MECHANISM-GROUNDED FLOW MATCHING FOR INTERPRETABLE SINGLE-CELL PERTURBATION PREDICTION WITH ADAPTIVE TRANSPORT DYNAMICS

Anonymous authors
Paper under double-blind review

ABSTRACT

Single-cell perturbation prediction is essential for systems biology and therapeutic discovery, where interpretability is crucial for both grounding transcriptional responses in biological mechanisms and understanding how mechanisms shape cellular state transitions. Recently, flow matching has emerged as an effective approach for modeling continuous transitions between cellular states. However, existing methods primarily use it to predict perturbed expression profiles, without explicitly grounding the learned transitions in underlying biological mechanisms, thus limiting interpretability. To solve this problem, in this paper, we propose **MECHFLOW**, a mechanism-grounded flow matching framework that adaptively models the dynamics of transport from untreated to perturbed states. Specifically, we propose three modules: *target-anchored mechanism grounding for perturbation drift* propagates perturbation effects through directed biological graphs to derive a mechanism drive toward transcriptional changes, termed the perturbation drift; *graph metric adaptation for state-responsive cellular geometry* calculates a graph metric to define an adaptive geometry that assigns different costs to directions of transcriptional change according to gene associations, and *transport dynamics for interpretable distribution flow* integrates the drift and graph geometry to construct interpretable transition flows for supervising velocity field learning in flow matching. Together, these modules enable **MECHFLOW** to ground predicted transcriptional responses in biological mechanisms and characterize how they unfold along cellular state transitions. Experiments on Tahoe-100M show that **MECHFLOW** achieves not only strong predictive performance under various settings but also interpretability at both the response and transition levels.

1 INTRODUCTION

Predicting single-cell responses to perturbations is a central goal of systems biology and virtual-cell modeling, with broad applications in perturbation screening and therapeutic discovery (Zhang et al., 2026). Existing methods learn perturbation effects from observed single-cell profiles through representation learning and generative modeling, including flow matching, to predict cellular responses under unmeasured conditions(Lotfollahi et al., 2023; Yu et al., 2026; Klein et al., 2025).

Interpretability, which helps to establish the biological plausibility of predicted responses, is essential for biomedical research and potential clinical translation. Such interpretability comprises two complementary levels. At the response level, it requires grounding intracellular transcriptional changes in candidate targets, pathways, and biological processes(Berger & Iyengar, 2009). At the transition level, it concerns interpreting how underlying mechanisms shape cellular transitions from untreated to perturbed states.

However, existing methods primarily focus on predicting perturbed expression profiles (Ahlmann-Eltze et al., 2025), with either biological priors, such as interaction graphs used to guide representation learning, or continuous flows learned without explicit mechanism grounding, thus failing to provide interpretability at both the response and transition levels. The problem of interpretable single-cell perturbation prediction is highly non-trivial due to the following challenges:

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.

1. How can diverse gene-level perturbation responses be grounded in candidate molecular targets, pathways, and regulatory mechanisms?
2. How can the coordination among gene expressions be modeled adaptively when their associations vary across cellular contexts?
3. How can mechanism-grounded effects and state-dependent gene coordination be integrated into an explicit, analyzable model of transitions between untreated and perturbed cell populations?

To solve this problem, in this paper, we propose **MechFlow**: **Mech**anism-Grounded **Flow** Matching for Interpretable Single-Cell Perturbation with Adaptive Transport Dynamics. It learns a continuous flow matching velocity field from cell distributions, not only grounds predicted perturbation effects in regulatory mechanisms but also adapts their transport to cellular context, thereby enabling interpretable prediction across both levels.

In concrete terms, the framework comprises three modules. First, *Target-Anchored Mechanism Grounding for Perturbation Drift* anchors perturbations to molecular targets and propagates their effects through a directed biological graph to derive a signed gene-level drive toward expression changes, termed the perturbation drift. Second, *Graph Metric Adaptation for State-Responsive Cellular Geometry* calculates a graph metric from cellular profiles and fixed prior graphs, defining an adaptive geometry that penalizes discordant transcriptional changes between structurally connected genes. Third, *Transport Dynamics for Interpretable Distribution Flow* integrates the signed perturbation drift and adaptive graph geometry to construct continuous transitions from untreated to perturbed cellular states, whose velocities supervise flow matching to learn the corresponding transport field. In summary, the main contributions of this paper lie as follows:

- We propose to study interpretable single-cell perturbation prediction at two levels: grounding predicted transcriptional responses in regulatory mechanisms, and explaining how mechanism-guided effects shape transitions from untreated to perturbed cellular states.
- We propose **MechFlow**, a flow matching framework comprising three modules: i) target-anchored mechanism grounding to identify perturbation-driven transcriptional effects, ii) graph metric adaptation to characterize state-dependent cellular geometry, and iii) the formulated transport dynamics to construct interpretable transcriptomic transition flows.
- We conduct extensive experiments to demonstrate not only the superior predictive performance of MECHFLOWbut also its interpretability in grounding predicted responses in perturbation-related regulatory mechanisms and characterizing the generated transport dynamics.

2 RELATED WORK

Distributional transport and flow matching. Because destructive single-cell assays observe untreated and perturbed cells as unpaired populations, perturbation prediction naturally admits a distributional transport formulation (Bunne et al., 2023). Recent methods therefore learn conditional transport maps or generative flows between cellular populations (Driessen et al., 2026; Yu et al., 2026; Klein et al., 2025), allowing perturbation effects to be modeled at the population level. More broadly, Flow Matching studies show that the choice of endpoint coupling and underlying geometry can directly alter the probability path and its velocity targets (Lipman et al., 2023; Tong et al., 2024; Chen & Lipman, 2024). These developments make the structure of the learned transport itself an important modeling choice rather than merely an implementation detail. However, existing transport-based approaches generally learn cellular transitions without explicitly grounding the transport dynamics in perturbation-specific biological mechanisms.

Biological graphs for single-cell perturbation prediction. Biological interaction graphs provide useful inductive priors for perturbation prediction by encoding known relationships among genes and other molecular entities. Existing methods incorporate gene or biomedical interaction graphs through message passing, graph-based representation learning, or mechanism-aware conditioning (Roohani et al., 2024; Yu et al., 2026; Alsulami et al., 2026; Feng et al., 2026). Such priors can constrain information propagation and provide a structured biological context for predicting perturbation responses. However, biological structure is typically used within representation or conditioning components rather than to explicitly structure the transition from untreated to perturbed cell distributions.

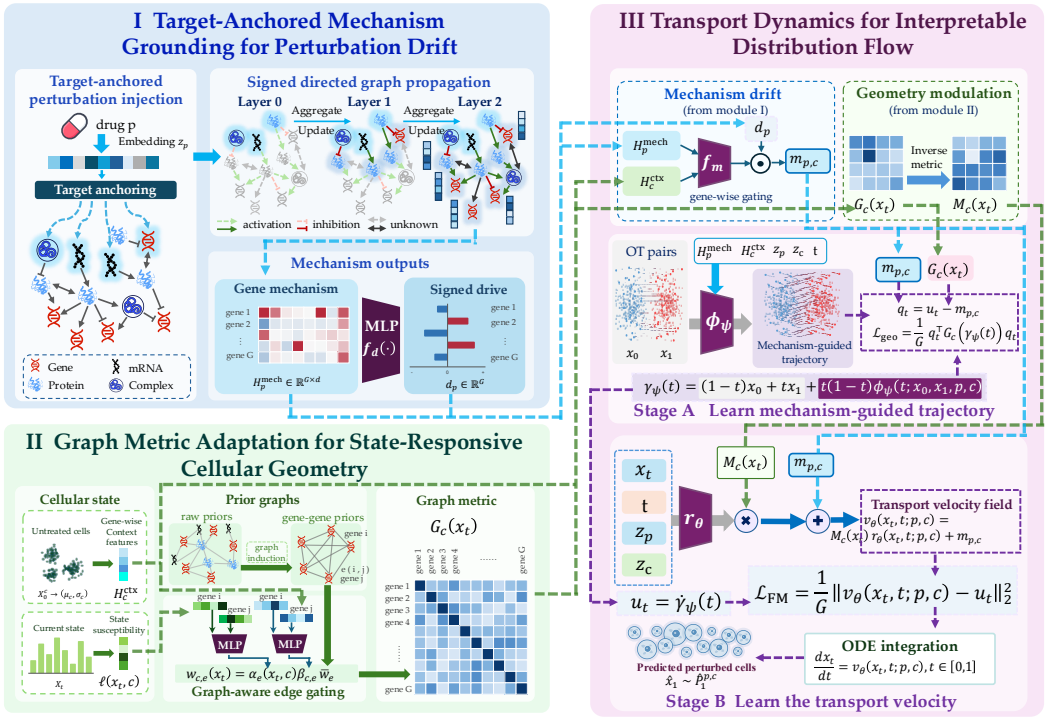

Figure 1: Overview of MECHFLOW. Our MECHFLOW enables interpretable single-cell perturbation prediction through three modules: (1) Target-Anchored Mechanism Grounding for Perturbation Drift; (2) Graph Metric Adaptation for State-Responsive Cellular Geometry; (3) Transport Dynamics for Interpretable Distribution Flow.

3 METHOD

3.1 PROBLEM FORMULATION AND FRAMEWORK

Let G be the number of measured genes and $x \in \mathbb{R}^G$ a single-cell expression vector. For cellular context c, untreated cells follow $x_0 \sim P_0^c$, while cells after perturbation p follow $x_1 \sim P_1^{p,c}$. We learn a conditional continuous-time transport

$$\frac{dx_t}{dt} = v_\theta(x_t, t; p, c), \qquad t \in [0, 1], \tag{1}$$

where $x_t \in \mathbb{R}^G$ is the transported expression state at time t and $v_\theta : \mathbb{R}^G \times [0,1] \to \mathbb{R}^G$ is the conditional velocity field with learnable parameters θ. The flow maps P_0^c toward $P_1^{p,c}$. Because single-cell assays provide unpaired populations, training endpoints $(x_0, x_1) \sim \pi_{p,c}$ are obtained with minibatch Sinkhorn optimal transport(Cuturi, 2013; Fatras et al., 2020).

MechFlow contains three modules. Module I propagates perturbation information through a directed biological graph to derive a signed gene-level perturbation drive, which is subsequently adapted to cellular context to form the mechanism-grounded drift. Module II combines structural gene associations with cellular context and the evolving expression state to define an adaptive graph metric. Module III integrates the mechanism-grounded drift with the adaptive geometry to construct endpoint-preserving training trajectories and learn the corresponding transport velocity field with Flow Matching. Figure 1 summarizes the complete framework.

3.2 TARGET-ANCHORED MECHANISM GROUNDING FOR PERTURBATION DRIFT

Let $\mathcal{G}_\to = (\mathcal{V}_\to, \mathcal{E}_\to)$ be a directed mechanism graph containing measured genes and biological intermediates such as proteins and transcription factors. Edge types encode biological relations and, when available, regulatory sign and evidence strength. Graph construction is detailed in Appendix B.

