BEYOND THE TRANSCRIPTOME: CHROMATIN-INFORMED PREDICTION OF CELL-STATE-DEPENDENT PERTURBATION RESPONSES

Anonymous authors
Paper under double-blind review

ABSTRACT

Predicting transcriptional responses to genetic perturbations is central to understanding gene function. Existing predictors primarily rely on transcriptomic measurements, although chromatin accessibility provides complementary information about the cellular context in which perturbations act. Using this information requires linking chromatin context to specific perturbations and accounting for baseline differences between independently sampled control and perturbed populations. We propose **ChromaPert**, a chromatin-informed framework for predicting state-dependent perturbation responses. ChromaPert combines molecular and DNA-informed locus priors with paired control RNA–ATAC features to jointly represent target identity and measured cellular context. Its Chromatin-Guided Response Router (CGRR) builds a source bank of control-derived expression offsets and retrieves relevant sources using perturbation/locus or control-ATAC similarity. These offsets account for baseline differences, while conditional transport flow learns the remaining response. Across three unseen-target and held-out-state settings on K562 CAT-ATAC and Perturb-Multiome, ChromaPert achieves the highest all-gene response correlation among evaluated methods. When transferring known perturbations to held-out states, it improves correlation by 23.2% across all genes and 46.2% for the twenty most responsive genes over the strongest respective baselines. Correct RNA–ATAC pairing improves recovery of chromatin-associated response slopes over shuffled pairing, while cross-state predictions retain state-specific responses to the same perturbation.

1 INTRODUCTION

Predicting transcriptional responses to perturbations is central to functional genomics and the systematic characterization of gene function. Single-cell CRISPR and combinatorial screens provide large-scale measurements (Dixit et al., 2016; Norman et al., 2019), but their experimental coverage remains incomplete across perturbations and cellular contexts, motivating generalizable *in silico* prediction for unseen perturbations and new cellular backgrounds.

Existing perturbation predictors primarily represent cellular state using transcriptomic profiles from single-cell RNA sequencing (scRNA-seq), which measures RNA abundance across genes as a readout of gene expression. They learn expression shifts, exploit gene relationships, or model mappings from control to perturbed cells (Lotfollahi et al., 2019; 2023; Roohani et al., 2024; Bunne et al., 2023; Klein et al., 2025). However, prior studies suggest that **chromatin accessibility**—the physical accessibility of DNA packaged with proteins—provides regulatory information that is not fully captured by current gene expression. The single-cell assay for transposase-accessible chromatin using sequencing (scATAC-seq) measures this accessibility (Buenrostro et al., 2015). This information may contribute to perturbation-response prediction in ***two complementary ways***. ***First***, pre-existing accessibility at response-relevant regulatory regions can influence transcription-factor recruitment and downstream transcription following stimulation. For example, pre-existing chromatin accessibility influences where the glucocorticoid receptor (GR) binds after hormone stimulation, thereby shaping the resulting gene-expression response (John et al., 2011; Biddie et al., 2011). ***Second***, chromatin accessibility can distinguish cellular backgrounds that appear similar at the transcriptomic level. Paired RNA–ATAC analysis of mouse skin, for instance, identified basal-cell popula-

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.

tions with similar RNA profiles but distinct accessibility patterns, including at regions containing p53 and CTCF motifs (Hao et al., 2021). Together, these findings suggest that chromatin accessibility may complement transcriptomic measurements by identifying response-relevant regulatory features and resolving cellular backgrounds with similar RNA profiles but distinct regulatory contexts, motivating its incorporation into models of perturbation response.

Using chromatin for perturbation prediction raises two challenges. **i) How can the RNA–chromatin background be linked to the response to a specific perturbation?** Control RNA–ATAC measurements describe gene expression and chromatin accessibility in unperturbed cells, whereas the prediction target is the expression change associated with a specific perturbation. The model must connect the RNA–chromatin background with the perturbation and learn how they jointly shape the transcriptional response. **However, because single-cell profiling is typically destructive, this relationship must be learned from separately sampled control and perturbed populations** rather than longitudinal before–after measurements of the same cells (Bunne et al., 2023). This introduces the second challenge: **ii) How can pre-existing expression differences be distinguished from perturbation-associated changes in these unpaired populations?** Control and perturbed populations may have different expression baselines. When reference controls do not represent the appropriate baseline, particularly when state-specific perturbed expression is compared with controls pooled across states, the observed difference can contain both pre-existing baseline shifts and perturbation-associated changes. The model must therefore account for these control-derived baseline differences rather than learning them as perturbation responses.

To address these challenges, we introduce **ChromaPert**, a chromatin-informed framework for predicting perturbation expression (Figure 1). To link cellular background to the specific perturbation, its **chromatin-informed perturbation condition** combines molecular and DNA-informed locus priors with paired control RNA–ATAC features. Protein and RNA-token embeddings encode target identity, while promoter-sequence and regulatory annotations describe its locus context. These complementary priors relate unseen targets to training perturbations; control RNA–ATAC features specify the cellular background in which they act. To account for baseline differences, its **Chromatin-Guided Response Router (CGRR)** builds a source bank of control-derived expression offsets. It routes queries by perturbation/locus similarity for unseen-target prediction or control-ATAC similarity for cross-state transfer. This decomposition separates baseline variation from the response learned by conditional transport flow, which reconstructs post-perturbation expression relative to matched controls for each perturbation in its corresponding cellular background.

ChromaPert achieves the highest all-gene response correlation among evaluated methods in unseen-target prediction on K562 CAT-ATAC (Shevade et al., 2025) and Perturb-Multiome (Martin-Rufino et al., 2025), and in cross-state transfer on the latter. Relative held-out-state gains reach **23.2%** across all genes and **46.2%** for the twenty most responsive genes over the strongest respective baselines (Table 3). Ablations support the model components; RNA–ATAC pairing and cross-state analyses show preserved chromatin-associated and state-dependent response structure (Table 4; Figure 2).

Our main contributions are summarized as follows:

- **Chromatin-informed perturbation conditioning.** We integrate molecular and DNA-informed locus priors with paired control RNA–ATAC features, allowing the perturbation condition to adapt to the measured RNA–chromatin context.
- **Chromatin-guided response routing.** CGRR builds response sources from control-derived expression offsets and performs query-dependent routing, separating pre-existing baseline variation from the response subsequently modeled by conditional transport flow.
- **Predictive and response-structure validation.** ChromaPert improves perturbation-response prediction across unseen-perturbation and cross-state settings, while controlled ATAC and response-structure analyses support the predictive value of chromatin-informed conditioning.

2 Related Work

Single-cell perturbation prediction. Latent-shift, compositional, and transport models learn intervention effects from independent control and perturbed populations (Lotfollahi et al., 2019; 2023; Bunne et al., 2023; Yu et al., 2025; Adduri et al., 2026). Gene relations, regulatory networks, causal structure, and sparse mechanism shifts provide biological constraints (Roohani et al., 2024; Kami-

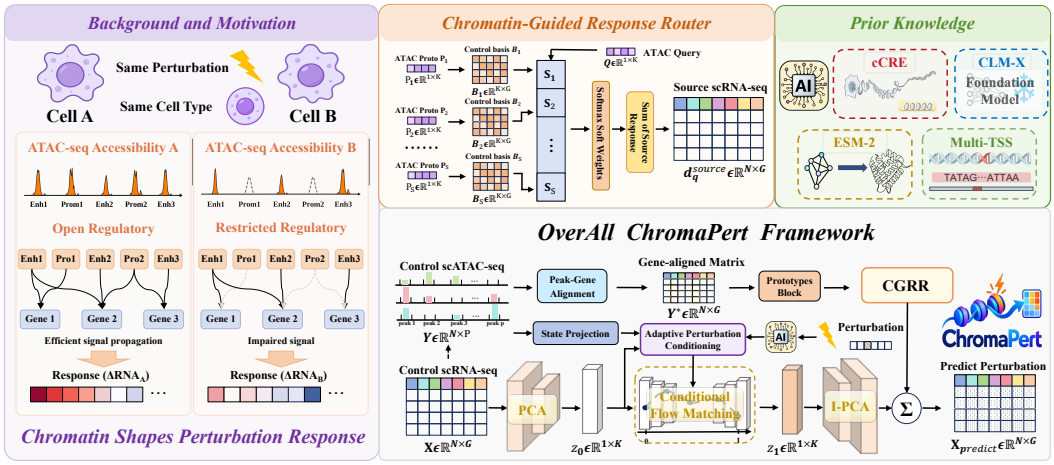

Figure 1: **ChromaPert overview.** §3 **Left:** chromatin context shapes the response to the same perturbation. **Top:** CGRR pairs control-RNA offsets B_s with ATAC prototypes P_s, using target/locus retrieval for Target-LOO and ATAC routing for Cross-State. **Bottom:** target/locus priors condition the flow, with paired control RNA–ATAC summaries added to the state-aware branch. **Right:** predicted responses are reconstructed relative to the matched-control RNA baseline.

moto et al., 2023; Wu et al., 2022; Bereket & Karaletsos, 2023; Chi et al., 2025b; Aibar et al., 2017). Foundation models supply pretrained gene or cell representations (Theodoris et al., 2023; Cui et al., 2024; Hao et al., 2024; Yang et al., 2024; Zeng et al., 2025; Wang et al., 2026). ChromaPert uses target representations and control chromatin for context-dependent source estimation and response prediction across perturbation targets and cellular backgrounds.

