999 for inference.
- **Backbone and integration.** The supplied configuration has four hidden SiLU layers and uses 10 explicit Euler steps for training endpoints and inference; see below for details.

Per-screen settings. Only the hidden width and the guidance scalar vary by screen. Width follows the data scale, 256 on the compact screens and 2048 on the genome-scale and drug screens. Guidance uses $\alpha = 0.5$ on Norman, 0.7 on Adamson and Replogle, and 1.0 on Schmidt and SciPlex. The guidance-selection protocol is described in §A.7. All other settings follow the common setup.

Network architecture and tensor shapes. Let n_b be the cell batch size and G the number of genes, with a fixed gene ordering within each screen. The velocity network receives $z_t \in \mathbb{R}^{n_b \times G}$ and produces a velocity of the same shape. A frozen gene embedding table $E_g \in \mathbb{R}^{G \times 512}$ is projected by a learned affine map to $q = E_g w_g + b_g \mathbf{1} \in \mathbb{R}^G$ and broadcast over the batch. The condition encoder produces $e_p \in \mathbb{R}^{n_b \times 256}$, and the sinusoidal time encoding has shape $n_b \times 64$. The concatenated input is therefore

$$H_0 = [z_t,\ \mathbf{1}_{n_b} q^\top,\ e_p,\ \tau(t)] \in \mathbb{R}^{n_b \times (2G+320)}. \tag{8}$$

The MLP processes each complete cell vector jointly, mixing gene coordinates through dense weights; it is not a coordinate-wise network shared across genes. In the supplied configuration, the layer count names hidden-to-hidden layers after the first input-to-hidden layer, so both velocity and potential trunks have four hidden layers of width H, each followed by SiLU, and a final linear output layer. The potential uses a separate condition encoder, takes a residual-state batch concatenated with e_p, of shape $n_b \times (G + 256)$, and outputs one scalar per cell, clamped to $[-6, 6]$. It receives neither q nor time.

For genetic perturbations, the raw condition features have dimension 512; double perturbations sum the constituent embeddings before projection. The default projection is affine to dimension 256. For chemical conditions, a 2048-bit Morgan fingerprint is projected to 256 dimensions, and a separate $1 \to 64 \to 256$ MLP with an intermediate SiLU maps $\log_{10} D$ to an additive dose embedding. This differs from the scaled log-dose concatenated in the mean head. An optional variant replaces the affine condition projection by an input-to-512-to-256 MLP with one SiLU; it is disabled by default. The gene table E_g is set to zero on chemical screens, while the affine gene projection retains its bias.

Role of the fixed gene channel. For fixed parameters and a fixed screen, q is constant across cells, conditions, and integration times. For one cell, writing its inputs as column vectors, the first-layer preactivation is

$$\begin{aligned} a &= W_z z_t + W_q q + W_p e_p + W_t \tau(t) + b \\ &= W_z z_t + W_p e_p + W_t \tau(t) + b', \qquad b' = b + W_q q. \end{aligned} \tag{9}$$

Thus, the fixed channel provides an alternative first-layer bias parameterization. Gene coordinates are distinguished by dense input and output weights, while the perturbation embedding varies with the condition.

This functional equivalence allows different optimization trajectories under the two parameterizations.

ODE integration and differentiation. The supplied configuration uses fixed-step explicit Euler on $[0, 1]$ with $N_{\text{ODE}} = 10$ steps during both training and inference:

$$\begin{aligned} z^{(n+1)} &= z^{(n)} + \Delta t \, v_\theta(z^{(n)}, n\Delta t, e_p), \\ \Delta t &= 1/N_{\text{ODE}} = 0.1, \qquad n = 0, \dots, N_{\text{ODE}} - 1. \end{aligned} \tag{10}$$

The endpoint is $T(z_0) = z^{(N_{\text{ODE}})}$. This ODE step size is distinct from the JKO proximal parameter h, even though both are set to 0.1. For velocity updates, endpoint losses differentiate directly through all ten unrolled Euler steps using reverse-mode automatic differentiation; no adjoint solver, adaptive stepping, or trajectory checkpointing is used. When updating the potential, the generated endpoint is computed without tracking gradients through the velocity network. The frozen previous-phase endpoint in the proximal component of $\mathcal{L}_{\text{WGF}}$ is also computed without gradients, using the same source samples. Flow matching itself requires no ODE rollout: it evaluates the velocity once per sampled interpolated state at $t \sim \mathcal{U}[0, 1]$. After the warm-up, the current-flow endpoint is shared by the mean loss and both components of $\mathcal{L}_{\text{WGF}}$ that are enabled. Inference uses the EMA velocity parameters and the same ten-step solver without gradients.