For perturbation p, a molecular fingerprint $\xi_p \in \mathbb{R}^{d_{\mathrm{fp}}}$ is encoded by the molecular encoder E_{mol} as $z_p = E_{\mathrm{mol}}(\xi_p) \in \mathbb{R}^{d_z}$, where d_{fp} and d_z are the fingerprint and perturbation-embedding dimensions, respectively. Each targetable node v has an identity representation $h_v^{\mathrm{id}} \in \mathbb{R}^{d_g}$, where d_g is the graph-node feature dimension. It combines learned node and node-type embeddings, with fixed pretrained protein representations incorporated when available. We score perturbation–target compatibility as

$$a_{p,v} = a_{p,v}^{\mathrm{mol}} + b_{\mathrm{tar}} T_{p,v}, \qquad a_{p,v}^{\mathrm{mol}} = \frac{\kappa}{\tau} \frac{(W_q z_p)^\top (W_k h_v^{\mathrm{id}})}{\|W_q z_p\|_2 \, \|W_k h_v^{\mathrm{id}}\|_2}, \tag{2}$$

where $a_{p,v}$ is the final perturbation–target score and $a_{p,v}^{\mathrm{mol}}$ is its molecular-compatibility component. Here $T_{p,v} \geq 0$ denotes optional curated target support, $b_{\mathrm{tar}} \geq 0$ its strength, W_q, W_k are learned query and key projections, $\tau > 0$ is the compatibility temperature, and $\kappa > 0$ is a logit scale. For the top-K targetable nodes S_p, we define the injection weights as

$$\omega_{p,v} = \begin{cases} \dfrac{\exp(a_{p,v})}{\sum_{u \in S_p} \exp(a_{p,u})}, & v \in S_p, \\ 0, & v \notin S_p. \end{cases} \tag{3}$$

We use $\omega_{p,v}$ to inject the perturbation representation into the selected targetable nodes:

$$h_{p,v}^{(0)} = \omega_{p,v}(W_{\mathrm{inj}} z_p + h_v^{\mathrm{id}} + s_{p,v}^{\mathrm{tar}} W_{\mathrm{sign}} z_p), \qquad A_{p,v}^{(0)} = \mathbf{1}[\omega_{p,v} > 0]. \tag{4}$$

Here, W_{inj} projects the perturbation embedding into the node-feature space, and $s_{p,v}^{\mathrm{tar}} \in \{-1, 0, +1\}$ specifies the curated target-interaction sign when available. Thus, only selected nodes receive an initial perturbation signal; the active mask $A_{p,v}^{(0)}$ then determines where directed propagation begins. The exact initialization and message rules are given in Appendix C.1. Subsequent sparse message passing propagates this signal through the directed mechanism graph:

$$\begin{aligned} \bar{h}_{p,v}^{(\ell)} &= \frac{\sum_{e:u \to v} a_{p,e} \widetilde{h}_e^{(\ell)}}{\max(1, \sum_{e:u \to v} a_{p,e})}, \\ h_{p,v}^{(\ell+1)} &= \mathrm{LN}\Big(h_{p,v}^{(\ell)} + U(\bar{h}_{p,v}^{(\ell)})\Big), \qquad \ell = 0, \ldots, L-1. \end{aligned} \tag{5}$$

Here $a_{p,e} \geq 0$ combines edge confidence, drug-dependent relevance, and source activity; $\widetilde{h}_e^{(\ell)}$ applies a relation-specific transform with a known regulatory sign or a separate unknown-sign transform. U is a learned update, LN is layer normalization, and inactive nodes remain zero.

After the final propagation layer, we denote the gated representation of measured gene i by $h_{p,i}^{\mathrm{mech}}$, obtained from the final message-passing state $h_{p,i}^{(L)}$. Collecting these gene-wise representations gives

$$H_p^{\mathrm{mech}} = [h_{p,1}^{\mathrm{mech}}, \ldots, h_{p,G}^{\mathrm{mech}}]^\top \in \mathbb{R}^{G \times d_g}, \qquad [d_p]_i = f_d(h_{p,i}^{\mathrm{mech}}), \tag{6}$$

where $f_d : \mathbb{R}^{d_g} \to \mathbb{R}$ is shared across genes, unreachable genes have zero drive, and $d_p \in \mathbb{R}^G$ is the resulting signed perturbation drive. Implementation details, including the final gate and the optional target-alignment objective, are given in Appendix C.1.

3.3 Graph Metric Adaptation for State-Responsive Cellular Geometry

For context c, let $X_0^c = \{x_0^{(n)}\}_{n=1}^{N_c}$ be the set of N_c untreated cells, with gene-wise basal mean $\mu_c \in \mathbb{R}^G$ and standard deviation $\sigma_c \in \mathbb{R}^G$. Over the undirected positive structural graph $\mathcal{G}_{\leftrightarrow}$, we encode each gene's basal statistics together with a trust-gated summary of its neighbors. This yields gene-resolved $H_c^{\mathrm{ctx}} = [h_{c,1}^{\mathrm{ctx}}, \ldots, h_{c,G}^{\mathrm{ctx}}]^\top \in \mathbb{R}^{G \times d_g}$; a separate global encoder gives $z_c = E_{\mathrm{ctx}}([\mu_c; \sigma_c]) \in \mathbb{R}^{d_c}$. The trust gate limits how much neighborhood information enters each gene, and isolated genes use only their own basal features. The exact aggregation and source balancing are given in Appendix C.2.

The STRING and Reactome views retain separate edges; their positive weights are mixed equally into $\bar{w}_e > 0$. We factor each edge weight as

$$w_{c,e}(x) = \bar{w}_e \, \beta_{c,e} \, \alpha_e(x, c) > 0, \tag{7}$$

The context factor $\beta_{c,e} > 0$ is learned from the untreated-context representations of the two endpoint genes through a symmetric edge scorer; its full parameterization is given in Appendix C.2. To

respond to the evolving expression state, we derive a bounded, zero-centered gene susceptibility $\ell_i(x, c)$ from each gene's deviation relative to untreated cells (Appendix C.2). The corresponding positive state factor is

$$\alpha_e(x, c) = \exp\left(\frac{\ell_i(x, c) + \ell_j(x, c)}{2}\right), \tag{8}$$

so the same structural edge can be reweighted differently across contexts and along the transport trajectory. The exact susceptibility parameterization and centering constraints are provided in Appendix C.2.

Let $E = |\mathcal{E}_{\leftrightarrow}|$, $B \in \mathbb{R}^{G\times E}$ be an oriented incidence matrix, and $w_c(x) \in \mathbb{R}^E$ collect the weights in Equation 7. We define

$$\boxed{\mathsf{G}_c(x) = I + \lambda B \,\mathrm{Diag}(w_c(x)) B^\top}, \qquad \lambda > 0, \tag{9}$$

where $I \in \mathbb{R}^{G\times G}$ is the identity matrix, $\mathrm{Diag}(w_c(x))$ places the adaptive edge weights on the diagonal, and $\lambda > 0$ controls the strength of graph-induced geometry. The metric is symmetric positive definite because every edge weight is positive. For any expression-space direction $q \in \mathbb{R}^G$,

$$q^\top \mathsf{G}_c(x) q = \|q\|_2^2 + \lambda \sum_{e=(i,j)} w_{c,e}(x)(q_i - q_j)^2. \tag{10}$$

Thus, strongly weighted edges make discordant motion between connected genes more costly, while the metric itself adapts with context and the evolving state.

3.4 Transport Dynamics for Interpretable Distribution Flow

We first adapt the perturbation drive to the context with a bounded gene-wise gate

$$g_{p,c,i} = 1 + \eta_m \tanh f_m([h_{p,i}^{\mathrm{mech}}, h_{c,i}^{\mathrm{ctx}}]), \qquad m_{p,c} = \gamma_m(g_{p,c} \odot d_p) + \delta_{p,c}^{\mathrm{cond}}, \tag{11}$$

where f_m is a shared gene-wise gating network, $g_{p,c} = [g_{p,c,1}, \ldots, g_{p,c,G}]^\top$ is the context-dependent gate, $\eta_m \in [0, 1]$ bounds its modulation amplitude, $\gamma_m \geq 0$ is a learned global drift scale, and $\odot$ denotes elementwise multiplication. The optional zero-initialized term $\delta_{p,c}^{\mathrm{cond}} \in \mathbb{R}^G$ captures residual condition-level effects only in the configurations specified in Appendix C. The resulting $m_{p,c} \in \mathbb{R}^G$ is the signed mechanism-grounded drift.

Stage A: graph-shaped probability path. For a coupled endpoint pair (x_0, x_1), we use an endpoint-preserving path

$$\gamma_\psi(t) = (1 - t)x_0 + tx_1 + t(1 - t)\phi_\psi(t; x_0, x_1, p, c), \tag{12}$$

where $\phi_\psi : \cdot \mapsto \mathbb{R}^G$ is a bounded path-correction network with parameters ψ, conditioned on the current linear interpolation, endpoint displacement, basal statistics, gene-wise mechanism/context features, transport time, and global perturbation/context representations. Its exact tangent is

$$\dot\gamma_\psi(t) = (x_1 - x_0) + (1 - 2t)\phi_\psi(t) + t(1 - t)\partial_t\phi_\psi(t), \tag{13}$$

with $\partial_t\phi_\psi(t)$ obtained by automatic differentiation. Let $u_t = \dot\gamma_\psi(t)$ and define the residual tangent $q_t = u_t - m_{p,c}$. Stage A samples $t \sim \mathrm{Unif}[0, 1]$ and minimizes

$$\boxed{\mathcal{L}_{\mathrm{geo}} = \mathbb{E}_{(x_0,x_1)\sim\pi_{p,c},\, t\sim\mathrm{Unif}[0,1]}\left[\frac{1}{G} q_t^\top \mathsf{G}_c(\gamma_\psi(t)) q_t\right].} \tag{14}$$

The path therefore preserves the observed endpoints while assigning a lower cost to residual motion compatible with the adaptive graph geometry after accounting for the signed drift.

Stage B: Flow Matching. After Stage A, the path-network and geometry parameters are frozen. At each Stage-B step, current condition encodings produce (x_t, u_t) through the frozen path network; this pair is detached before the Flow Matching update. The residual branch centers x_t by the untreated basal mean and combines the centered state with transport time to obtain a hidden representation h_t. We decompose the residual force $r_\theta(x_t, t; p, c) \in \mathbb{R}^G$ into shared, perturbation-specific, and context-specific components,

$$r_\theta(x_t, t; p, c) = f_0(h_t) + f_p([h_t, z_p]) + f_c([h_t, z_c]). \tag{15}$$

where f_0, f_p, and f_c are learned output heads for the shared, perturbation-conditioned, and context-conditioned residual components, respectively, and θ denotes their collective parameters. The grounded drift is excluded from this residual branch and is added only after the optional metric transformation:

$$\boxed{v_\theta(x_t, t; p, c) = \mathcal{M}_c(x_t) r_\theta(x_t, t; p, c) + m_{p,c},} \qquad \mathcal{M}_c(x) \in \{I, \mathsf{G}_c(x)^{-1}\}, \tag{16}$$

where $\mathcal{M}_c(x)$ is the mobility operator: I leaves the residual velocity unchanged, while $\mathsf{G}_c(x)^{-1}$ applies inverse-metric geometry. The chosen operator for each evaluation setting is specified in Appendix E. The primary Flow-Matching loss is

$$\mathcal{L}_{\mathrm{FM}} = \mathbb{E}\left[\frac{1}{G}\|v_\theta(x_t, t; p, c) - u_t\|_2^2\right]. \tag{17}$$

Population-mean and task-specific weak auxiliary terms, initialization details, and the two-stage training algorithm are deferred to Appendices C–D.