Generative modeling of unpaired responses. Generative models, including diffusion models, predict responses to unseen interventions (Qi et al., 2024; Chi et al., 2025a; Yuan et al., 2026). CellFlow uses flow matching, and scDFM adds distribution-level alignment (Klein et al., 2025; Yu et al., 2026). Conditional flow matching parameterizes transport between distributions (Lipman et al., 2022), but its learning target depends on the response definition and control reference. ChromaPert defines responses relative to a pooled training-control reference, subtracts a source component derived from control-state offsets, and fits a train-only conditional transport flow to the remaining response. Its source-on candidate combines the routed source component with the decoded transport-flow displacement before reliability gating and matched-control reconstruction.

Chromatin-aware and multimodal modeling. Chromatin accessibility complements RNA-derived cell state (Buenrostro et al., 2015; Ma et al., 2020). ArchR, weighted-neighbor analysis, and CLM-X integrate regulatory or multimodal information (Granja et al., 2021; Hao et al., 2021; Li et al., 2026); Perturb-Multiome and CAT-ATAC provide paired perturbation profiles (Martin-Rufino et al., 2025; Shevade et al., 2025). CellOracle uses multimodal regulatory networks to simulate transcription-factor perturbations (Kamimoto et al., 2023). ChromaPert uses target/locus or control-ATAC similarity, according to the evaluation setting, to route each query to expression offsets derived from training control states. These control-derived offsets, not known perturbation effects, form the source prior for conditional transport-flow prediction.

3 METHOD

Overview. As shown in Figure 1, ChromaPert predicts post-perturbation expression by connecting perturbation identity with measured RNA–chromatin context. It comprises three components: (i) **chromatin-informed perturbation conditioning** combines molecular and locus priors with control RNA–ATAC summaries; (ii) **chromatin-guided response routing** represents control-derived baseline variation and selects relevant training sources; and (iii) **conditional transport flow** learns the remaining response and reconstructs post-perturbation expression relative to matched controls.

Prediction setup. Let p denote a perturbation target and q a query cellular context. Control and perturbed cells are independently sampled populations, without before–after cell pairing. For train-

ing condition i with perturbed mean μ_i^{pert}, we define the G-gene response relative to the mean training-control expression $\bar{b}_{\text{tr}}$:

$$\Delta_i = \mu_i^{\text{pert}} - \bar{b}_{\text{tr}}. \tag{1}$$

At inference, matched-control RNA b_q, control-ATAC context Q_q, target p, and static target/locus resources Π_p determine a response $\widehat{\Delta}_{p,q}$ and post-perturbation expression $\widehat{\mu}_{p,q}^{\text{pert}} = b_q + \widehat{\Delta}_{p,q}$. Target-LOO holds out p; Cross-State holds out state q while retaining p in other training states, with query controls available only at prediction.

3.1 Chromatin-Informed Perturbation Conditioning

Target and locus priors. For target p, frozen 64-dimensional ESM-2 protein (Lin et al., 2023) and CLM-X RNA-token (Li et al., 2026) embeddings form u_p. A 96-dimensional sequence/locus block and a 128-dimensional regulatory projection form ℓ_p:

$$\begin{aligned} u_p &= [e_p^{\text{ESM2}} \| e_p^{\text{CLMX}}] \in \mathbb{R}^{128}, \quad \ell_p = [f_p^{\text{seq}} \| f_p^{\text{reg}}] \in \mathbb{R}^{224}, \\ x_p &= [u_p \| \ell_p] \in \mathbb{R}^{352}. \end{aligned} \tag{2}$$

The sequence block uses TSS-centered features from GENCODE/RefTSS on GRCh38 (Frankish et al., 2021; Abugessaisa et al., 2019; Schneider et al., 2017); the regulatory block combines guide-library descriptors, ENCODE cCREs, and JASPAR motifs (Martin-Rufino et al., 2025; Moore et al., 2020; Castro-Mondragon et al., 2022). These priors are fixed across cells for a given target.

Control context and condition assembly. Paired control RNA–ATAC measurements (Buenrostro et al., 2015; Hao et al., 2021) provide the context $h_s = [h_s^{\text{RNA}} \| h_s^{\text{ATAC}}] \in \mathbb{R}^{160}$ for state or control group s: 96 RNA dimensions and 64 chromatin dimensions, the latter obtained by control-only TF–IDF/LSI (Stuart et al., 2021). The router and pooled flow append the explicit locus block to x_p; the state-aware branch additionally receives h_s:

$$c_p^{\text{pool}} = [x_p \| \ell_p] \in \mathbb{R}^{576}, \qquad c_{p,s}^{\text{state}} = [x_p \| \ell_p \| h_s] \in \mathbb{R}^{736}. \tag{3}$$

Thus, the state-aware condition adapts the same target priors to the measured RNA–chromatin environment. All blocks use training-fold standardization; missing priors are zero-filled.

3.2 Chromatin-Guided Response Routing

Control-derived source bank. Pooled responses can mix perturbation effects with pre-existing control-state differences. CGRR represents the latter through control-RNA offsets:

$$B_s = b_s - \bar{b}_{\text{tr}} \in \mathbb{R}^G, \tag{4}$$

where b_s is the control mean of training group s. These offsets are fixed during flow training and contain no observed perturbation effects. Let $\mathcal{P}_{\text{tr}}$ and $\mathcal{S}_{\text{tr}}$ denote the training targets and source groups.

Target-based routing. For Target-LOO, CGRR retrieves training perturbations by target/locus similarity and combines their observed state compositions $m_i(s)$:

$$w_i^{\text{feat}}(p) = \text{softmax}_i\left(-\frac{d(c_p^{\text{pool}}, c_i^{\text{pool}})}{\tau_{\text{feat}}}\right), \qquad m_p^{\text{feat}}(s) = \sum_{i \in \mathcal{P}_{\text{tr}}} w_i^{\text{feat}}(p) m_i(s), \tag{5}$$

where d is Euclidean distance after training-fold standardization and $\tau_{\text{feat}} = 0.20$.

Chromatin-based routing. For Cross-State, clean-control ATAC Q_q routes the query among training-state ATAC prototypes P_s:

$$m_q^{\text{ATAC}}(s) = \text{softmax}_s\left(\frac{\cos(Q_q, P_s)}{\tau_{\text{ATAC}}}\right), \qquad \tau_{\text{ATAC}} = 0.15. \tag{6}$$

Baseline decomposition. The routed source offset is

$$d_{p,q}^{\text{source}} = \sum_{s \in \mathcal{S}_{\text{tr}}} m_q(s) B_s, \tag{7}$$

with $m_q = m_p^{\text{feat}}$ for Target-LOO and $m_q = m_q^{\text{ATAC}}$ for Cross-State. Training uses observed compositions $m_i(s)$, with control frequencies as the fallback for insufficient support, to define the source-adjusted response:

$$\widetilde{\Delta}_i = \Delta_i - d_i^{\text{source}}. \tag{8}$$

This removes the composition-weighted control offset; for a single-state condition, $\widetilde{\Delta}_i = \mu_i^{\text{pert}} - b_s$. Cross-State reuses the ATAC weights to transfer state-centered responses, without source addback in the final readout.