A.2 The adversarial component as a variational bound

For any potential v and distributions P and Q with density ratio $r = \mathrm{d}P/\mathrm{d}Q$, the pointwise inequality $r\,v + e^{-v} \geq r - r \log r$ holds because the exponential function is convex, with equality at $e^{-v^*} = r$. Taking expectations under Q yields an upper bound on the negative KL divergence,

$$\mathbb{E}_P[v] + \mathbb{E}_Q\left[e^{-v} - 1\right] \;\geq\; -\mathrm{KL}(P \,\|\, Q), \tag{11}$$

or, after negation, a lower bound on KL. Equality is attained by the unrestricted potential $v^* = \log(\mathrm{d}Q/\mathrm{d}P)$ when this potential is admissible and the expectations are well defined. For fixed condition p and phase k, Eq. (4) uses this expression with P the law of $\tilde{\tilde{z}}$ and Q the law of $\tilde{z}_1$: the generated and target residual distributions after instance noise, respectively. Its gradient penalty is evaluated at $\tilde{z}_1$. The supplied implementation omits the additive constant -1, which does not affect gradients. The implemented objective uses a bounded neural potential and gradient regularization to optimize a variational lower bound on KL. With the potential held fixed, the generator minimizes $-\mathbb{E}[v_\phi(\tilde{\tilde{z}})]$, the term of the negated bound that depends on its samples. This variational construction motivates the adversarial component of $\mathcal{L}_{\text{WGF}}$ and combines with the transport-cost approximation below to form a JKO-inspired update (Choi et al., 2024).

The proximal component uses a coupling inherited from the transport maps. Two consecutive iterates are pushforwards of the same source, $\rho = T_{\#}\rho_{\text{src}}$ and $\rho_k = (T_{\text{old}})_{\#}\rho_{\text{src}}$, so the inherited coupling $(T(z_0),\, T_{\text{old}}(z_0))$ is an admissible coupling and

$$W_2^2(\rho, \rho_k) \;\leq\; \mathbb{E}\big\|T(z_0) - T_{\text{old}}(z_0)\big\|^2. \tag{12}$$

Algorithm 1 Training one screen as a smoothed JKO scheme

Input: controls X_0, responses $\{X_p\}$, head predictions $\hat{\mu}_p$, embeddings (e_p, e_g)
shift responses into deviation space $z_1 = x - \hat{\mu}_p$; initialize velocity field v_θ, potential v_ϕ, and an EMA copy of v_θ
for phase $k = 0, \ldots, K_{\text{phase}} - 1$ **do**
    snapshot $T_{\text{old}} \leftarrow T$; set instance-noise scale $\sigma_k \leftarrow 0.05/(k+1)$
    for $4{,}000$ iterations **do**
        sample a condition p, then independently sample $c_x \sim X_0$, $x \sim X_p$, and $\xi \sim \mathcal{N}(0, I_G)$; set $z_0 = c_x + \sigma\Sigma\xi - \hat{\mu}_p$ and $z_1 = x - \hat{\mu}_p$
        potential step. With T frozen, form the endpoint $\hat{z} = T(z_0)$ without gradients; form $\tilde{\hat{z}}$ and $\tilde{z}_1$ with independent noise $\sigma_k\Sigma\xi_g$ and $\sigma_k\Sigma\xi_r$; descend $\mathcal{L}_\phi$ in Eq. (4)
        velocity step. Sample $t \sim \mathcal{U}[0,1]$; form $z_t = (1-t)z_0 + tz_1$ and initialize $\ell = \mathcal{L}_{\text{TS}}$
        if at least $1{,}000$ global iterations have elapsed **then**
            recompute $\hat{z} = T(z_0)$ with gradients and obtain $\hat{z}_{\text{old}} = T_{\text{old}}(z_0)$ without gradients
            add $\mathcal{L}_{\text{WGF}} + \lambda_m \mathcal{L}_{\text{mean}}$ to ℓ, using fresh instance noise only in the adversarial component of $\mathcal{L}_{\text{WGF}}$
        end if
        descend ℓ with respect to θ; update its EMA copy
    end for
end for
Output: the EMA velocity field used at inference