Inference. Starting from untreated cells, predictions are generated by integrating Equation 1 with the learned velocity field in Equation 16. Metric-related linear systems are solved without explicitly forming $\mathsf{G}_c(x)^{-1}$; implementation details are provided in Appendix E.

4 Experiments

4.1 Experimental Setup

Dataset and splits. We evaluate on the middle-dose subset of Tahoe-100M (Zhang et al., 2026) using a fixed ordered 2,000-gene panel and matched untreated controls. We consider P-holdout (unseen perturbations) and C-holdout (unseen cellular contexts) as the primary settings; the frozen test sets contain 2,998 and 3,197 perturbation–context conditions, respectively.

Baselines. We compare with Identity and Linear-MLP, a matched Vanilla FM backbone, and external perturbation/transport models whose interfaces support each holdout, including chemCPA (Hetzel et al., 2022), PRnet (Qi et al., 2024), STATE (Adduri et al., 2026), CellFlow (Klein et al., 2025), scDFM (Yu et al., 2026), and PerturbDiff (Yuan et al., 2026). Simple baselines are retained because recent benchmarking shows that sophisticated perturbation predictors can fail to outperform carefully constructed linear references (Ahlmann-Eltze et al., 2025).

Metrics. For each perturbation–context condition, Mean-MSE compares predicted and observed treated-cell means, while Δ-Pearson correlates their changes relative to matched controls. DE50 denotes the 50 genes with the largest absolute observed changes; on these genes we report rank correlation (DE-Spearman), sign agreement (DE50 direction), and Pearson correlation (DE50 Pearson). RBF MMD^2 measures the discrepancy between predicted and observed cell distributions in training-derived PCA50 space. All metrics are macro-averaged across conditions. More details are included in Appendix F.

4.2 Main Results

Table 1: P-holdout results. Best results are **bold**, and second-best results are underlined.

Method	MSE↓	Δ-Pear.↑	DE-Spear.↑	DE50 Dir.↑	DE50 r↑	MMD^2↓
Identity	0.01897	-0.0328	0.0016	0.4729	-0.0030	0.01761
Linear-MLP	**0.01580**	0.3672	0.3675	0.7167	0.4431	0.01649
chemCPA	0.14893	0.2755	0.2817	0.6626	0.3118	0.30545
PRnet	0.11369	0.2709	0.2832	0.6671	0.3154	0.23171
PerturbDiff	0.07302	0.2803	0.2734	0.6841	0.3029	0.10077
Vanilla FM	0.01800	0.1789	0.2217	0.6291	0.3102	0.01643
MechFlow	0.01716	**0.4144**	**0.4418**	**0.7669**	**0.5440**	**0.01577**

Table 2: C-holdout results. Best results are **bold**, and second-best results are underlined.

Method	MSE↓	Δ-Pear.↑	DE-Spear.↑	DE50 Dir.↑	DE50 r↑	MMD^2↓
Identity	0.01812	-0.0370	-0.0141	0.4727	-0.0176	0.01623
P-Mean	0.01565	0.3225	0.3603	0.7098	0.4605	0.01325
Linear-MLP	0.01752	0.3081	0.3150	0.6911	0.3495	0.01608
PRnet	0.18191	0.2890	0.3067	0.6819	0.3310	0.26076
STATE	0.18894	0.2962	0.3134	0.6825	0.3394	0.29042
CellFlow	0.02313	0.0272	0.0765	0.5168	0.1101	0.01756
scDFM	0.33586	0.2205	0.2655	0.6008	0.2613	0.55740
Vanilla FM	0.01775	0.1299	0.1792	0.6001	0.2312	0.01518
MECHFLOW	**0.01413**	**0.4062**	**0.3834**	**0.7811**	**0.4828**	**0.01147**

P-holdout. **MechFlow** achieves the strongest performance across all four effect-oriented metrics and the best distributional agreement measured by MMD^2. Although Linear-MLP attains the lowest Mean-MSE, **MechFlow** more accurately recovers perturbation-specific differential-expression patterns and consistently outperforms the matched Vanilla FM backbone. These results suggest that mechanism-grounded transport is particularly beneficial for recovering the direction and structure of unseen perturbation effects rather than merely minimizing global reconstruction error.

C-holdout. **MechFlow** achieves the best performance across all six metrics, improving both global prediction accuracy and perturbation-specific differential-expression recovery. Perturbation-Mean remains a strong baseline under context holdout, indicating that perturbation-level average responses transfer substantially across cellular contexts. Nevertheless, **MechFlow** consistently improves upon this baseline, suggesting that explicitly adapting the transport geometry to cellular context captures context-specific response variation beyond a perturbation-level mean response.

4.3 INTERPRETABILITY: MECHANISM GROUNDING AND TRANSPORT DYNAMICS

We analyze interpretability at two complementary levels. At the *response level*, we trace perturbation-specific propagation through the directed biological graph. At the *transition level*, we examine the learned transport from three views: its evolution over flow time, the composition of mechanism-grounded and residual dynamics, and its dependence on cellular context.

Response-level mechanism grounding. **MechFlow** first selects perturbation-dependent graph entry nodes and propagates their signals through directed biological relations before constructing d_p. This makes it possible to inspect where the perturbation signal is injected and which candidate graph routes receive high relevance during propagation. Figure 2 illustrates this process for Trimetrexate, Pralsetinib, and LB-100. The three perturbations activate different entry nodes and downstream neighborhoods, while retaining the regulatory signs available in the prior graph. The mechanism representation supplied to the transport model is traceable to perturbation-specific routes in the directed graph, providing gene-level candidate mechanistic support for each prediction.

Transition-level organization over flow time. We track the learned velocity field over normalized flow time to examine how predicted transport evolves across genes and pathways. In the C-holdout LB-100/CVCL_1635 case (Figure 3), different target genes peak at different stages of the trajectory. Reactome summaries show a similar temporal redistribution: RHOA GTPase remains prominent throughout, while interferon signaling is stronger earlier and FOXO transcription becomes more prominent later. These patterns expose how predicted transport changes as cells move toward the perturbed endpoint.

Mechanism–residual composition. Equation 16 separates the final velocity into mechanism-grounded drift $m_{p,c}$ and effective residual transport $\mathcal{M}_c(x_t)r_\theta(x_t, t; p, c)$, allowing their interaction to be inspected directly. In Figure 4, several genes with decreasing predicted expression show positive residual velocity but negative total velocity, indicating a stronger negative mechanism-grounded drift partially compensated by the residual. Across genes and flow times, the two components can

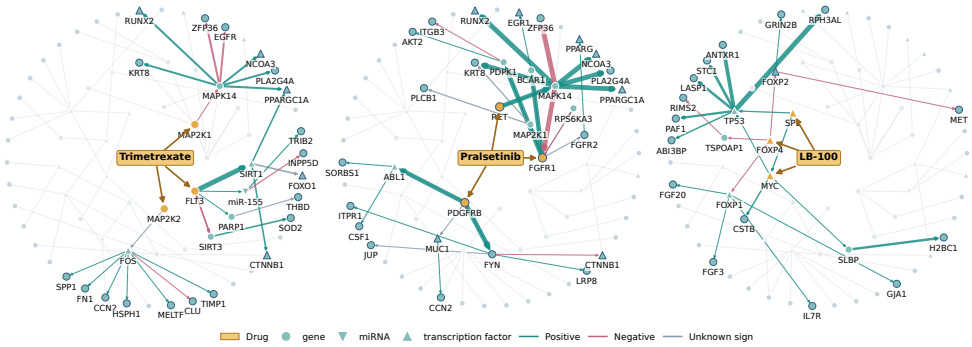

Figure 2: **Perturbation-specific directed mechanism propagation.** Local two-hop subgraphs are shown for Trimetrexate, Pralsetinib, and LB-100. Brown arrows denote model injection into graph entry nodes; edge color denotes known positive, known negative, or unknown regulatory sign; and edge width reflects active-source relevance averaged over the two propagation layers and rescaled within each panel. The visualization exposes perturbation-dependent routes used by Module I rather than establishing causal regulatory interactions.

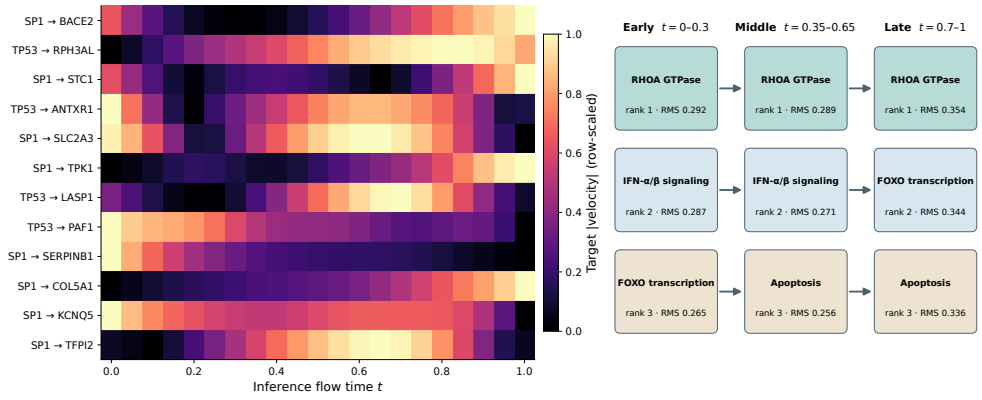

Figure 3: **Flow-time variation of target-level transport and pathway summaries.** For C-holdout LB-100/CVCL_1635, the left panel shows independently rescaled population-mean velocity magnitudes for genes reached by high relevance SP1/TP53 edges. The right panel shows the top Reactome gene sets across early, middle, and late flow intervals. Flow time denotes inference time.

reinforce or offset each other, exposing how biological structure and data-driven correction jointly shape the prediction.

Context-dependent transport. Appendix H compares LB-100 trajectories across three held-out contexts in a shared PCA space. Distinct displacement patterns show that the learned transport changes with cellular context through context-conditioned residual dynamics and adaptive geometry.

Together, these analyses connect candidate mechanism propagation with the gene-, pathway-, and context-level dynamics of the predicted transition. Additional visualizations and definitions are provided in Appendix H.

4.4 Ablation Study

We examine the mechanism and geometry components in separate 512-condition proxy studies with fixed cell sampling and unified metrics. Fig. 5 shows three representative metrics; full proxy results appear in Tables 6 and 7. The interventions behind each table row are specified in Appendix G.