3.3 Conditional Transport and Response Reconstruction

Pooled response branch. A conditional flow learns $\widetilde{\Delta}_i$. Its decoded prediction a_p, blended with the training mean using $\rho = 0.18$, gives two Target-LOO candidates:

$$\Delta_{p,q}^{\text{on}} = \mathcal{C}(d_{p,q}^{\text{source}} + a_p), \qquad \Delta_{p,q}^{\text{off}} = a_p. \tag{9}$$

Source addback restores the pooled-response convention. The stabilization operator $\mathcal{C}$ preserves supported genes and blends unsupported genes toward the pooled training response with $\alpha = 0.10$ (Equation 34). A separate flow trained directly on Δ_i supplies a stabilized clean-flow candidate.

State-centered response branch. When state structure is available, a separate flow uses $c_{p,s}^{\text{state}}$ and matched-control response targets:

$$r_{i,s} = \mu_{i,s}^{\text{pert}} - b_s, \qquad A_{p,s} = (1-\rho)a_{p,s}^{\text{state}} + \rho\bar{r}_s, \tag{10}$$

where $a_{p,s}^{\text{state}}$ is the decoded displacement and $\bar{r}_s$ the training mean response for state s. Rows with fewer than 20 perturbed cells fall back to Δ_i. Aggregating $A_{p,s}$ yields the *statebase*, which uses no source addback.

Conditional flow matching. All branches share a fixed PCA map fitted on training controls. For a branch target $y_j \in \{\widetilde{\Delta}_i, r_{i,s}, \Delta_i\}$, the latent endpoints are

$$z_0 = E_{\text{PCA}}(\bar{b}_{\text{tr}}), \qquad z_{1,j} = E_{\text{PCA}}(\bar{b}_{\text{tr}} + y_j). \tag{11}$$

With $t \sim \mathcal{U}(0,1)$, conditional flow matching (Lipman et al., 2022) uses

$$\begin{aligned} z_{t,j} &= (1-t)z_0 + tz_{1,j}, \qquad u_j^* = z_{1,j} - z_0, \\ \mathcal{L}_{\text{CFM}} &= \mathbb{E}_{j,t}\left[\left\| v_\phi(z_{t,j}, t \mid c_j) - u_j^* \right\|_2^2\right]. \end{aligned} \tag{12}$$

Training uses fold-derived reliability weights. At inference, eight Euler steps from z_0 give the displacement $D_{\text{PCA}}(z^{(8)}) - D_{\text{PCA}}(z_0)$, canceling the affine PCA offset.

Protocol-specific reconstruction. For Target-LOO, a train-only reliability gate combines statebase, source-on, source-off, and clean-flow candidates into $\Delta_{p,q}^{\text{gate}}$. It uses performance on nearby training targets, agreement with a training-response estimate, and source-shift norm guards. Cross-State bypasses the gate and transfers state-centered responses through ATAC routing:

$$\widehat{\mu}_{p,q}^{\text{pert}} = b_q + \begin{cases} \Delta_{p,q}^{\text{gate}}, & \text{Target-LOO}, \\ \displaystyle\sum_{s \in \mathcal{S}_{\text{tr}}} m_q^{\text{ATAC}}(s) A_{p,s}, & \text{Cross-State}. \end{cases} \tag{13}$$

The Cross-State branch adds the held-out state's matched-control mean b_q once, with no separate routed-source term.

Population readout. For Target-LOO, a common gate-induced adjustment preserves differences among state-specific predictions. Supplied state proportions $\pi_q(s)$ then give

$$\widehat{\mu}_{p,q}^{\text{pert}} = \sum_s \pi_q(s)\left[b_{q,s} + \widehat{\Delta}_{p,q,s}\right], \qquad \sum_s \pi_q(s) = 1. \tag{14}$$

State labels and counts supply these proportions; perturbed expression is used only for evaluation. All model fitting, PCA, standardization, and source bank construction use the outer-training split. Query controls provide the RNA baseline and, for Cross-State, the ATAC route. This gives the predicted post-perturbation expression; further details appear in Appendices C.1–C.3.

Table 1: Single-cell-line unseen-target perturbation prediction on K562. Metrics are averaged over the 18 held-out targets. §4.2

Family	Method	Public.	Mean-Response Shape		Pointwise Accuracy			Transport / Direction	
			$\text{Pearson}_\Delta \uparrow$	$\text{Pearson}_{\Delta,\text{DE}} \uparrow$	$L2 \downarrow$	MSE $\downarrow$	MAE $\downarrow$	W1 $\downarrow$	Disc. Cos. $\uparrow$
Statistical	*Control*	–	-0.0001	-0.0346	1.6135	0.0013	0.0194	6.9185	0.0001
Graph	GEARS	*Nat. Biotechnol.'24*	0.0878	0.1308	2.4023	0.0029	0.0371	6.9013	0.0825
	CellOracle	*Nature'23*	0.0572	0.0872	1.9800	0.0020	0.0245	7.6820	0.0525
	GraphVCI	*ICLR'23*	0.0657	0.1111	2.0108	0.0021	0.0248	7.6848	0.0600
Foundation	scGPT	*Nat. Methods'24*	0.0677	0.3090	2.7581	0.0038	0.0331	7.0620	0.0650
	scFoundation	*Nat. Methods'24*	0.0442	0.1681	4.5387	0.0106	0.0650	7.1529	0.0404
	GeneCompass	*Cell Res.'24*	0.1174	0.2153	2.1901	0.0024	0.0352	6.9284	0.1118
	CellFM	*Nat. Commun.'25*	0.0634	**0.3357**	5.3577	0.0143	0.0887	7.6679	0.0432
	CellVQ	*Nat. Commun.'26*	0.0430	0.0518	2.4619	0.0031	0.0353	6.9286	0.0432
Generative	CPA	*Mol. Syst.'23*	0.0972	0.2230	1.9208	0.0019	0.0276	**6.6606**	0.0958
	CellFlow	*bioRxiv'25*	**0.1437**	0.0309	2.1637	0.0024	**0.0238**	7.0127	**0.1468**
	scDFM	*ICLR'26*	0.0100	0.2373	3.6133	0.0068	0.0516	8.3818	0.0025
	PerturbDiff	*ICML'26*	-0.0300	0.2342	9.7288	0.0470	0.1123	11.2291	-0.0258
	DeSCOPE	*bioRxiv'26*	0.1286	0.3220	**1.7480**	**0.0015**	0.0282	6.9186	0.0994
	STATE	*Cell'26*	0.0316	0.1051	4.1675	0.0087	0.0554	7.1523	0.0276
	ChromaPert	*Proposed*	**0.2210**	**0.3235**	**1.5875**	**0.0013**	**0.0188**	**6.8687**	**0.2188**

4 Experiments

4.1 Experimental Setup

Datasets and protocols. We use paired RNA–ATAC from Perturb-Multiome (Martin-Rufino et al., 2025) (GSE274113; 19 targets across differentiating hematopoietic states) and CAT-ATAC (Shevade et al., 2025) (GSE288996; 18 targets in DMSO-treated K562). Both test unseen targets; GSE also evaluates 19 known targets across seven held-out states (133 conditions). Supplied test cell-type counts weight population means (Appendix B.4). Data composition, preprocessing, splits, coupling, and query controls are detailed in Appendices B.1–B.2.

Comparison methods. The 14 learned baselines comprise (1) **graph-based models**: **GEARS**, **CellOracle**, and **GraphVCI** (Roohani et al., 2024; Kamimoto et al., 2023; Wu et al., 2022); (2) **foundation models**: **scGPT**, **scFoundation**, **GeneCompass**, **CellFM**, and **CellVQ** (Cui et al., 2024; Hao et al., 2024; Yang et al., 2024; Zeng et al., 2025; Wang et al., 2026); and (3) **generative models**: **CPA**, **CellFlow**, **scDFM**, **PerturbDiff**, **DeSCOPE**, and **STATE** (Lotfollahi et al., 2023; Klein et al., 2025; Yu et al., 2026; Yuan et al., 2026; Wu et al., 2026; Adduri et al., 2026). scGPT uses its native perturbation decoder; other foundation models use frozen embeddings with a GEARS-style adapter. Training and prediction details appear in Appendix B.6.

Metrics and reporting. Methods share a gene panel and evaluator; unperturbed *Control* is excluded from learned-model rankings. Pearson_Δ correlates condition-mean responses across genes relative to matched controls; $\text{Pearson}_{\Delta,\text{DE}}$ uses the 20 largest absolute observed responses. $L2$, MSE, and MAE measure mean-expression error, discriminative cosine measures direction, and supplementary W1 evaluates exported predictions (Appendix B.6). Scores average held-out conditions (four decimals); see Appendices B.5 and A.3 for definitions and three-seed summaries.