Let $c(x, y) = \frac{1}{G}\|x - y\|^2$ denote the squared displacement per gene. The proximal component in Eq. (6) can then be written as $\lambda_p \, \mathbb{E}[c(\hat{z}, \hat{z}_{\text{old}})]/(2h)$, with the expected cost providing an upper bound on W_2^2/G. Together with the step size h, this normalization rescales the proximal weight while keeping consecutive iterates close as Eq. (3) prescribes. With the annealed instance noise, the finite potential class, and the transport-supervision and mean losses, the scheme is a regularized S-JKO-inspired approximation of the JKO step.

The smoothing is the annealed instance noise. At zero-indexed phase $k = 0, \ldots, K_{\text{phase}} - 1$, independent noise with covariance $\sigma_k^2 \Sigma\Sigma^\top$ is added to generated and target residuals before potential evaluation, with $\sigma_k = 0.05/(k+1)$. Here Σ is the diagonal matrix of control-gene standard deviations used for source noise. This gives progressively less smoothed distribution-matching objectives across phases. The mean loss and the proximal component use residuals before this additional noise. Together with the potential clamp, the gradient penalty, and the warm start, this is the stabilization recipe behind the smoothed scheme of §3.2.

A.3 Evaluation protocol and metrics

Protocol. Every method is evaluated through a single pipeline. Each method writes one result file that holds three categories of cells, control, stimulated, and imputed, and all metrics are computed from these files. For each held-out condition the generated population matches the measured population size. A control-scaled jitter of 10^{-3} guards zero-variance degeneracies in the differential-expression pipeline. The main-text tables, figures, and ablations use the official train and test splits with seed 1, with 93 held-out conditions on Norman, 15 on Adamson, 323 on Replogle, 31 on Schmidt, and 36 on SciPlex. Table 7 reports results across official split seeds 1–3.

Per-method evaluation conventions. The following conventions specify how model outputs enter evaluation.

- **scGPT, CellFlow (genetic), chemCPA.** Their outputs are near point masses, with an internal dispersion of 0.03–0.04 (scGPT), 0.17–0.25 (CellFlow), and 0.06–0.09 (chemCPA) of the real populations. Point metrics come from the raw outputs, distributional metrics from a *Sampled* variant that wraps the per-condition mean in control-scaled Gaussian noise.
- **CellFlow (drug).** Predicts through a 50-dimensional PCA of the training expression, and for unseen molecules the linear decode leaves the valid range. We clamp coordinates to the per-component training range before inversion; even clamped, E-dist stays above 11.

- **scDFM.** Generates on a designed width of 1,000 genes within a 2,000-gene space. Its raw decode overshoots the response amplitude by a median 1.3–1.8$\times$ across screens while leaving per-condition Pearson delta unchanged, so its rows rescale each deviation to the median training-response norm; raw numbers are retained in the released tables.
- **CPA.** Excluded. Its released routine passes arguments the pinned package version does not accept, and our reproduction maps every unseen perturbation to the same average response (pairwise mean separation 0.4–1.1 against a real 6.3, unchanged after 160 epochs) and near-zero profiles for unseen drugs.

Visualization protocol. Figure 3 uses the common genes across result files. For display, each predicted population is scaled to match the observed population's median per-cell expression sum. Observed and predicted cells are pooled for gene-wise standardization, clipping at ± 10, and 30-component PCA. UMAP is then fitted on the observed cells only (15 neighbors, minimum distance 0.2, random seed 0), and predictions are transformed with the fitted mapper. These display transformations are separate from the expression-space metric calculations below.