P-holdout: perturbation mechanism ablations. For unseen perturbations, we ablate the signed graph mechanism, perturbation–target prior, and context-dependent modulation. All three interventions reduce perturbation-sensitive metrics while leaving global expression similarity comparatively

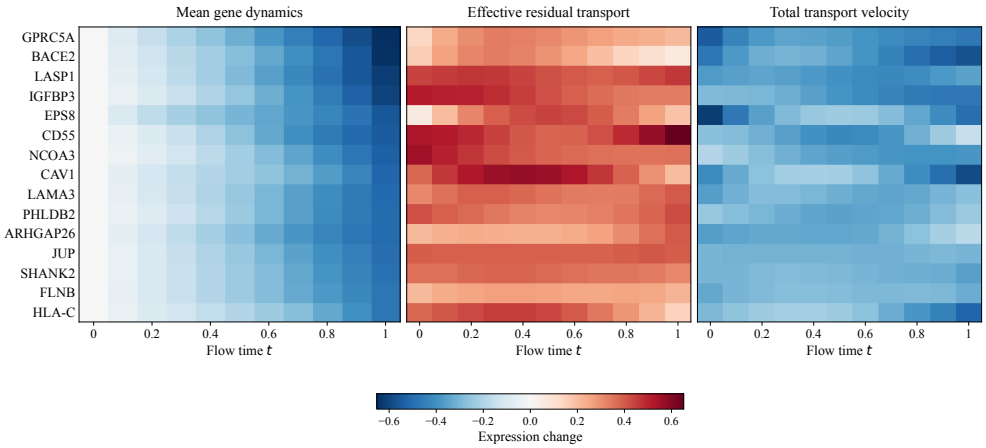

Figure 4: **Gene-level expression and transport dynamics.** Predicted expression displacement, effective residual transport $\mathcal{M}_c(x_t)r_\theta(x_t, t; p, c)$, and total velocity $v_\theta(x_t, t; p, c)$ for LB-100 in CVCL_1635. Their comparison reveals the contribution of mechanism-grounded drift to the final transport.

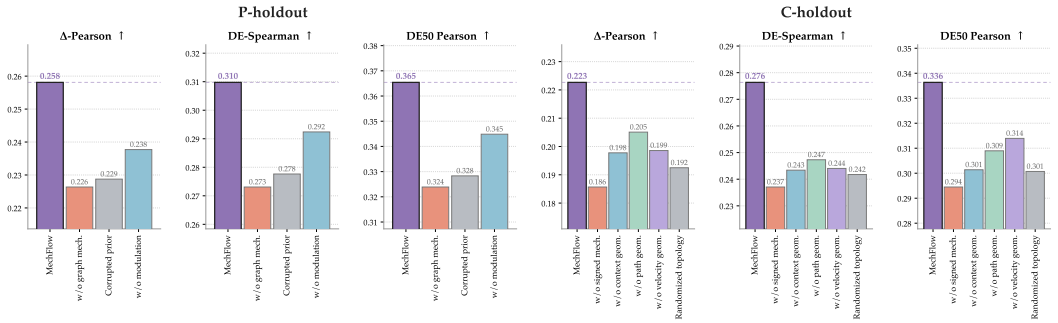

Figure 5: Ablation results on P-holdout and C-holdout for three representative metrics showing the most pronounced differences across variants.

stable (Table 6). Corrupting the perturbation–target mapping degrades performance similarly to removing the graph mechanism, showing that biologically meaningful graph entry points contribute directly to generalization. Removing context modulation yields a smaller but consistent decline.

C-holdout: context and geometry ablations. For unseen contexts, we ablate the signed mechanism, graph-induced context geometry, path geometry, velocity geometry, and graph topology. Removing the signed mechanism and randomizing topology causes the clearest degradation in effect-level metrics (Table 7), supporting the importance of biologically structured propagation. Geometry ablations produce smaller, metric-dependent changes, indicating complementary contributions from context, path, and velocity geometry.

5 Conclusion

We propose **MechFlow**, a mechanism-grounded flow matching framework for interpretable single-cell perturbation prediction. By grounding predicted transcriptional responses in biological mechanisms and adaptively modeling the dynamics of transport from untreated to perturbed cellular states, **MechFlow** enables interpretability at both the response and transition levels. Across extensive experiments on Tahoe-100M, the model not only improves distributional prediction over matched baselines, but also enables inspection of candidate mechanisms and of how grounded and residual dynamics jointly shape predicted transitions.

Reproducibility Statement

We provide the model formulation and training procedure in Section 3, with additional algorithmic and implementation details in the appendix. The supplementary material documents biological graph construction, dataset preprocessing and splits, task-specific configurations, training and inference procedures, and the unified evaluation protocol used for the reported experiments. We additionally provide anonymized source code, including graph- and data-processing utilities, model and training implementations, evaluation routines, experiment configurations, and scripts for reproducing the reported analyses and visualizations. Random seeds and relevant task-specific implementation choices are documented where applicable. We provide implementation details, experimental protocols, and additional analyses in the appendix. An anonymized core implementation of **MechFlow**, including preprocessing, training, and evaluation code, is available at `https://anonymous.4open.science/r/MechFlow-2027-0747`.

AI Use Statement

Generative AI tools were used to assist with preliminary literature search and exploration of related work, code development and debugging, and language polishing of the manuscript. The research questions, core methodological ideas, experimental decisions, execution and evaluation of the experiments, interpretation of the final evidence, and decisions regarding the content and claims of the paper were made by the authors. AI-assisted code and text were reviewed, tested, revised, and integrated by the authors rather than accepted without verification. References and bibliographic metadata used in the manuscript were checked against primary or official sources before inclusion. The authors take full responsibility for the correctness, originality, and final content of the submission, including all material produced with the assistance of generative AI.

References

Abhinav K. Adduri et al. Predicting cellular responses to perturbation across diverse contexts with State. *Cell*, 2026. doi: 10.1016/j.cell.2026.07.052.

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. doi: 10.1038/s41592-025-02772-6.

Reem Alsulami, Robert Lehmann, Sumeer A. Khan, et al. Predicting and interpreting cell-type-specific drug responses in the small-data regime using inductive priors. *Nature Machine Intelligence*, 8:461–473, 2026. doi: 10.1038/s42256-026-01202-2.

Sorin Avram, Cristian G. Bologa, Jayme Holmes, Giovanni Bocci, Thomas B. Wilson, Dac-Trung Nguyen, Ramona Curpan, Liliana Halip, Alina Bora, Jeremy J. Yang, Jeffrey Knockel, Suman Sirimulla, Oleg Ursu, and Tudor I. Oprea. DrugCentral 2021 supports drug discovery and repositioning. *Nucleic Acids Research*, 49(D1):D1160–D1169, 2021. doi: 10.1093/nar/gkaa997.

Seth I. Berger and Ravi Iyengar. Network analyses in systems pharmacology. *Bioinformatics*, 25 (19):2466–2472, 2009. doi: 10.1093/bioinformatics/btp465.

Charlotte Bunne, Stefan G. Stark, Gabriele Gut, et al. Learning single-cell perturbation responses using neural optimal transport. *Nature Methods*, 20:1759–1768, 2023. doi: 10.1038/s41592-023-01969-x.

Ricky T. Q. Chen and Yaron Lipman. Flow matching on general geometries. In *International Conference on Learning Representations*, 2024.

Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In *Advances in Neural Information Processing Systems*, volume 26, 2013.

Alice Driessen, Dhruva Abhijit Rajwade, Benedek Harsanyi, et al. Conditional monge gap enables generalizable single-cell perturbation modelling. *Nature Machine Intelligence*, 8:984–996, 2026. doi: 10.1038/s42256-026-01242-8.

Kilian Fatras, Younes Zine, Rémi Flamary, Rémi Gribonval, and Nicolas Courty. Learning with minibatch wasserstein: Asymptotic and gradient properties. In *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108, pp. 2131–2141. PMLR, 2020.

Jinghao Feng, Ziheng Zhao, Xiaoman Zhang, et al. A knowledge-driven framework for predicting single-cell responses for unprofiled drugs. *Nature Machine Intelligence*, 8:1478–1491, 2026. doi: 10.1038/s42256-026-01286-w.

Leon Hetzel, Simon Böhm, Niki Kilbertus, Stephan Günnemann, Mohammad Lotfollahi, and Fabian J. Theis. Predicting Cellular Responses to Novel Drug Perturbations at a Single-Cell Resolution. In *Advances in Neural Information Processing Systems*, volume 35, pp. 26711–26722, 2022. doi: 10.52202/068431-1937.

Dominik Klein, Jonas Simon Fleck, Daniil Bobrovskiy, et al. CellFlow enables generative single-cell phenotype modeling with flow matching. *bioRxiv*, 2025. doi: 10.1101/2025.04.11.648220.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. Flow matching for generative modeling. In *International Conference on Learning Representations*, 2023.

Mohammad Lotfollahi et al. Predicting cellular responses to complex perturbations in high-throughput screens. *Molecular Systems Biology*, 19(6):e11517, 2023. doi: 10.15252/msb.202211517.

Marija Milacic, Deidre Beavers, Patrick Conley, Chuqiao Gong, Marc Gillespie, Johannes Griss, Robin Haw, Bijay Jassal, Lisa Matthews, Bruce May, Robert Petryszak, Eliot Ragueneau, Karen Rothfels, Cristoffer Sevilla, Veronica Shamovsky, Ralf Stephan, Krishna Tiwari, Thawfeek Varusai, Joel Weiser, Adam Wright, Guanming Wu, Lincoln Stein, Henning Hermjakob, and Peter D'Eustachio. The Reactome pathway knowledgebase 2024. *Nucleic Acids Research*, 52(D1): D672–D678, 2024. doi: 10.1093/nar/gkad1025.

Sophia Müller-Dott, Eirini Tsirvouli, Miguel Vazquez, Ricardo O. Ramirez Flores, Pau Badia-I-Mompel, Robin Fallegger, Dénes Türei, Astrid Lægreid, and Julio Saez-Rodriguez. Expanding the coverage of regulons from high-confidence prior knowledge for accurate estimation of transcription factor activities. *Nucleic Acids Research*, 51(20):10934–10949, 2023. doi: 10.1093/nar/gkad841.

Xiaoning Qi, Lianhe Zhao, Chenyu Tian, et al. Predicting transcriptional responses to novel chemical perturbations using deep generative model for drug discovery. *Nature Communications*, 15: 9256, 2024. doi: 10.1038/s41467-024-53457-1.

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

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

Prisca Lo Surdo, Marta Iannuccelli, Klas Karis, Eleonora Meo, Parnian Omidi, Marta Tosoni, Simone Graziosi, Simona Panni, Maria Teresa Fiorillo, Luana Licata, Francesca Sacco, Benjamin M. Gyori, and Livia Perfetto. SIGNOR 4.0: the 2025 update with focus on phosphorylation data. *Nucleic Acids Research*, 54(D1):D682–D690, 2026. doi: 10.1093/nar/gkaf1237.

Damian Szklarczyk, Rebecca Kirsch, Mikaela Koutrouli, Katerina Nastou, Farrokh Mehryary, Radja Hachilif, Annika L. Gable, Tao Fang, Nadezhda T. Doncheva, Sampo Pyysalo, Peer Bork, Lars J. Jensen, and Christian von Mering. The STRING database in 2023: protein–protein association networks and functional enrichment analyses for any sequenced genome of interest. *Nucleic Acids Research*, 51(D1):D638–D646, 2023. doi: 10.1093/nar/gkac1000.

Alexander Tong, Kilian Fatras, Nikolay Malkin, Guillaume Huguet, Yanlei Zhang, Jarrid Rector-Brooks, Guy Wolf, and Yoshua Bengio. Improving and generalizing flow-based generative models with minibatch optimal transport. *Transactions on Machine Learning Research*, 2024.