4.2 Single-cell-line unseen-target perturbation prediction

Fixed-background target generalization. This setting isolates unseen-target generalization at a fixed cellular background. On K562 (Table 1), ChromaPert improves all-gene Pearson by 53.8% and discriminative cosine by 49.0% over CellFlow, while leading the learned predictors in pointwise errors. DE correlation and W1 retain specialized leaders, so the main advantage is broad response-shape and expression accuracy.

Chromatin-dependent biological responses. True cell–ATAC pairing improves response-slope recovery over shuffled context for all 18 targets (Figure 2a): mean all-gene Pearson is 0.2120 versus

Table 2: Multi-cell-state unseen-target perturbation prediction on GSE. Metrics are averaged over the 19 held-out targets. §4.3

Family	Method	Public.	Mean-Response Shape		Pointwise Accuracy			Transport / Direction	
			$\text{Pearson}_\Delta \uparrow$	$\text{Pearson}_{\Delta,\text{DE}} \uparrow$	$L2 \downarrow$	MSE $\downarrow$	MAE $\downarrow$	W1 $\downarrow$	Disc. Cos. $\uparrow$
Statistical	*Control*	–	0.0146	0.0333	1.2374	0.0008	0.0164	8.8965	0.0132
Graph	GEARS	*Nat. Biotechnol.'24*	0.0263	0.0199	1.8074	0.0018	0.0248	7.1760	0.0383
	CellOracle	*Nature'23*	0.0897	0.1909	1.7686	0.0016	0.0235	10.1314	0.0963
	GraphVCI	*ICLR'23*	0.0073	-0.0281	1.6742	0.0014	0.0223	10.1687	0.0070
Foundation	scGPT	*Nat. Methods'24*	0.0416	0.0537	2.3209	0.0027	0.0289	7.3371	0.0436
	scFoundation	*Nat. Methods'24*	0.0275	0.1114	9.7229	0.1753	0.1444	10.2412	0.0379
	GeneCompass	*Cell Res.'24*	**0.1030**	**0.3843**	6.7853	0.0229	0.1170	9.5319	**0.1085**
	CellFM	*Nat. Commun.'25*	0.0716	0.2386	7.2340	0.0261	0.1255	9.8175	0.0873
	CellVQ	*Nat. Commun.'26*	0.0059	-0.0427	14.7377	0.1999	0.2162	12.8551	0.0112
Generative	CPA	*Mol. Syst.'23*	0.0331	0.1913	1.5268	0.0012	0.0208	8.8482	0.0113
	CellFlow	*bioRxiv'25*	0.0698	0.1319	**1.5066**	**0.0012**	**0.0193**	**7.1029**	0.0712
	scDFM	*ICLR'26*	-0.0374	-0.0513	3.9002	0.0085	0.0606	9.1254	-0.0220
	PerturbDiff	*ICML'26*	0.0314	0.1682	8.9828	0.0402	0.0913	11.1395	0.0074
	DeSCOPE	*bioRxiv'26*	0.0268	-0.0231	1.8125	0.0018	0.0232	7.2850	-0.0214
	STATE	*Cell'26*	-0.0383	-0.0638	3.0031	0.0046	0.0430	7.3296	-0.0594
	ChromaPert	*Proposed*	**0.1735**	**0.3697**	**1.4354**	**0.0011**	**0.0186**	**7.1014**	**0.1779**

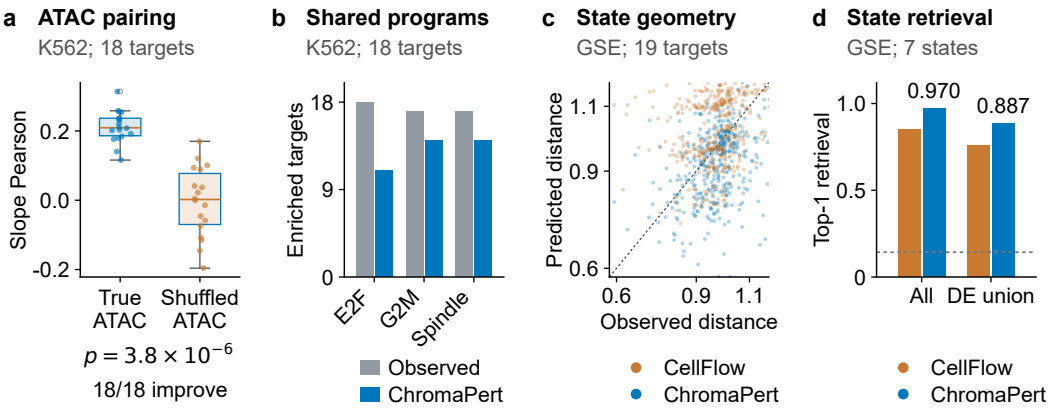

Figure 2: **Chromatin-dependent and state-resolved response structure.** §4.2 §4.4 **(a)** K562 response-slope Pearson under true/shuffled ATAC; points denote held-out targets. **(b)** K562 targets with enriched Hallmark programs in observed/predicted responses (BH $q < 0.05$). **(c)** Observed/predicted GSE response distances for 19 targets with 21 state pairs each; dashed line, equality. **(d)** Seven-state top-1 retrieval using all genes or the target-specific DE union; dashed line, $1/7$ chance retrieval.

-0.0043 (paired one-sided Wilcoxon $p = 3.8 \times 10^{-6}$). Both evaluations hold the frozen model and input dimensions fixed and vary only cell–chromatin correspondence. Top-20-gene Spearman also increases from 0.1001 to 0.2944, with improvement for 15 of 18 targets. This diagnostic evaluates chromatin-associated response variation, not benchmark Pearson_Δ, and does not establish causality. Hallmark enrichment (Liberzon et al., 2015) identifies shared proliferative programs (Figure 2b): G2M checkpoint and mitotic spindle are enriched for 14 predicted versus 17 observed targets each; E2F targets for 11 versus 18. These counts capture shared proliferative and cell-cycle responses, rather than target-wise pathway recovery (Appendix A.2).

4.3 Multi-cell-state unseen-target perturbation prediction

Multi-state target generalization. On GSE held-out-target evaluation (Table 2), ChromaPert leads five of six condition-mean metrics. Relative to the strongest learned competitors, it improves all-gene Pearson by 68.4% and discriminative cosine by 64.0%, while attaining the best pointwise errors and W1. GeneCompass retains the best DE-focused correlation. Control's lower MSE but near-zero response correlation further shows that absolute fit alone can miss response structure.

Table 3: Cross-cell-state perturbation transfer on GSE. Metrics are averaged over 133 held-out state–target conditions; STATE is unavailable in this setting. §4.4

Family	Method	Public.	Mean-Response Shape		Pointwise Accuracy			Transport / Direction	
			Pearson_{Δ} ↑	$\text{Pearson}_{\Delta,\text{DE}}$ ↑	$L2$ ↓	MSE ↓	MAE ↓	W1 ↓	Disc. Cos. ↑
Statistical	*Control*	–	0.0009	-0.0040	1.4238	0.0011	0.0169	7.0702	0.0008
Graph	GEARS	*Nat. Biotechnol.'24*	0.1013	0.1812	3.0964	0.0056	0.0439	7.9511	0.0891
	CellOracle	*Nature'23*	0.0999	0.2060	1.9807	0.0021	0.0244	7.5019	0.0942
	GraphVCI	*ICLR'23*	0.0906	0.1851	**1.9190**	**0.0020**	**0.0236**	**7.4817**	0.0852
Foundation	scGPT	*Nat. Methods'24*	0.0978	0.1686	9.4233	0.0518	0.1074	11.6138	0.0971
	scFoundation	*Nat. Methods'24*	0.0479	0.1100	11.1176	0.0689	0.1631	10.7156	0.0423
	GeneCompass	*Cell Res.'24*	0.0787	0.1555	7.6563	0.0302	0.1177	9.6319	0.0694
	CellFM	*Nat. Commun.'25*	0.0696	0.1428	7.9967	0.0327	0.1240	9.9307	0.0590
	CellVQ	*Nat. Commun.'26*	0.0228	0.0344	6.5700	0.0258	0.0977	7.8899	0.0241
Generative	CPA	*Mol. Syst.'23*	0.1236	**0.2262**	5.3573	0.0183	0.0565	8.7267	0.1215
	CellFlow	*bioRxiv'25*	**0.1484**	0.2197	4.1406	0.0236	0.0469	8.4091	**0.1478**
	scDFM	*ICLR'26*	0.0812	0.1312	9.2382	0.0490	0.1109	12.4967	0.0812
	PerturbDiff	*ICML'26*	0.0742	0.1588	12.4345	0.0825	0.1181	13.8396	0.0696
	DeSCOPE	*bioRxiv'26*	0.1010	0.1693	9.3057	0.0506	0.1042	11.6300	0.1006
	STATE	*Cell'26*	–	–	–	–	–	–	–
	ChromaPert	*Proposed*	**0.1829**	**0.3307**	**1.4309**	**0.0011**	**0.0173**	**7.0648**	**0.1809**