Metric definitions. The five reported metrics are computed through a common evaluation pipeline. Pearson delta, MAE, and MSE follow the cell-eval implementations (Adduri et al., 2025); L2 follows Ahlmann-Eltze et al. (2025); and E-dist follows the population-level energy-distance metric used in scPerturBench (Wei et al., 2026). Let P be the held-out conditions, with N_p predictions and M_p measured cells for condition p. Write the final evaluated predictions as $x_{p,i}^{\text{pred}}$, including mean guidance for scPertWGF, and the measured responses as $x_{p,j}^{\text{obs}}$. Their empirical means are $\bar{x}_p^{\text{pred}} = N_p^{-1} \sum_{i=1}^{N_p} x_{p,i}^{\text{pred}}$ and $\bar{x}_p^{\text{obs}} = M_p^{-1} \sum_{j=1}^{M_p} x_{p,j}^{\text{obs}}$. The final prediction mean need not equal the head reference $\hat{\mu}_p$. Define expression changes around the control mean as $\hat{d}_p = \bar{x}_p^{\text{pred}} - \bar{c}$ and $d_p = \bar{x}_p^{\text{obs}} - \bar{c}$.

Pearson correlation of expression changes (Pearson delta). We compute the correlation across genes between predicted and measured expression changes for each held-out condition and average over conditions:

$$\text{Pearson } \Delta = \frac{1}{|P|} \sum_{p \in P} \text{corr}\big(\hat{d}_p,\ d_p\big), \tag{13}$$

evaluating agreement in the pattern of perturbation-induced expression changes across genes.

MSE and MAE (mean-profile error). The per-gene squared and absolute errors of the predicted mean profile,

$$\text{MSE} = \frac{1}{|P|} \sum_{p \in P} \frac{1}{G} \big\| \bar{x}_p^{\text{pred}} - \bar{x}_p^{\text{obs}} \big\|^2, \tag{14}$$

$$\text{MAE} = \frac{1}{|P|} \sum_{p \in P} \frac{1}{G} \big\| \bar{x}_p^{\text{pred}} - \bar{x}_p^{\text{obs}} \big\|_1. \tag{15}$$

L2 (mean-level perturbation distance). The average Euclidean distance between the predicted and measured mean profiles,

$$\text{L2} = \frac{1}{|P|} \sum_{p \in P} \big\| \bar{x}_p^{\text{pred}} - \bar{x}_p^{\text{obs}} \big\|_2, \tag{16}$$

evaluating the magnitude of the mean-profile prediction error in gene expression space.

E-dist (population energy distance). We follow the energy-distance evaluation used in scPerturBench (Wei et al., 2026), whose scripts select genes from a condition-specific DEG ranking and compute E-dist through pertpy. Our implementation uses the same Euclidean pairwise-distance estimator on all genes retained after preprocessing, without condition-specific DEG subsetting. Let $x_1^{\text{pred}}, x_2^{\text{pred}}$ be independent draws from the final predicted population and $x_1^{\text{obs}}, x_2^{\text{obs}}$ from the measured population. Using up to 2,000 cells from each population under a fixed seed,

$$\text{E-dist} = 2\,\mathbb{E}\|x_1^{\text{pred}} - x_1^{\text{obs}}\| - \mathbb{E}\|x_1^{\text{pred}} - x_2^{\text{pred}}\| - \mathbb{E}\|x_1^{\text{obs}} - x_2^{\text{obs}}\|, \tag{17}$$

At the population level, this quantity is zero only when the two distributions coincide (for finite first moments). We estimate it using empirical pairwise distances, including self-pairs, and average condition-level estimates across held-out conditions.

A.4 Additional ablation analyses

Component definitions. The variants in Figure 4 use $\alpha = 0.5$, except w/o mean guidance ($\alpha = 0$) and w/o mean head. The former retains the trained residual flow and omits inference-time calibration; the latter trains the flow end-to-end in gene space, without the head or guidance. The w/o WGF variant retains the head and flow while removing both the adversarial and proximal components of the WGF objective. The w/o TS variant removes the transport-supervision (flow matching) signal, so after the adversarial warm-up the velocity is trained solely by the endpoint losses. Percentage labels report relative changes from the full model, with signs reversed for L2 and E-dist so that negative values indicate worse performance.

Residual parameterization under matched guidance. To distinguish the effect of residual training from inference-time mean calibration, Table 4 compares residual and end-to-end flows with the same architecture, width, optimizer settings, training budget, and post-hoc guidance sweep. At $\alpha = 1$, guidance fixes both generated populations to the same head mean, making their E-dist difference attributable to the centered population distributions. Residual training reduces E-dist at every tested α, including this matched-mean setting (Figure 5a). The mean paired E-dist improvement at $\alpha = 1$ is 0.060 (95% CI $[0.036, 0.083]$), with a median of 0.033. Residual training lowers E-dist on 77 of 93 conditions (Figure 5b).