Dénes Türei, Alberto Valdeolivas, Lejla Gul, Nicolàs Palacio-Escat, Michal Klein, Olga Ivanova, Márton Ölbei, Attila Gábor, Fabian J. Theis, Dezső Módos, Tamás Korcsmáros, and Julio Saez-Rodriguez. Integrated intra- and intercellular signaling knowledge for multicellular omics analysis. *Molecular Systems Biology*, 17(3):e9923, 2021. doi: 10.15252/msb.20209923.

Susan Tweedie, Bryony Braschi, Kristian Gray, Tamsin E. M. Jones, Ruth L. Seal, Bethan Yates, and Elspeth A. Bruford. Genenames.org: the HGNC and VGNC resources in 2021. *Nucleic Acids Research*, 49(D1):D939–D946, 2021. doi: 10.1093/nar/gkaa980.

UniProt Consortium. UniProt: the universal protein knowledgebase in 2023. *Nucleic Acids Research*, 51(D1):D523–D531, 2023. doi: 10.1093/nar/gkac1052.

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.

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

Jesse Zhang, Airol A. Ubas, Valentine Svensson, et al. Tahoe-100M: Mapping drug-induced molecular phenotypes at single-cell resolution. *Cell*, 189(19):5945–5961.e8, 2026. doi: 10.1016/j.cell.2026.08.035.

A NOTATION AND SCOPE

Table 3 summarizes the principal symbols used throughout the paper. The current formulation separates three distinct components: a perturbation-grounded signed drift produced from the directed mechanism graph, a context- and state-dependent metric constructed from the undirected structural graph, and an endpoint-preserving probability path used to supervise the final Flow-Matching vector field.

Table 3: Principal notation used in MECHFLOW.

Symbol	Meaning
G	number of measured genes
E	number of edges in the undirected structural graph
p, c	perturbation and cellular context
$x_0, x_1 \in \mathbb{R}^G$	untreated and perturbed expression states
$t \in [0, 1]$	normalized flow time
$\pi_{p,c}$	empirical endpoint coupling used for training
$\mathcal{G}_{\rightarrow}$	directed heterogeneous mechanism graph
$\mathcal{G}_{\leftrightarrow}$	undirected positive structural graph
ξ_p	molecular fingerprint of perturbation p
z_p	global perturbation representation
H_p^{mech}	gene-wise mechanism representation
d_p	signed perturbation drive from Module I
μ_c, σ_c	untreated-context gene-wise mean and standard deviation
H_c^{ctx}	gene-wise context representation
z_c	global context representation
$\bar{w}_e$	fixed positive prior weight of structural edge e
$\beta_{c,e}$	positive context-dependent edge factor
$\alpha_e(x, c)$	positive state-dependent structural-edge factor
$w_{c,e}(x)$	adaptive structural-edge weight
$B \in \mathbb{R}^{G \times E}$	oriented incidence matrix of the structural graph
$\mathsf{G}_c(x)$	context- and state-dependent positive-definite metric
$g_{p,c}$	gene-wise context gate on the perturbation drive
$\delta_{p,c}^{\mathrm{cond}}$	optional condition-residual drift
$m_{p,c}$	context-adapted mechanism-grounded drift
$\phi_\psi(t)$	learned endpoint-preserving path correction
$\gamma_\psi(t)$	learned training probability path
$u_t = \dot{\gamma}_\psi(t)$	exact tangent of the training path
$q_t = u_t - m_{p,c}$	residual tangent used in the geometric path objective
r_θ	residual neural vector field
$\mathcal{M}_c(x)$	optional metric mobility operator
v_θ	inference-time conditional Flow-Matching vector field

B BIOLOGICAL GRAPH CONSTRUCTION

B.1 COMPLEMENTARY BIOLOGICAL GRAPH PRIORS

MechFlow uses a *directed mechanism graph* and an *undirected structural graph* as complementary biological priors. The directed graph represents where a perturbation may enter the molecular system and how signed effects can propagate through biological relations. The undirected graph defines the positive structural backbone of the adaptive metric over measured genes, specifying which coordinated expression changes receive graph-geometric transport costs. Activation and inhibition are represented only in the directed graph and the resulting signed mechanism drift, while the structural graph contains non-negative support to preserve positive definiteness of the graph-induced metric.

B.2 DIRECTED HETEROGENEOUS MECHANISM GRAPH

The directed mechanism graph integrates CollecTRI (Müller-Dott et al., 2023), OmniPath (Türei et al., 2021), and SIGNOR 4.0 (Surdo et al., 2026), together with perturbation–target priors from

View	Reduced edges	Gene coverage	Reduction setting
STRING v12.0	20,933	1,690/2,000	score ≥ 700
Reactome	15,479	1,287/2,000	pathway elimination

Table 4: Frozen undirected structural backbones used by Module II. The two reduced views share 3,694 edges and remain separately identified in the model.

Tahoe metadata (Zhang et al., 2026) and DrugCentral (Avram et al., 2021). The frozen heterogeneous graph contains 13,819 nodes and 150,952 directed edges across eight node types: genes, proteins, transcription factors, complexes, miRNAs, protein families, fusion proteins, and ncRNAs. The final artifact contains 42 biological relation types; source provenance is not retained as a separate edge attribute in this heterogeneous representation.

For a directed edge $e = (u \rightarrow v)$, the graph stores a biological relation type ρ_e, sign $s_e \in \{-1, 0, +1\}$, sign-known indicator k_e, positive static graph weight w_e, and evidence confidence q_e when available. Unknown signs are represented by $s_e = 0$ with $k_e = 0$ instead of being converted to activation. Among the frozen heterogeneous edges, 112,388 are positive, 20,256 negative, and 18,308 sign-unknown, giving an 87.9% sign-known fraction.

Multiple observations with the same source, target, and relation type are collapsed by confidence-weighted aggregation. For signed observations, activation and inhibition are encoded as $+1$ and -1, and the retained sign is determined by their confidence-weighted signed mean; values with absolute mean below 0.05 are stored as sign-unknown. The resulting positive base weights w_e are normalized over outgoing edges of each source node. The evidence factor q_e is retained separately when observed; missing evidence uses the learned positive substitute described in Appendix C.1.

During message passing, known-sign and unknown-sign edges use separate transformations. Each message is weighted by w_e, q_e, perturbation-specific relevance, and source-node activity, and incoming messages are normalized by their accumulated weight before updating the target node.

Chemical perturbations use Morgan/ECFP-style circular fingerprints (Rogers & Hahn, 2010) with radius 2 and 1024 bits, mapped to a 128-dimensional perturbation representation. Targetable graph nodes are genes, transcription factors, proteins, and protein families. Molecular-to-node compatibility uses cosine similarity with top-32 selection, temperature 0.25, logit scale 2.0, and a curated-target logit boost of 4.0. Curated target support is available for 271/379 perturbations (71.5%); when no curated target is available, injection is determined solely by molecular compatibility. For P-holdout, perturbation identity, molecular features, and available target annotations are treated as input-side priors; held-out treated-expression data are not used to construct these priors.

Of the 2,000 modeled genes, 1,408 have at least one incident edge in the heterogeneous graph. Starting from non-measured targetable entry nodes, 1,277/2,000 modeled genes are reachable within the two directed propagation layers used by the final model.

B.3 Undirected structural graph for adaptive geometry

The undirected structural graph is constructed from two positive prior views, STRING v12.0 (Szklarczyk et al., 2023) and Reactome (Milacic et al., 2024), and restricted to edges connecting genes in the modeled expression panel. Its topology is fixed: the model only reweights retained edges and does not introduce new dense connections.

Each retained edge has a fixed positive prior weight $\bar{w}_e$, obtained from its source-specific structural evidence. During transport, this prior is modulated by context- and state-dependent factors as in Equation 7. In the P- and C-holdout experiments, the structural metric depends only on cellular context and the evolving state, with no additional perturbation-specific edge modulation.

B.4 Identifier harmonization

Measured-gene matching uses upper-case HGNC gene symbols (Tweedie et al., 2021) as the working namespace. Direct gene-level matches unresolved to the 2,000-gene panel are omitted; STRING proteins outside the panel remain as hidden nodes until sparse reduction. In the heterogeneous graph,

additional supported biological entities can remain as typed nodes; complexes and protein families are kept distinct even when their display labels overlap with simpler entities. Protein identifiers and sequence records are resolved through UniProt accessions where available (UniProt Consortium, 2023).

B.5 Sparse reduction to measured genes

The raw structural resources contain nodes that are absent from the G-gene expression vector. Conceptually, if a graph Laplacian is partitioned into measured (M) and hidden (H) nodes,

$$L = \begin{bmatrix} L_{MM} & L_{MH} \\ L_{HM} & L_{HH} \end{bmatrix}, \tag{18}$$

then, when L_{HH} is invertible, exact elimination of hidden nodes gives the Schur-complement Laplacian

$$L_{\text{eff}} = L_{MM} - L_{MH} L_{HH}^{-1} L_{HM}. \tag{19}$$

The implementation avoids forming the dense inverse in Equation 19 and uses source-specific sparse reductions.

STRING reduction. STRING v12.0 interactions are filtered at a combined score ≥ 700. Measured-gene-associated proteins are terminals, and proteins outside the expression panel are treated as hidden nodes. Let D_H contain the full degrees of hidden nodes and W_{HH} their mutual adjacency, with $A_H = D_H^{-1/2} W_{HH} D_H^{-1/2}$. The hidden-node fill-in uses the truncated Jacobi/Neumann expansion

$$L_{HH}^{-1} \approx D_H^{-1/2} \left(I + A_H + A_H^2\right) D_H^{-1/2}. \tag{20}$$

The frozen construction retains terms corresponding to at most three hidden proteins, prunes each hidden frontier to the strongest 64 measured-terminal entries, and finally keeps the union of each measured gene's strongest 16 incident edges.

Reactome reduction. Reactome pathways are represented as pathway-centered stars over their retained measured member genes. For a pathway containing d measured genes, exact elimination of its unit-conductance pathway node contributes $1/d$ conductance to each measured-gene pair. Contributions from overlapping pathways are accumulated, followed by the same union top-16 sparsification. Thus, STRING uses a sparse truncated-Neumann approximation to hidden-node Schur reduction, while Reactome pathway-node elimination is exact before terminal sparsification.

B.6 Positive support and view normalization

Module II uses only positive undirected structural support. For each retained edge e within a structural view, let $w_e > 0$ denote its static edge weight. Activation or inhibition signs from the directed mechanism graph are not transferred into this module.

To place the STRING and Reactome views on a common numerical scale, the edge weights of each view are rescaled independently as

$$w_e \leftarrow w_e \frac{G}{2 \sum_{e'} w_{e'}}, \tag{21}$$

where the sum is taken over all retained edges in that view. For an undirected weighted graph,

$$\operatorname{tr}(L) = 2 \sum_e w_e,$$

so this normalization gives $\operatorname{tr}(L) = G$ for each structural view and prevents differences in raw edge-weight scale from determining their relative contribution.

The normalized STRING and Reactome views are retained separately and combined using fixed mixture coefficients of $1/2$. The resulting static support weight is denoted by $\bar{w}_e$ and is used in Equation 7.