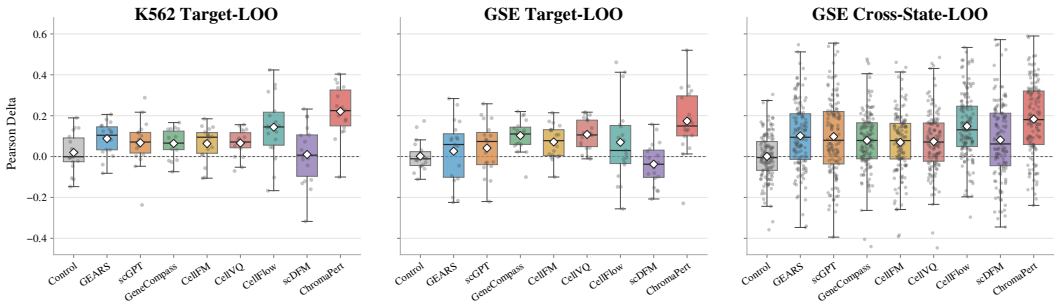

Figure 3: Fold-level distribution of the response correlation (Pearson_{Δ}) across the three held-out evaluation settings. Each box summarizes completed folds for the corresponding setting; the center line denotes the median and whiskers show the non-outlier range. §4.4

4.4 Cross-cell-state perturbation transfer

Cross-state transfer. The GSE held-out-state setting (Table 3) tests transfer of known perturbations to an unseen state using its unperturbed RNA–ATAC context. This protocol requires the model to reuse perturbation knowledge while adapting the predicted response to a cellular background absent from training. ChromaPert leads all six condition-mean metrics, improving all-gene and DE-focused Pearson by 23.2% and 46.2%, respectively, over the strongest learned baselines. Its lower expression errors and W1 show that this gain does not trade off pointwise or distributional accuracy.

State-resolved response structure. Across the 21 state pairs per target, predicted and observed response distances have target-mean Spearman 0.385 (Figure 2c). Correct-state retrieval improves over CellFlow from 0.850 to 0.970 across all genes and from 0.759 to 0.887 on the target-specific DE union (Figure 2d), supporting state-resolved response consistency rather than cell-type classification (Appendix A.2). Together, these analyses test both relative geometry between states and whether each predicted response remains closest to its matching state. Figure 3 summarizes fold-level response-correlation variability across all three evaluation settings.

4.5 Ablation Analysis

Structured use of ATAC. Table 4(a) compares ChromaPert with CellFlow and CellFlow + ATAC on 133 held-out-state conditions. ATAC concatenation lowers Pearson_{Δ} from 0.1484 to 0.0842 and $\text{Pearson}_{\Delta,\text{DE}}$ from 0.2197 to 0.0841; ChromaPert remains best on all four criteria. This favors context-aware organization over direct RNA–ATAC concatenation.

Table 4: ChromaPert ablations. §4.5 Higher is better for Pearson_Δ and $\text{Pearson}_{\Delta,\text{DE}}$; lower is better for MSE and MAE. Bold shaded rows mark the primary configuration in each panel.

Method	$\text{Pearson}_\Delta \uparrow$	$\text{Pearson}_{\Delta,\text{DE}} \uparrow$	MSE↓	MAE↓
CellFlow	0.1484	0.2197	0.0236	0.0469
CellFlow + ATAC	0.0842	0.0841	0.0073	0.0336
ChromaPert	**0.1829**	**0.3307**	**0.0011**	**0.0173**

(a) ATAC-augmented control on GSE held-out-state transfer. §4.5

Variant	$\text{Pearson}_\Delta \uparrow$	$\text{Pearson}_{\Delta,\text{DE}} \uparrow$	MSE↓	MAE↓
Direct MLP	0.1534	0.3017	0.0011	0.0186
w/o ATAC/locus match	0.1393	0.3048	0.0011	0.0190
w/o source decomposition	0.1016	0.2193	0.0011	0.0190
Full	**0.1735**	**0.3697**	**0.0011**	**0.0186**

(b) Source and flow ablations on GSE held-out-target evaluation. §4.5

Temperature	$\text{Pearson}_\Delta \uparrow$	$\text{Pearson}_{\Delta,\text{DE}} \uparrow$	MSE↓	MAE↓
0.05	0.1616	0.3363	0.0011	0.0188
0.10	0.1605	0.3359	0.0011	0.0188
0.20 (default)	**0.1735**	**0.3697**	**0.0011**	**0.0186**
0.40	0.1516	0.3272	0.0011	0.0188
0.80	0.1382	0.3091	0.0011	0.0191

(c) CGRR routing temperature on GSE held-out-target evaluation. §4.5

Aggregation	$\text{Pearson}_\Delta \uparrow$	$\text{Pearson}_{\Delta,\text{DE}} \uparrow$	MSE↓	MAE↓
Official/test	**0.1735**	**0.3697**	**0.0011**	**0.0186**
Training-control	0.1710	0.3670	0.0011	0.0187
Uniform	0.1533	0.3113	0.0011	0.0198

(d) Aggregation weights on GSE held-out-target evaluation. §4.5

Source decomposition and flow prediction. In Table 4(b), replacing the conditional flow with Direct MLP lowers Pearson_Δ from 0.1735 to 0.1534 and $\text{Pearson}_{\Delta,\text{DE}}$ from 0.3697 to 0.3017; removing routing or source decomposition also lowers correlation. Flow therefore improves response shape, while source decomposition separates control variation from perturbation changes.

Routing and aggregation sensitivity. Table 4(c) shows that intermediate routing gives the best correlations. Training-control aggregation retains $\text{Pearson}_\Delta = 0.1710$ and $\text{Pearson}_{\Delta,\text{DE}} = 0.3670$ versus 0.1735 and 0.3697 under official weights, whereas uniform weights are lower, supporting composition-aware routing. Further controls appear in Appendices A.1, A.5, and A.6.

External-prior ablation. In Table 7, removing ESM-2 gives the largest loss (Pearson_Δ 0.1735→0.1569; DE 0.3697→0.3268; cosine 0.1779→0.1637), identifying protein sequence as the main target cue among the evaluated external priors. CLM-X mainly raises MSE/MAE to 0.001101/0.018953, while cCRE and Multi-TSS removal affect all-gene and DE correlation, respectively. Pretrained-only and regulatory-only variants reach 0.1086 and 0.1585 versus 0.1735 for Full. The complementary molecular and locus priors therefore contextualize the perturbation, with conditional flow learning the remaining displacement (Table 7; Appendix A.4).

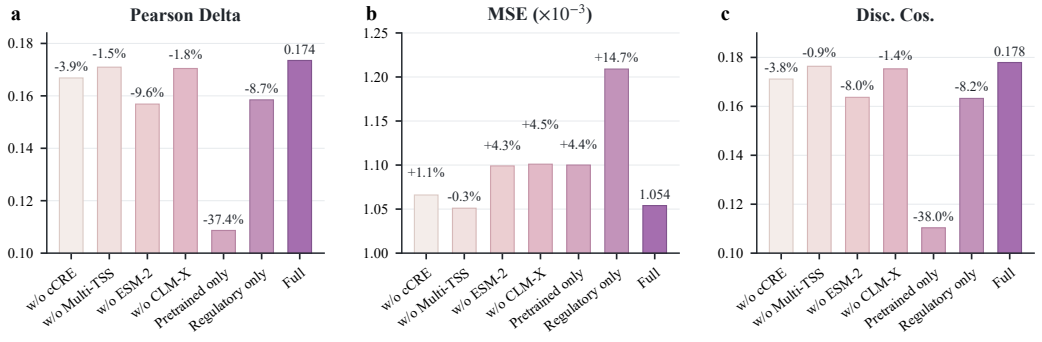

Figure 4: Prior ablations on GSE target-LOO; changes are relative to Full. §4.5

5 Conclusion

ChromaPert separates pooled baseline offsets from target-specific perturbation responses and conditions transport on paired control RNA–ATAC context. Its structured target/locus conditioning and chromatin-guided routing preserve state-dependent response structure, while conditional flow models the remaining displacement. This decomposition lets the same perturbation acquire distinct response representations across cellular contexts without treating pre-existing control differences as perturbation effects. Together, these components support expression prediction for unseen perturbations and transfer of known perturbations to new cellular states within a common formulation for learning from unpaired cell populations, making measured chromatin context directly actionable for response prediction across diverse cellular backgrounds and targets.