At the reported Norman operating point, $\alpha = 0.5$, the paired E-dist improvement is 0.041 (95% CI $[0.017, 0.064]$; 66 of 93 conditions), whereas the Pearson delta difference is 0.003 ($[-0.004, 0.009]$). The comparison therefore supports a distributional benefit from residual training without establishing an improvement in Pearson delta under matched guidance.

	End-to-end flow					Residual flow				
α	Pearson $\Delta\uparrow$	MSE$\downarrow$	MAE$\downarrow$	L2$\downarrow$	E-dist$\downarrow$	Pearson $\Delta\uparrow$	MSE$\downarrow$	MAE$\downarrow$	L2$\downarrow$	E-dist$\downarrow$
0.0	0.582	0.0034	0.0261	3.88	0.835	0.600	0.0036	0.0271	3.94	0.810
0.3	0.644	0.0030	0.0239	3.64	0.771	0.649	0.0032	0.0246	3.69	0.737
0.5	0.661	**0.0030**	**0.0232**	**3.60**	**0.763**	0.664	**0.0031**	0.0237	**3.64**	**0.722**
0.7	**0.664**	0.0031	0.0233	3.66	0.782	**0.666**	0.0031	**0.0236**	3.68	0.734
1.0	0.656	0.0035	0.0249	3.92	0.861	0.656	0.0035	0.0249	3.92	0.802

Table 4: **Decomposition under matched guidance** (Norman, seed 1). The same post-hoc guidance sweep is applied to an end-to-end trained flow and to the residual model (same backbone, width, and budget); best per column in bold. Point metrics nearly coincide and are identical at $\alpha = 1$; the residual parameterization improves E-dist at every α.

Mean anchoring across conditions. Removing the mean anchor from the Norman model reduces Pearson delta by 0.020 and worsens it for 70 of 93 conditions. The aggregate change is therefore accompanied by deterioration across most held-out conditions, supporting the role of the training-time anchor in controlling the generated population mean even when inference-time guidance is retained.

Number of training phases. With the total number of iterations held fixed, E-dist decreases from 0.77 with $K = 1$ to 0.75 with $K = 3$ and 0.72 with $K = 5$ on Norman. This comparison supports the phased training scheme at a fixed optimization budget. Changing the number of phases updates both the proximal reference and the instance-noise schedule, so the comparison does not isolate their individual contributions.



Figure 5: **Residual modeling under matched guidance on Norman.** (a) Mean E-dist across guidance strengths for the two flow parameterizations. (b) Paired E-dist reduction (end-to-end minus residual) against the average of the two errors for all 93 test perturbations at $\alpha = 1$, when the predicted means coincide. Positive values favor the residual flow (red), and negative values favor the end-to-end flow (blue). The labelled perturbation is the example shown in Figure 3.

A.5 Population controls using the same mean head

To examine whether a predicted mean together with control-derived variation is sufficient to recover the response population, Table 5 compares the full model at $\alpha = 1$ against two populations built from the same closed-form head. HeadGauss draws cells from $\hat{\mu}_p + \mathcal{N}(0, \mathrm{diag}(\sigma^2_{\mathrm{ctrl}}))$, whereas HeadResamp adds centered control cells resampled per output cell, $\hat{\mu}_p + (c_i - \bar{c})$. The former retains control-gene variances independently, while the latter samples the joint variation present in control cells.

Table 5: **Population controls using the same mean head** (E-dist, ↓, seed 1); lowest reported value per screen in bold. The controls share the head mean in expectation; only the full model at $\alpha = 1$ fixes the empirical mean exactly.

Screen	scPertWGF ($\alpha = 1$)	HeadGauss	HeadResamp
Norman	**0.80**	0.98	1.00
Adamson	**0.46**	0.49	0.50
Replogle	1.74	**1.03**	1.04
Schmidt	**2.24**	3.00	2.90
SciPlex	**2.44**	2.72	2.77

The full model has lower E-dist on four screens, indicating that these two uses of control variation do not reproduce the agreement achieved by the learned generator on those screens. On Replogle, both controls instead outperform the full model; Appendix A.6 examines its population objective in more detail. The controls share the head mean in expectation but retain finite-sample variation in their empirical means, whereas guidance at $\alpha = 1$ fixes the full model's empirical mean to $\hat{\mu}_p$. These comparisons therefore assess population shape together with empirical-mean calibration, rather than isolating shape alone.