Context modulation and state-dependent susceptibility only reweight edges already present in the frozen structural support; they do not introduce new edges. Because the normalized static weights, mixture coefficients, and adaptive factors are positive, the effective weights $w_{c,e}$ in Equation 7 remain positive. Consequently, the corresponding weighted graph Laplacian is symmetric positive semidefinite, and the regularized metric in Equation 9 is symmetric positive definite.

C Model Details

This appendix records implementation and auxiliary details omitted from the main derivation. The principal transport quantities d_p, $m_{p,c}$, $\mathsf{G}_c(x)$, $\gamma_\psi(t)$, and v_θ are defined in Section 3 and are not redefined here.

C.1 Implementation Details of Mechanism Grounding

Node identity. Every graph node has a learned node embedding e_v^{node} and a learned embedding $e_{\mathrm{type}(v)}^{\mathrm{type}}$ for its node type. In both cases, the embedding is randomly initialized (normal standard deviation 0.02) and optimized with the model; it is not derived from the molecular fingerprint, which encodes the perturbation instead. In the formal configurations, its identity representation is

$$h_v^{\mathrm{id}} = e_v^{\mathrm{node}} + e_{\mathrm{type}(v)}^{\mathrm{type}} + \begin{cases} W_{\mathrm{prot}} f_v, & \text{if a sequence-derived feature is available,} \\ e_v^{\mathrm{fallback}}, & \text{otherwise.} \end{cases} \tag{22}$$

The fixed feature f_v comes from a pretrained ESM-2 encoder of UniProt protein sequences. It can represent a gene, transcription factor, or protein node; family and complex features average their protein-member features. W_{prot} and the fallback embeddings are learned, while f_v remains fixed.

Sparse propagation implementation. Let $s_{p,v}^{\mathrm{tar}}$ be the curated target sign, set to zero when unknown. The injected state and active mask are

$$h_{p,v}^{(0)} = \omega_{p,v}(W_{\mathrm{inj}} z_p + h_v^{\mathrm{id}} + s_{p,v}^{\mathrm{tar}} W_{\mathrm{sign}} z_p), \qquad A_{p,v}^{(0)} = \mathbf{1}[\omega_{p,v} > 0]. \tag{23}$$

Here, z_p is the perturbation embedding, h_v^{id} is the identity representation of node v, and $\omega_{p,v}$ is the perturbation-specific target weight that controls the strength of the initial injection. W_{inj} projects the perturbation embedding into the node-feature space, while W_{sign} encodes the direction of a curated target interaction through $s_{p,v}^{\mathrm{tar}} \in \{-1, 0, +1\}$. The binary mask $A_{p,v}^{(0)}$ marks nodes that receive a nonzero initial perturbation signal and therefore participate in subsequent sparse propagation. For an edge $e = (u \to v)$ with relation ρ_e, the message in Equation 5 is

$$\widetilde{h}_e^{(\ell)} = r_{\rho_e} \odot \begin{cases} s_e W_{\mathrm{known}} h_{p,u}^{(\ell)}, & \text{known sign,} \\ W_{\mathrm{unknown}} h_{p,u}^{(\ell)}, & \text{unknown sign,} \end{cases} \qquad a_{p,e}^{(\ell)} = w_e q_e R_{p,e}^{(\ell)} A_{p,u}^{(\ell)}. \tag{24}$$

Here $w_e > 0$ is the graph weight, $q_e > 0$ is observed evidence or a learned positive substitute, and $R_{p,e}^{(\ell)} \in (0, 1)$ is a drug-conditioned source–target relevance gate. r_{ρ_e} is a learned feature-wise relation scale; $s_e \in \{-1, +1\}$ is used only for known signs. The update U in Equation 5 is a two-layer MLP. Nodes become active when they receive an active incoming edge; inactive states stay zero. The measured-gene representation uses the gated active state after layer two, and the drive head is zeroed on unreachable genes. Sparse graph-index operations avoid a dense heterogeneous adjacency.

Target alignment. For perturbations with curated target annotations, let $\mathcal{T}_p$ denote the annotated targetable nodes. The auxiliary alignment loss acts on the molecular compatibility logits before the target-prior boost:

$$\mathcal{L}_{\mathrm{align}} = -\frac{1}{|\mathcal{T}_p|} \sum_{v \in \mathcal{T}_p} \log \frac{\exp a_{p,v}^{\mathrm{mol}}}{\sum_{u \in \mathcal{V}_{\mathrm{targetable}}} \exp a_{p,u}^{\mathrm{mol}}}. \tag{25}$$

It is used only in the perturbation-transfer configurations described in Appendix E.

C.2 Implementation Details of Graph Metric Adaptation

Untreated-context aggregation. For gene i in context c, local untreated statistics first give

$$r_{c,i} = E_{\text{basal}}([\mu_{c,i}, \sigma_{c,i}]) + e_i. \tag{26}$$

E_{basal} maps the basal mean and standard deviation to the node feature space, and e_i is the learned gene identity embedding. We then summarize neighboring basal features. Let $\mathcal{E}(i)$ denote the retained structural edges incident to gene i, and let $n_i(e)$ denote the opposite endpoint of edge e. The aggregated feature is

$$\bar{r}_{c,i} = \frac{\sum_{e\in\mathcal{E}(i)} \bar{w}_e^{\text{enc}} \left(r_{c,n_i(e)} - e_{n_i(e)}\right)}{\max\left(\epsilon, \sum_{e\in\mathcal{E}(i)} \bar{w}_e^{\text{enc}}\right)}. \tag{27}$$

Here $\epsilon = 10^{-8}$ is used for numerical stability, and $\bar{w}_e^{\text{enc}} > 0$ assigns each retained STRING or Reactome edge one half of its stored view-specific weight. Subtracting the identity embedding leaves the basal feature for aggregation. Genes without structural neighbors bypass this neighborhood correction and retain $r_{c,i}$. The neighborhood correction is

$$\Delta_{c,i} = \bar{r}_{c,i} - (r_{c,i} - e_i). \tag{28}$$

It measures the difference between neighboring and local basal features. Its contribution is controlled by

$$\tau_{c,i} = \text{sigmoid}\, f_{\text{trust}}([r_{c,i}, \Delta_{c,i}]). \tag{29}$$

The learned gate $\tau_{c,i} \in (0, 1)$ uses both the local feature and the neighborhood difference. The resulting context feature is

$$h_{c,i}^{\text{ctx}} = r_{c,i} + \tau_{c,i}\Delta_{c,i}. \tag{30}$$

Genes without structural neighbors retain $r_{c,i}$.

For context-dependent edge scoring, we remove the identity embedding, $\widetilde{h}_{c,i} = h_{c,i}^{\text{ctx}} - e_i$, so the score reflects basal context rather than a fixed gene prior. For edge $e = (i, j)$, the symmetric features $[\widetilde{h}_{c,i} + \widetilde{h}_{c,j}; |\widetilde{h}_{c,i} - \widetilde{h}_{c,j}|]$ produce

$$b_{c,e} = \eta_c \tanh f_\beta\left([\widetilde{h}_{c,i} + \widetilde{h}_{c,j}; |\widetilde{h}_{c,i} - \widetilde{h}_{c,j}|]\right). \tag{31}$$

We remove its static-weighted edge mean,

$$\widehat{b}_{c,e} = b_{c,e} - \frac{\sum_{e'} \bar{w}_{e'} b_{c,e'}}{\sum_{e'} \bar{w}_{e'}}, \qquad \beta_{c,e} = \exp(\widehat{b}_{c,e}), \tag{32}$$

which gives the context-dependent factor used in Equation 7.

Stage-B auxiliary objectives. In addition to the pointwise Flow-Matching objective, the reported configurations supervise the population-mean velocity error. For a coupled minibatch with N cells, let

$$\bar{e}_v = \frac{1}{N}\sum_{n=1}^{N}\left(v_\theta^{(n)} - u_t^{(n)}\right), \tag{33}$$

where $u_t = \dot{\gamma}_\psi(t)$ is the Stage-A path tangent. The corresponding mean-velocity objective is

$$\mathcal{L}_{\text{mean}} = \mathbb{E}_{(p,c),t}\left[\frac{1}{G}\|\bar{e}_v\|_2^2\right]. \tag{34}$$

P-holdout additionally uses a weak condition-level effect-correlation objective. Define

$$\widehat{\Delta}_v = \frac{1}{N}\sum_{n=1}^{N} v_\theta^{(n)}, \qquad \Delta_{\text{pair}} = \frac{1}{N}\sum_{n=1}^{N}\left(x_1^{(n)} - x_0^{(n)}\right). \tag{35}$$

The auxiliary correlation loss is

$$\mathcal{L}_{\text{effect}} = \mathbb{E}_{(p,c),t}\left[1 - \text{corr}_g\left(\widehat{\Delta}_v, \Delta_{\text{pair}}\right)\right], \tag{36}$$

where corr_g denotes Pearson correlation across the G measured genes. These objectives supervise instantaneous velocities at the sampled training time; they do not use the final ODE-integrated population as a training target.

Algorithm 1 Two-stage training of MECHFLOW

Require: coupled minibatches $(x_0, x_1) \sim \pi_{p,c}$
encode perturbation and context to obtain $(H_p^{\mathrm{mech}}, d_p, z_p)$ and $(H_c^{\mathrm{ctx}}, z_c)$
construct the mechanism-grounded drift $m_{p,c}$
Stage A: learned training path
repeat
    construct the context- and state-dependent metric $\mathsf{G}_c(\cdot)$
    sample $t \sim \mathcal{U}[0,1]$
    compute $(x_t, u_t) = (\gamma_\psi(t), \dot{\gamma}_\psi(t))$
    update Stage-A parameters using $\mathcal{L}_{\mathrm{geo}}$ and the configured regularizers
until early stopping
freeze the learned path and geometric components
Stage B: Flow Matching
repeat
    encode the current perturbation and context
    sample $t \sim \mathcal{U}[0,1]$
    compute (x_t, u_t) with the frozen path construction and detach the resulting training pair
    compute the residual field $r_\theta(x_t, t; p, c)$
    compute $v_\theta(x_t, t; p, c) = \mathcal{M}_c(x_t) r_\theta(x_t, t; p, c) + m_{p,c}$
    update Stage-B parameters using $\mathcal{L}_{\mathrm{FM}}$ and the configured auxiliary objectives
until early stopping

C.3 Implementation Details of Transport Dynamics

Condition residual. For P-holdout, the condition residual in Equation 11 is

$$\delta_{p,c}^{\mathrm{cond}} = \rho_{\mathrm{res}} f_{\mathrm{res}}([z_p, z_c, \mu_c]), \tag{37}$$

where $\rho_{\mathrm{res}} = 0.5$ in the frozen configuration. The output layer of f_{res} is initialized to zero, so optimization begins from the target-grounded graph contribution. The C-holdout configuration has no independent condition-residual head.

Path correction network. The path correction ϕ_ψ in Equation 12 receives the current linear interpolation, endpoint displacement, basal mean and standard deviation, gene-wise mechanism and context features, transport time, and global perturbation/context representations. Its output layer is initialized to zero, so Stage A begins from the linear endpoint path. The time derivative in Equation 13 is evaluated with a forward-mode JVP with respect to the scalar transport time.