Reproducibility Statement

Section 3 describes the model formulation, training target, inference procedure, and information boundaries. Dataset provenance, preprocessing, fixed evaluation splits, prediction-time control pools, response construction, aggregation, and metric definitions are documented in Appendices B.1–B.5. Comparator interfaces are described in Appendix B.6; prior construction and numerical flow details appear in Appendices C.1–C.2; and the architecture, optimization settings, and primary hyperparameters are given in Appendix C.3. Repeat-run summaries and controlled ablations are reported in Appendices A.3, A.5, and A.6. Our code is provided in the anonymous supplementary material and includes the complete implementation framework. We will release the complete codebase, model checkpoints, and documentation. All datasets used in this work are publicly available through their original repositories, with accession information in Appendix B.1.

Ethics Statement

This work is a secondary computational analysis of publicly available single-cell datasets. We recruited no new human participants, performed no clinical intervention, and accessed no direct personal identifiers. Raw data are not redistributed; access and reuse remain subject to the terms of the original studies and repositories. ChromaPert is intended for computational research and hypothesis generation, not for clinical diagnosis or treatment selection. Its predictions may inherit technical or biological biases from the source assays, cell-state annotations, and external reference resources, and therefore require independent experimental validation before biomedical use.

AI Use Statement

Generative AI tools assisted manuscript editing and organization, discussion of author-supplied results, and figure preparation. The authors are responsible for checking AI-assisted content against the underlying data, code, and references, and for the final text, claims, equations, figures, and results.

References

Imad Abugessaisa, Shuhei Noguchi, Akira Hasegawa, Atsushi Kondo, Hideya Kawaji, Piero Carninci, and Takeya Kasukawa. reftss: a reference data set for human and mouse transcription start sites. *Journal of molecular biology*, 431(13):2407–2422, 2019.

Abhinav K Adduri, Dhruv Gautam, Beatrice Bevilacqua, Mohsen Naghipourfar, Alishba Imran, Rohan Shah, Noam Teyssier, Rishi Verma, Christopher Carpenter, Basak Eraslan, et al. Predicting cellular responses to perturbation across diverse contexts with state. *Cell*, 2026.

Sara Aibar, Carmen Bravo González-Blas, Thomas Moerman, Vân Anh Huynh-Thu, Hana Imrichova, Gert Hulselmans, Florian Rambow, Jean-Christophe Marine, Pierre Geurts, Jan Aerts, et al. Scenic: single-cell regulatory network inference and clustering. *Nature methods*, 14(11): 1083–1086, 2017.

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.

Simon C Biddie, Sam John, Pete J Sabo, Robert E Thurman, Thomas A Johnson, R Louis Schiltz, Tina B Miranda, Myong-Hee Sung, Saskia Trump, Stafford L Lightman, et al. Transcription factor ap1 potentiates chromatin accessibility and glucocorticoid receptor binding. *Molecular cell*, 43(1):145–155, 2011.

Jason D Buenrostro, Beijing Wu, Ulrike M Litzenburger, Dave Ruff, Michael L Gonzales, Michael P Snyder, Howard Y Chang, and William J Greenleaf. Single-cell chromatin accessibility reveals principles of regulatory variation. *Nature*, 523(7561):486–490, 2015.

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.

Jaime A Castro-Mondragon, Rafael Riudavets-Puig, Ieva Rauluseviciute, Roza Berhanu Lemma, Laura Turchi, Romain Blanc-Mathieu, Jeremy Lucas, Paul Boddie, Aziz Khan, Nicolás Manosalva Pérez, et al. Jaspar 2022: the 9th release of the open-access database of transcription factor binding profiles. *Nucleic acids research*, 50(D1):D165–D173, 2022.

Changxi Chi, Jun Xia, Yufei Huang, Jingbo Zhou, Siyuan Li, Yunfan Liu, Chang Yu, and Stan Z Li. Unlasting: Unpaired single-cell multi-perturbation estimation by dual conditional diffusion implicit bridges. *arXiv e-prints*, pp. arXiv–2506, 2025a.

Changxi Chi, Jun Xia, Jingbo Zhou, Jiabei Cheng, Chang Yu, and Stan Z Li. Grape: Heterogeneous graph representation learning for genetic perturbation with coding and non-coding biotype. *arXiv preprint arXiv:2505.03853*, 2025b.

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.

Atray Dixit, Oren Parnas, Biyu Li, Jenny Chen, Charles P Fulco, Livnat Jerby-Arnon, Nemanja D Marjanovic, Danielle Dionne, Tyler Burks, Raktima Raychowdhury, et al. Perturb-seq: dissecting molecular circuits with scalable single-cell rna profiling of pooled genetic screens. *cell*, 167(7): 1853–1866, 2016.

Adam Frankish, Mark Diekhans, Irwin Jungreis, Julien Lagarde, Jane E Loveland, Jonathan M Mudge, Cristina Sisu, James C Wright, Joel Armstrong, If Barnes, et al. Gencode 2021. *Nucleic acids research*, 49(D1):D916–D923, 2021.

Jeffrey M Granja, M Ryan Corces, Sarah E Pierce, S Tansu Bagdatli, Hani Choudhry, Howard Y Chang, and William J Greenleaf. Archr is a scalable software package for integrative single-cell chromatin accessibility analysis. *Nature genetics*, 53(3):403–411, 2021.

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.

Yuhan Hao, Stephanie Hao, Erica Andersen-Nissen, William M Mauck, Shiwei Zheng, Andrew Butler, Maddie J Lee, Aaron J Wilk, Charlotte Darby, Michael Zager, et al. Integrated analysis of multimodal single-cell data. *Cell*, 184(13):3573–3587, 2021.

Sam John, Peter J Sabo, Robert E Thurman, Myong-Hee Sung, Simon C Biddie, Thomas A Johnson, Gordon L Hager, and John A Stamatoyannopoulos. Chromatin accessibility pre-determines glucocorticoid receptor binding patterns. *Nature genetics*, 43(3):264–268, 2011.

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

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.

Bowen Li, Ziqiang Liu, Zhen Wang, Zhenyu Xu, Yantao Li, Chulin Sha, and Xiaolin Li. Clm-x: A multimodal single-cell foundation model with flexible multi-way transformer for unified scrna-seq and scatac-seq analysis. *bioRxiv*, pp. 2026–02, 2026.

Arthur Liberzon, Chet Birger, Helga Thorvaldsdóttir, Mahmoud Ghandi, Jill P Mesirov, and Pablo Tamayo. The molecular signatures database hallmark gene set collection. *Cell systems*, 1(6): 417–425, 2015.

Zeming Lin, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Nikita Smetanin, Robert Verkuil, Ori Kabeli, Yaniv Shmueli, et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379(6637):1123–1130, 2023.

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.

Sai Ma, Bing Zhang, Lindsay M LaFave, Andrew S Earl, Zachary Chiang, Yan Hu, Jiarui Ding, Alison Brack, Vinay K Kartha, Tristan Tay, et al. Chromatin potential identified by shared single-cell profiling of rna and chromatin. *Cell*, 183(4):1103–1116, 2020.

Jorge Diego Martin-Rufino, Alexis Caulier, Seayoung Lee, Nicole Castano, Emily King, Samantha Joubran, Marcus Jones, Seth R Goldman, Uma P Arora, Lara Wahlster, et al. Transcription factor networks disproportionately enrich for heritability of blood cell phenotypes. *Science*, 388(6742): 52–59, 2025.

Jill E Moore, Michael J Purcaro, Henry E Pratt, Charles B Epstein, Noam Shoresh, Jessika Adrian, Trupti Kawli, Carrie A Davis, Alexander Dobin, et al. Expanded encyclopaedias of dna elements in the human and mouse genomes. *Nature*, 583(7818):699–710, 2020.