A.6 Population-objective analysis on Replogle

Replogle separates accurate mean prediction from distributional agreement: the full model leads the point metrics in Table 1, but not E-dist. To investigate this discrepancy, Table 6 compares variants at $\alpha = 0.7$ with the same width, condition set, and training budget. Removing the adversarial component alone lowers E-dist from 1.73 to 1.03 and raises the median ratio of generated to measured total variance from 0.54 to 0.78.

Table 6: **Population-objective analysis on Replogle** (seed 1). Flow variants use $\alpha = 0.7$. HeadGauss samples around the same head prediction; its expected mean is the head mean.

Replogle variant or reference	E-dist ↓
Full model	1.73
w/o adversarial component	**1.03**
w/o WGF objective	1.89
HeadGauss control	**1.03**

For the full model, the interdecile range of this ratio is 0.47–0.58 over the 323 conditions, indicating consistently reduced total variation in the generated populations. The improvement without the adversarial component coincides with a closer match in total variance.

Removing the full WGF objective instead raises E-dist to 1.89. The effect of removing the adversarial component thus differs from that of removing both the adversarial and proximal components. This result cautions against attributing the effect of the combined objective to each component individually: on this screen, the full objective improves on its removal even though its adversarial component can worsen population agreement.

A.7 Guidance selection

The guidance sweep in Table 3 evaluates the relative contributions of the flow mean and the head mean with training fixed. At $\alpha = 1$, the generated empirical mean equals the head prediction, so the point metrics also provide a mean-head-only reference; the residual flow still determines the centered population distribution. At $\alpha = 0$, the head-centered residual model remains in place, but its generated mean is not calibrated after sampling. This distinguishes removal of guidance from removal of the mean head during training.

The reported guidance values are accompanied by validation analyses. On the genetic screens we carve a stratified 15% of training conditions as validation and retrain without them (Norman: 20 of 133, 12 singles and 8 doubles; Adamson: 9 of 61 singles; Schmidt: 18 of 116, 9 singles and 9 doubles); on SciPlex the official *seen* conditions contribute held-out cells that never enter training, so no retraining is required. The validation Pearson delta curve selects the reported operating point on every screen where it is computed: $\alpha = 0.5$ on Norman, 0.7 on Adamson, and 1.0 on Schmidt and SciPlex. The Replogle operating point follows the Adamson value. The hidden width follows a data-scale prior, 256 on the compact screens and 2048 on the genome-scale and drug screens, and is not tuned per screen.

A.8 Robustness across data splits

Table 7 reports the mean and standard deviation over three official data splits to assess performance beyond the main-table partition. The guidance scalar and hidden width follow the seed-1 selection ($\alpha = 0.5$ Norman, 0.7 Adamson and Replogle, 1.0 Schmidt and SciPlex) and are kept fixed across seeds.

The reported variation combines changes in the held-out conditions with variation across training runs. On SciPlex, the three splits contain 36, 148, and 152 held-out conditions, respectively, so their scores average over different test populations. These results characterize sensitivity to the benchmark partition as well as training variability; they do not isolate sensitivity to random initialization.

	scPertWGF				
Screen	Pearson $\Delta \uparrow$	MSE$\downarrow$	MAE$\downarrow$	L2$\downarrow$	E-dist$\downarrow$
Norman	0.699±0.034	0.003±0.001	0.023±0.001	3.43±0.20	0.66±0.05
Adamson	0.583±0.031	0.002±0.001	0.020±0.001	3.02±0.10	0.49±0.04
Replogle	0.309±0.008	0.004±0.001	0.040±0.001	4.38±0.07	1.73±0.11
Schmidt	0.373±0.014	0.007±0.001	0.045±0.001	5.72±0.16	2.22±0.04
SciPlex	0.473±0.039	0.008±0.001	0.024±0.003	6.13±0.19	2.42±0.47

Table 7: **Seed robustness.** Mean±standard deviation of scPertWGF over the official split seeds 1, 2, and 3 on all screens, with α and the hidden width kept at the seed-1 selection.