Residual vector field. For the residual decomposition in Equation 15, the state is centered by the untreated basal mean before being combined with transport time to obtain h_t. The signed drift $m_{p,c}$ is not provided to this residual branch and is added only after the optional metric transformation in Equation 16. For the inverse-metric option, the corresponding linear system is solved without explicitly materializing $\mathsf{G}_c(x)^{-1}$.

D Algorithms

Algorithms 1 and 2 summarize the two-stage training and inference procedures. Mechanism and context encoding follow Appendix C.

The Stage-A path network is used only to construct training paths and tangents; it is not evaluated during inference. The reported solver settings and stopping criteria are given in Appendix E.

E Engineering Details and Training Stabilization

E.1 Stage-A stabilization objective

Section 3 defines the primary graph-geometric path objective $\mathcal{L}_{\mathrm{geo}}$. The implemented Stage-A loss adds weak regularization of the path correction, its temporal derivative, state susceptibility, and—for

Algorithm 2 Inference with MECHFLOW

Require: untreated cells X_0^c, perturbation p, number of ODE steps N_{ode}
1: encode the perturbation and the untreated context
2: construct the mechanism-grounded drift $m_{p,c}$
3: $x^{(0)} \leftarrow X_0^c$
4: **for** $k = 0, \dots, N_{\text{ode}} - 1$ **do**
5: $\quad t_k \leftarrow k/N_{\text{ode}}$
6: $\quad$ compute $v_k \leftarrow \mathcal{M}_c(x^{(k)}) r_\theta(x^{(k)}, t_k; p, c) + m_{p,c}$
7: $\quad x^{(k+1)} \leftarrow x^{(k)} + N_{\text{ode}}^{-1} v_k$
8: **end for**
9: **return** $x^{(N_{\text{ode}})}$

P-holdout—molecular target alignment:

$$\begin{aligned}\mathcal{L}_A = \mathcal{L}_{\text{geo}} + \lambda_{\text{corr}} \mathbb{E}[\|\phi_\psi(t)\|_2^2] + \lambda_{\text{smooth}} \mathbb{E}[\|\partial_t \phi_\psi(t)\|_2^2] \\ + \lambda_{\text{state}} \mathbb{E}[\text{Std}(\ell(x_t, c))^2] + \lambda_{\text{align}} \mathcal{L}_{\text{align}}.\end{aligned} \tag{38}$$

The frozen P/C configurations use $\lambda_{\text{corr}} = 10^{-3}$, $\lambda_{\text{smooth}} = 10^{-4}$, and $\lambda_{\text{state}} = 10^{-4}$; P-holdout uses $\lambda_{\text{align}} = 10^{-2}$ and C-holdout sets it to zero. Stage A evaluates both $t \sim \mathcal{U}(0, 1)$ and its antithetic counterpart $1 - t$ and averages the two losses.

E.2 ENDPOINT COUPLING

Untreated and perturbed cells are population samples and are not assumed to be biologically paired. During training, both stages construct minibatch couplings using log-domain Sinkhorn optimal transport with regularization $\epsilon = 0.1$ and 30 iterations. The coupling is recomputed per minibatch and is used only to construct training pairs; validation and inference do not use Sinkhorn pairing.

E.3 MATRIX-FREE ADAPTIVE GRAPH GEOMETRY

The implementation never materializes the dense metric in Equation 9. Its quadratic form is evaluated through sparse edge differences, with structural edges processed in chunks of 4096. When the inverse-metric mobility in Equation 16 is used, the model solves

$$\mathsf{G}_c(x)\widetilde{r} = r_\theta \tag{39}$$

using eight fixed Jacobi-preconditioned conjugate-gradient iterations. No residual-tolerance early stopping is used; the relative residual is recorded only as a diagnostic. P-holdout disables inverse-metric velocity geometry during generation, whereas C-holdout recomputes the state-dependent geometry and applies the solve at every ODE step.

E.4 OPTIMIZATION AND GRADIENT CONTROL

Path construction, JVP tangents, state-responsive geometry, graph-energy evaluation, and PCG are executed in FP32. Both stages use AdamW with weight decay 10^{-4} and gradient-norm clipping at 1.0. Stage A uses learning rates 2×10^{-4} for the main parameters, 10^{-3} for the path-correction network, and 2×10^{-5} for state susceptibility; Stage B uses 2×10^{-4}. During Stage A, the global graph-strength parameter and structural-view mixture logits are frozen, while context-dependent edge modulation and state susceptibility remain trainable. During Stage B, the path-network, graph-geometry, and state-geometry parameters are frozen; the perturbation mechanism encoder and untreated-context encoder remain trainable. The path and tangent are recomputed from their current encodings under no-gradient mode for each batch, so detached Flow Matching targets can change as the encoders are updated. Context-dependent geometry weights are likewise recomputed from current encodings, although the geometry module's parameters are frozen. Pretrained protein representations are fixed buffers, while their projection layers remain trainable.

Setting	P-holdout	C-holdout
Initial mechanism-drift scale	0.25	0.50
Inference velocity geometry	No	Yes
Basal-aware condition residual	Yes	No
Condition-residual scale	0.5	0
Target-alignment weight	0.01	0
Effect-correlation weight	0.05	0
Mean-velocity weight	2.0	2.0

Table 5: Task-specific settings of the frozen full-scale MECHFLOWconfigurations. The mechanism-drift values are initialization values; the scalar remains trainable during Stage B.

E.5 STAGE-B AUXILIARY SUPERVISION

The implemented Stage-B objective combines the primary Flow-Matching loss with the auxiliary objectives defined in Appendix C:

$$\mathcal{L}_B = \mathcal{L}_{\mathrm{FM}} + \lambda_{\mathrm{mean}}\mathcal{L}_{\mathrm{mean}} + \lambda_{\mathrm{effect}}\mathcal{L}_{\mathrm{effect}} + \lambda_{\mathrm{align}}\mathcal{L}_{\mathrm{align}}. \tag{40}$$

Here $\mathcal{L}_{\mathrm{mean}}$, $\mathcal{L}_{\mathrm{effect}}$, and $\mathcal{L}_{\mathrm{align}}$ are defined in Equations 34, 36, and 25, respectively. Both configurations set $\lambda_{\mathrm{mean}} = 2.0$. P-holdout sets $(\lambda_{\mathrm{effect}}, \lambda_{\mathrm{align}}) = (0.05, 0.01)$; C-holdout sets both to zero.

E.6 FROZEN TASK-SPECIFIC CONFIGURATIONS

Both formal P- and C-holdout configurations use a 2,000-gene panel, batch size 8 conditions, $d_g = 48$, $d_z = d_c = 128$, two directed mechanism layers, top-32 molecular-to-node compatibility, temperature 0.25, compatibility logit scale 2.0, target-prior boost 4.0, and context-only structural mobility. Training cycles through 64-cell blocks within each condition, whereas validation/checkpoint selection uses a deterministic fixed set of 64 cells per condition.

The P-holdout condition-residual head receives $[z_p, z_c, \mu_c]$ and is initialized with a zero output layer; C-holdout does not instantiate this head. The mechanism-drift scale remains trainable during Stage B in both configurations. Route-level context modulation is disabled in both formal configurations.

E.7 VALIDATION AND GENERATION

Stage A and Stage B have maximum budgets of 51 and 201 epochs, with early-stopping patience of 10 epochs. Stage A is selected by the condition-macro path objective subject to the geometry-feasibility checks used by the training code; Stage B is selected by condition-macro mean MSE on the designated validation split. Generation starts from untreated control cells and uses 20 explicit-Euler steps with step size $1/20$; the Stage-A path network is not evaluated during inference.

E.8 COMPLEXITY

We summarize the graph-related computational cost. Let E_s and E_d denote the number of structural and directed mechanism edges, respectively, and let L be the number of directed propagation layers. Stage-A graph-energy evaluation requires $\mathcal{O}(E_s)$ sparse edge operations per sampled cell and transport time. Directed mechanism propagation is evaluated at the condition level with $\mathcal{O}(LE_d)$ edge-wise operations; in the reported configurations, $L = 2$.

During generation, P-holdout does not apply the state-dependent inverse-metric mobility and therefore avoids the corresponding sparse linear solve. For C-holdout, each velocity evaluation recomputes the state-dependent edge factors and solves the metric system using $J = 8$ fixed PCG iterations. The dominant graph-related cost is therefore $\mathcal{O}(JE_s)$ per cell and ODE step.

All graph operations are implemented sparsely. No $G \times G$ dense metric or its inverse is materialized, giving graph-storage complexity $\mathcal{O}(E_s + G)$ instead of $\mathcal{O}(G^2)$.

F EXPERIMENTAL DETAILS

Tahoe subset and expression scale. We use the Tahoe-100M sample metadata to select one concentration per drug: among its distinct numeric concentrations, we retain the middle-ranked value only when their number is odd, and their units agree. Treated cells at that dose and untreated cells on the corresponding plates are retained. The ordered $2{,}000$-gene panel follows the PerturbDiff panel in the released graph prior; gene symbols are mapped to Tahoe token IDs, and all other genes are discarded. Counts are stored before normalization. At loading time, each cell's retained gene counts are divided by their sum, multiplied by $10{,}000$, and transformed with $\log(1 + x)$; cells with zero retained counts remain zero.

Quality control and condition pools. Samples pass the metadata-level filters of mean detected genes ≥ 500, mean transcripts ≥ 500, and mean mitochondrial fraction ≤ 0.20; samples with missing quality-control values are excluded. No additional per-cell quality filter or minimum treated-cell threshold is applied. A treated condition is defined by a drug–plate–cell-line combination, and its untreated reference is matched by plate and cell line. All passing cells are retained in their corresponding condition pools.

Holdouts and leakage prevention. Unique drugs and cell lines are sorted and shuffled independently with seeds 42 and 43, respectively, and divided into 70% train, 15% validation, and 15% test entity partitions, with counts rounded to whole entities. P-holdout partitions conditions by drug, whereas C-holdout partitions them by cell line; the complementary entity is not held out by that task. Consequently, no held-out drug is used to train the corresponding P-holdout model, and no held-out cell line is used to train the corresponding C-holdout model.

Each task uses its own validation partition for checkpoint selection and a frozen test-condition list containing $2{,}998$ P-holdout or $3{,}197$ C-holdout conditions. Test-treated cells are excluded from model fitting, checkpoint selection, PCA fitting, and bandwidth selection. Untreated cells matched to test conditions are used as inference inputs and control references; they are never used as test-treated targets.

Cell sampling and checkpoint selection. Training visits each condition's treated cells in successive 64-cell blocks, wrapping only the final incomplete block. Validation uses a deterministic, fixed set of 64 treated and 64 matched untreated cells per condition, sampling with replacement when a pool contains fewer than 64 cells.

Stage A is selected by the condition-macro path objective subject to the feasibility checks in Appendix E, whereas Stage B is selected by validation condition-macro mean MSE. Both stages use an early-stopping patience of 10 epochs. The selected Stage-B checkpoint is evaluated once on the frozen test partition. At test time, the evaluator visits every treated cell in each condition in 64-cell chunks; the final incomplete chunk is padded for generation, but padded cells are excluded from metric accumulation. Matched untreated cells provide the model inputs.

Distribution reference and metric aggregation. For each task, PCA is fitted using fixed 64-cell samples from 320 randomly selected training conditions, with condition-selection seed 4201. A seeded rank-58 low-rank decomposition with seed 4200 supplies the first 50 principal components.