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.

Xiaoning Qi, Lianhe Zhao, Chenyu Tian, Yueyue Li, Zhen-Lin Chen, Peipei Huo, Runsheng Chen, Xiaodong Liu, Baoping Wan, Shengyong Yang, et al. Predicting transcriptional responses to novel chemical perturbations using deep generative model for drug discovery. *Nature Communications*, 15(1):9256, 2024.

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

Valerie A Schneider, Tina Graves-Lindsay, Kerstin Howe, Nathan Bouk, Hsiu-Chuan Chen, Paul A Kitts, Terence D Murphy, Kim D Pruitt, Françoise Thibaud-Nissen, Derek Albracht, et al. Evaluation of grch38 and de novo haploid genome assemblies demonstrates the enduring quality of the reference assembly. *Genome research*, 27(5):849–864, 2017.

Kaivalya Shevade, Yeqing Angela Yang, Kevin Feng, Karl Mader, Volkan Sevim, Jacob Parsons, Gunisha Arora, Hasnaa Elfawy, Rachel Mace, Scot Federman, et al. Simultaneous capture of single cell rna-seq, atac-seq, and crispr perturbation enables multiomic screens to identify gene regulatory relationships. *Cell Reports Methods*, 5(12), 2025.

Tim Stuart, Avi Srivastava, Shaista Madad, Caleb A Lareau, and Rahul Satija. Single-cell chromatin state analysis with signac. *Nature methods*, 18(11):1333–1341, 2021.

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.

Jue Wang, Cheng Tan, Zhangyang Gao, Sida Shao, Shiping Liu, and Stan Z Li. Illuminating cell states by a comprehensive and interpretable single cell foundation model. *Nature Communications*, 17(1):4037, 2026.

Pengpeng Wu, Hailin Wei, Yazi Li, Xinwei Zheng, Caibin Zhou, Xihao Hu, and Chenfei Wang. Decoding single-cell omics of perturbation responses using descope. *bioRxiv*, pp. 2026–04, 2026.

Yulun Wu, Robert A Barton, Zichen Wang, Vassilis N Ioannidis, Carlo De Donno, Layne C Price, Luis F Voloch, and George Karypis. Predicting cellular responses with variational causal inference and refined relational information. *arXiv preprint arXiv:2210.00116*, 2022.

Xiaodong Yang, Guole Liu, Guihai Feng, Dechao Bu, Pengfei Wang, Jie Jiang, Shubai Chen, Qinmeng Yang, Hefan Miao, Yiyang Zhang, et al. Genecompass: deciphering universal gene regulatory mechanisms with a knowledge-informed cross-species foundation model. *Cell Research*, 34 (12):830–845, 2024.

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.

Yuansong Zeng, Jiancong Xie, Ningyuan Shangguan, Zhuoyi Wei, Wenbing Li, Yun Su, Shuangyu Yang, Chengyang Zhang, Jinbo Zhang, Nan Fang, et al. Cellfm: a large-scale foundation model pre-trained on transcriptomics of 100 million human cells. *Nature Communications*, 16(1):4679, 2025.

A SUPPLEMENTARY EXPERIMENTS

A.1 DIRECT REGRESSION AND AGGREGATION CONTROLS

Matched residual-prediction task. We compare ChromaPert with Direct MLP on all 19 GSE held-out-target folds using the configuration in Table 2. Both retain the same target/locus and control-state conditions, CGRR source decomposition, and PCA residual target. This comparison isolates the contribution of flow relative to direct regression under the same source-estimation procedure.

Aggregation and evaluation. Three rules aggregate state-specific responses: **Official/test** uses each target's supplied full test cell-type counts; **Training-control** uses training-control frequencies; and **Uniform** weights the aggregation groups equally. Group support, the matched-control baseline, test cells, and the evaluator remain fixed. After response aggregation and baseline addition, each condition mean is exported as 32 identical rows. The row count is an export convention, not the number of observed test cells.

Table 5: Flow versus direct regression under three aggregation rules on GSE held-out-target evaluation (19 targets). Bold marks the better value per pair; displayed ties are both bold.

Aggregation	Backbone	Pearson_{Δ} ↑	$\text{Pearson}_{\Delta,\text{DE}}$ ↑	$L2$ ↓	MSE↓	MAE↓
Official/test	Direct MLP	0.153359	0.301672	1.437177	0.001057	**0.018598**
	ChromaPert	**0.173501**	**0.369660**	**1.435411**	**0.001054**	0.018632
Training-control	Direct MLP	0.150976	0.298344	1.439621	0.001060	**0.018680**
	ChromaPert	**0.170987**	**0.366970**	**1.438524**	**0.001058**	0.018739
Uniform	Direct MLP	0.138576	0.263558	**1.465007**	**0.001096**	**0.019482**
	ChromaPert	**0.153270**	**0.311312**	1.465979	**0.001096**	0.019792

Flow improves response correlations beyond source estimation. ChromaPert has higher Pearson_{Δ} and $\text{Pearson}_{\Delta,\text{DE}}$ than Direct MLP under every aggregation rule. With official weights, the relative gains are 13.1% and 22.5%, respectively. Absolute errors are close: Direct MLP has slightly lower MAE in all three comparisons and slightly lower $L2$ under uniform weights. Flow's advantage in this comparison is therefore concentrated in the cross-gene mean-response pattern, with no uniform improvement in absolute error.

Sensitivity to composition weights. Training-control aggregation retains 98.6% of ChromaPert's official Pearson_{Δ} and 99.3% of its $\text{Pearson}_{\Delta,\text{DE}}$. Uniform aggregation lowers these correlations to 0.153270 and 0.311312, showing sensitivity to population weighting. Training-control frequencies retain most of the correlation advantage, and flow has higher response correlations than MLP under all three rules. These comparisons vary only the response weights: group support and the composition-weighted control reference remain fixed.

A.2 BIOLOGICAL RESPONSE AND CHROMATIN-STATE AUDITS

Chromatin pairing and shared functional programs. For each of the 18 held-out K562 targets, the response-slope diagnostic compares true control cell–chromatin pairing against ATAC context shuffled across cells. Both conditions use the same frozen model and input dimensions. All-gene slope Pearson averages 0.2120 with true pairing and -0.0043 after shuffling, with an improvement for every target and a paired one-sided Wilcoxon $p = 3.81 \times 10^{-6}$. Top-20-gene Spearman averages 0.2944 versus 0.1001, with improvements for 15 of 18 targets. These statistics quantify recovery of chromatin-associated response slopes, whereas benchmark Pearson_{Δ} measures condition-mean expression changes. Figure A1a shows the full target-wise comparison. The accompanying context-swap audit reports nonzero prediction changes for all 18 targets, averaging 9.080×10^{-3} in absolute magnitude.

Hallmark gene sets (Liberzon et al., 2015) summarize recurring functional programs in observed and predicted responses. Figure A1c expands the three proliferation-related programs shown in the

main figure. Predictions contain G2M-checkpoint and mitotic-spindle enrichment for 14 targets each, compared with 17 in the observations, and E2F enrichment for 11 versus 18. The original pathway audit also reports mTORC1 enrichment for three predicted versus six observed targets, and apoptosis and p53 enrichment for one versus two each. These counts summarize how often each program is enriched; they do not measure agreement for individual targets.