To determine the RBF bandwidth, fixed 64-cell samples are drawn from 128 validation conditions using condition-selection seed 4202. We then sample $100{,}000$ index pairs in the resulting PCA50 space using seed 4203, discard identical-index pairs, and set σ_0 to the median Euclidean distance of the remaining pairs.

The biased empirical MMD^2 estimator averages three Gaussian kernels with bandwidths $\{\sigma_0/2, \sigma_0, 2\sigma_0\}$. For each condition, predicted and observed populations are subsampled to at most 128 cells under each of five fixed evaluation-sampling seeds $0, \ldots, 4$, and the resulting MMD^2 values are averaged. These seeds are used only for the frozen distributional evaluation procedure.

All reported benchmark metrics are macro-averaged over conditions. Every compared method uses the same task-specific test-condition list and the same frozen evaluation reference; no method-specific test-condition subset contributes to the main results.

G Full Ablation Study

P-holdout interventions. The *w/o graph mechanism* variant zeros the explicit signed mechanism drift but keeps the other graph-conditioned components. *Corrupted prior* reassigns drug targets within node-type/sign groups while preserving target counts and graph structure, testing whether the drug–target correspondence matters. *w/o modulation* fixes the context-dependent mechanism gate to one, retaining target injection and signed propagation.

C-holdout interventions. *w/o signed mechanism* zeros the signed drift scale while retaining other graph-conditioned paths. *w/o context geometry* replaces the graph metric with the identity, removing its PCG contribution as well as context edge weighting. *w/o path geometry* and *w/o velocity geometry* disable the corresponding geometry branch separately. *Randomized topology* permutes gene labels while preserving graph size and degree structure.

Method	MSE↓	Mean-Pear.↑	Δ-Pear.↑	DE-Spear.↑	DE50 Dir.↑	DE50 r↑	MMD^2↓
MechFlow	0.03071	0.94510	0.25810	0.30976	0.69363	0.36541	0.01844
w/o graph mechanism	0.03102	0.94464	0.22635	0.27308	0.67266	0.32389	0.01892
Corrupted prior	0.03099	0.94456	0.22874	0.27764	0.67426	0.32831	0.01890
w/o modulation	0.03092	0.94468	0.23771	0.29239	0.67852	0.34485	0.01894

Table 6: Exploratory straight-interpolation P-holdout ablations on 512 held-out conditions.

Method	MSE↓	Mean-Pear.↑	Δ-Pear.↑	DE-Spear.↑	DE50 Dir.↑	DE50 r↑	MMD^2↓
MechFlow	0.03115	0.95198	0.22266	0.27635	0.70383	0.33634	0.01670
w/o signed mechanism	0.03157	0.95033	0.18568	0.23707	0.68313	0.29448	0.01751
w/o context geometry	0.03144	0.94990	0.19776	0.24339	0.70383	0.30138	0.01708
w/o path geometry	0.03121	0.95012	0.20508	0.24730	0.70688	0.30896	0.01688
w/o velocity geometry	0.03142	0.95083	0.19857	0.24404	0.69184	0.31393	0.01738
Randomized topology	0.03139	0.95020	0.19249	0.24176	0.69481	0.30070	0.01711

Table 7: C-holdout proxy ablations on 512 held-out conditions.

H Additional Interpretability Studies

The following analyses complement the main-text mechanism and transport visualizations with additional readout definitions, response-space comparisons, and cross-context views of the frozen model. As in Section 4.3, normalized flow time denotes ODE integration time rather than experimentally measured biological time. All analyses in this section are descriptive views of the learned predictor and are not independent performance tests.

H.1 Main-Text Figure Readouts

Directed mechanism propagation. Figure 2 visualizes perturbation-conditioned routing through the directed mechanism graph. For each displayed edge, we summarize its layer-specific relevance over the sparse propagation layers and retain edges activated by the corresponding perturbation. The resulting subgraphs highlight perturbation-specific routes through which the model propagates mechanism information from high-confidence entry nodes toward measured genes.

Edge widths encode relative routing relevance within each perturbation panel, making it possible to compare which interactions are emphasized by the model for a given perturbation. Because the widths are normalized separately within each panel, their absolute magnitudes are not intended for quantitative comparison across perturbations.

Flow-time gene and pathway dynamics. For the Reactome summary in Figure 3, let

$$\bar{v}_i(t_k) = \frac{1}{N}\sum_{n=1}^{N}\left[v_\theta(x_{t_k}^{(n)}, t_k; p, c)\right]_i \tag{41}$$

denote the population-mean velocity of $N = 64$ cells at $t_k = k/20$, $k = 0, \ldots, 20$. For each Reactome set S containing 10–150 measured genes, we compute

$$r_S(t_k) = \left(\frac{1}{|S|} \sum_{i \in S} \bar{v}_i(t_k)^2 \right)^{1/2}, \qquad R_S(W) = \frac{1}{|W|} \sum_{k \in W} r_S(t_k). \tag{42}$$

Pathways are ranked separately by $R_S(W)$ for the early ($k = 0$–6), middle ($k = 7$–13), and late ($k = 14$–20) windows, with the top three reported for each window. The RMS is taken across genes before averaging over flow-time points, and Reactome sets may overlap in gene membership.

Mechanism–residual decomposition. Figure 4 summarizes three population-level quantities for the C-holdout LB-100/CVCL_1635 case. For a generated trajectory $x_{t_k}^{(n)}$, the predicted expression displacement of gene i is

$$\Delta \bar{x}_i(t_k) = \frac{1}{N} \sum_{n=1}^{N} \left[x_{t_k}^{(n)} - x_0^{(n)} \right]_i. \tag{43}$$

The corresponding population-mean effective residual transport is

$$\bar{r}_i^{\text{eff}}(t_k) = \frac{1}{N} \sum_{n=1}^{N} \left[\mathcal{M}_c(x_{t_k}^{(n)}) r_\theta(x_{t_k}^{(n)}, t_k; p, c) \right]_i, \tag{44}$$

and the population-mean total velocity is

$$\bar{v}_i(t_k) = \frac{1}{N} \sum_{n=1}^{N} \left[v_\theta(x_{t_k}^{(n)}, t_k; p, c) \right]_i. \tag{45}$$

By Equation 16, the difference between total velocity and effective residual transport reflects the mechanism-grounded drift $m_{p,c}$. The heatmaps therefore expose whether the two components reinforce or oppose one another over normalized flow time.

H.2 Predicted Response Geometry

To compare the response distributions produced by **MechFlow** and the matched Vanilla FM backbone, we use the same 192 untreated cells from the P-holdout Trimetrexate/CVCL_0397 case. Each predicted endpoint is first projected using the training-derived PCA50 reference, and the per-cell response is represented by its endpoint-minus-control displacement in this space. A single two-dimensional UMAP is then fitted jointly to the predicted responses from both models, so the two panels share the same display coordinates.

The two models produce visibly different response clouds in the shared embedding. Because UMAP is nonlinear and no treated reference population is shown, distances, cluster shape, and apparent separation in Figure 6 are used only as qualitative summaries of the generated response distributions. Quantitative predictive performance is assessed by the unified metrics in Section 4.

H.3 Context-Specific Transport

To inspect context dependence while holding the perturbation fixed, we propagate LB-100 from untreated cells in three held-out cellular contexts. Each panel contains 128 control-cell trajectories generated by the frozen C-holdout model using the same 20-step Euler solver. All trajectories are displayed in a common training-derived PCA coordinate system with shared axis limits; no context-specific projection, recentering, or rescaling is applied.

Figure 7 shows that the same perturbation is realized through different population-level displacement patterns across held-out cellular contexts. These trajectories reflect the combined effect of context-conditioned residual dynamics and adaptive transport geometry. Because the visualization is two-dimensional and contains no treated reference population, it is intended to expose context dependence of the learned flow rather than establish predictive accuracy on its own.

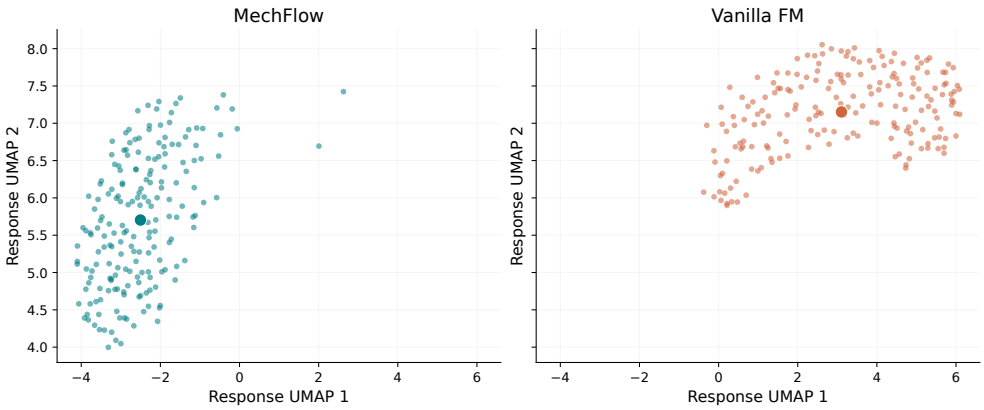

Figure 6: **Predicted response geometry for MECHFLOW and Vanilla FM.** Endpoint-minus-control responses for the same P-holdout Trimetrexate/CVCL_0397 control cells are shown in a shared two-dimensional UMAP fitted jointly to the two models' predicted responses after projection with the training-derived PCA50 basis. The larger marker denotes the mean location in the two-dimensional embedding. Observed treated cells are neither used to fit the UMAP nor displayed; the separation therefore reflects differences between the two predicted response geometries and is not a standalone accuracy comparison.

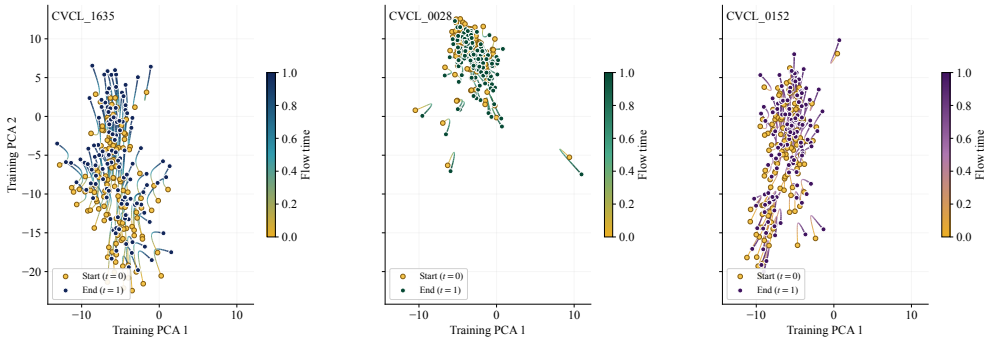

Figure 7: **Context-specific transport for a shared perturbation.** Frozen MECHFLOW trajectories for LB-100 in CVCL_1635, CVCL_0028, and CVCL_0152 are shown in the same training-derived PCA coordinates. Yellow points mark untreated states at $t = 0$, darker points mark generated endpoints at $t = 1$, and each connecting trajectory follows one initial cell through normalized flow time. The figure visualizes context-dependent predicted transport and does not represent longitudinal single-cell measurements.