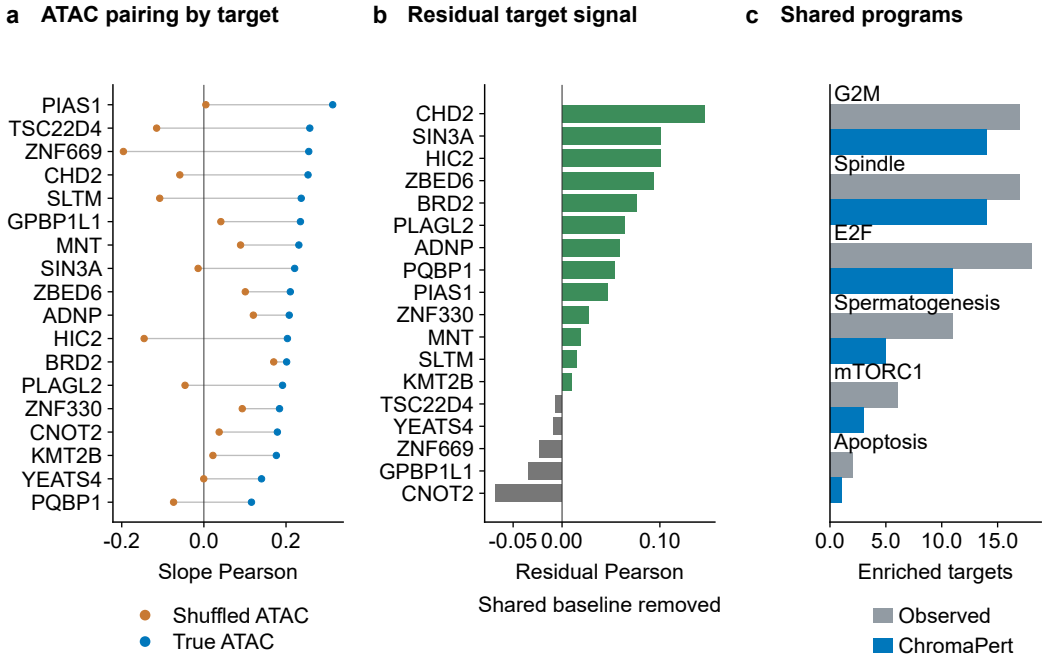

Figure A1: **Extended K562 response audits. (a)** Target-wise response-slope Pearson under true and shuffled ATAC pairing, with lines connecting results for the same target. **(b)** Residual target signal after removing the shared response baseline. Positive and negative values are shown in green and gray. **(c)** The expanded set of shared response programs displayed in the state-discovery analysis. Bars count enriched targets in observed and predicted responses. E2F, G2M, and spindle abbreviate E2F targets, G2M checkpoint, and mitotic spindle.

Candidate target-specific signals. After shared-baseline removal, CHD2, SIN3A, HIC2, ZBED6, and BRD2 have the largest positive residual signals in the target audit (Figure A1b). The ranking identifies candidate chromatin-sensitive perturbations for follow-up. It remains exploratory because target-label permutation provides limited support for pathway specificity. The functional interpretation therefore rests primarily on the shared proliferation programs.

Held-out-state response geometry and retrieval. The GSE analysis uses the seven annotated states from held-out-state evaluation. For each of 19 targets, all $\binom{7}{2} = 21$ state pairs contribute an observed and a predicted response distance. Figure A2 reports the pooled comparison and the target-wise Spearman correlations. ChromaPert's mean target-wise Spearman is 0.385 and its pooled Spearman is 0.396. The per-target results vary across perturbations, and CellFlow has the higher correlation for some targets. We report the target-stratified summary because pairs sharing a state or target are dependent.

The retrieval audit matches each predicted state–target response against the seven observed state responses for that same target. All-gene top-1 retrieval is 0.970 for ChromaPert and 0.850 for CellFlow. Restricting the comparison to the target-specific DE union gives 0.887 and 0.759, respectively. The reference level for uniform retrieval among seven states is $1/7$. Retrieval measures retention of state information in predicted responses, complementing the comparison of relative distances.

Label-free control-chromatin variation. The state-discovery analysis is a post-hoc audit of control chromatin, carried out without retraining ChromaPert. It uses a control-fitted, technically residualized 32-axis ATAC representation combining fixed-bin, cCRE, gene-activity, and motif-activity

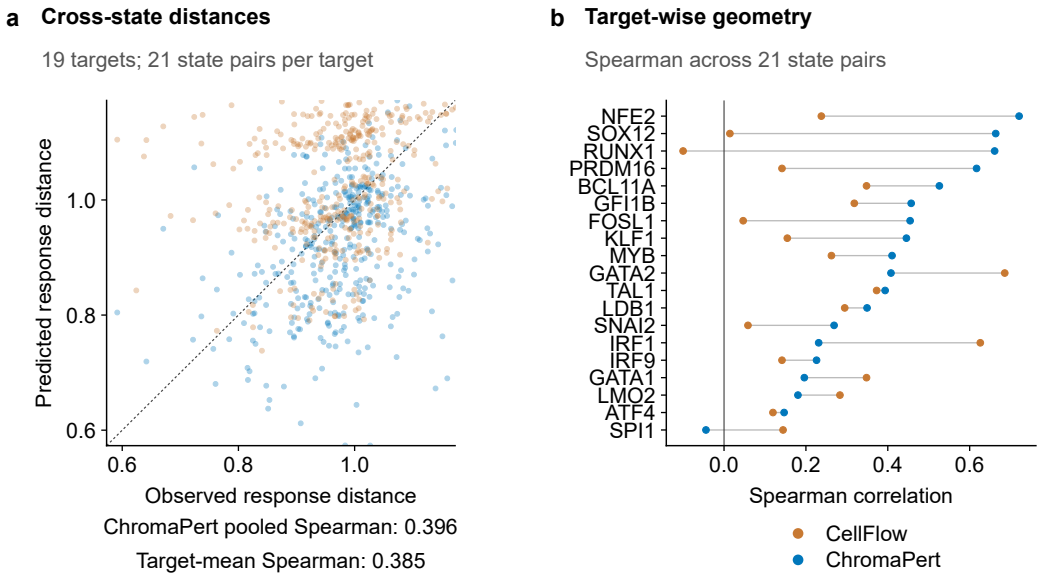

Figure A2: **GSE held-out-state response geometry at target resolution. (a)** Observed and predicted response distances, with each point representing one of 21 state pairs for a target. The dashed diagonal denotes equal distance. **(b)** Spearman correlation over the 21 pairs, shown separately for all 19 targets and both methods. Lines connect the same target, and the vertical reference marks zero correlation.

views. Of the 5,412 metadata-aligned cells in this analysis subset, 1,261 are controls: 547 from DMSO1 and 714 from DMSO2. These counts describe the audit subset, which is distinct from the benchmark partitions in Table 11. After unit-norm normalization, KMeans with the exploratory choice $k = 4$ defines four candidate chromatin regimes using control ATAC alone. Cell-state annotations and post-perturbation ATAC are excluded from this clustering step. Figure A3 shows the candidate partitions, their representation in both replicates, and bidirectional replicate-transfer agreement. Transfer ARI is 0.860 from DMSO1 to DMSO2 and 0.886 in the reverse direction, averaging 0.873. This agreement across replicates supports using control-chromatin variation as context for response prediction.

A.3 Repeat-run performance

Table 6 summarizes protocol-level scores across three random seeds as means and standard deviations. These summaries complement the primary evaluation tables in the main text. The percentage gains in the abstract and introduction use the primary evaluation tables, not these three-seed means. The standard deviations describe between-run variation; they are not confidence intervals over held-out targets or states.

Table 6(a) shows higher mean-response correlations and discriminative cosine, together with lower $L2$, MSE, and MAE, for ChromaPert than CellFlow on K562. The same pattern holds on GSE target-LOO (b) and cross-state-LOO (c). ChromaPert also has lower supplementary W1 on K562 and cross-state transfer; CellFlow retains a slightly lower W1 on GSE target-LOO.

A.4 External-prior ablation

Table 7 reports the complete external-prior ablation.

Protein and RNA target representations. ESM-2 removal reduces Pearson_{Δ} from 0.173501 to 0.156857 and $\text{Pearson}_{\Delta,\text{DE}}$ from 0.369660 to 0.326840, the largest correlation losses among single-prior removals. Protein language models encode sequence-derived structural information (Lin et al., 2023). These representations connect held-out protein targets to training targets with similar molecular properties. CLM-X removal has a smaller effect on these correlations but raises MSE from

Table 6: Three-seed comparison of CellFlow and ChromaPert (mean and SD). Bold marks the better mean.

Metric	CellFlow		ChromaPert	
	Mean	SD	Mean	SD
Pearson_{Δ} ↑	0.1038	0.0117	**0.230283**	0.001164
$\text{Pearson}_{\Delta,\text{DE}}$ ↑	0.0486	0.0173	**0.325833**	0.006795
$L2$ ↓	2.1344	0.0715	**1.588487**	0.000941
MSE ↓	0.002318	0.000153	**0.001289**	0.000001
MAE ↓	0.02363	0.00064	**0.018822**	0.000005
W1 ↓	7.0740	0.0179	**6.869656**	0.000102
Disc. Cos. ↑	0.1047	0.0140	**0.228007**	0.000967

(a) K562 target-LOO (18 targets).

Metric	CellFlow		ChromaPert	
	Mean	SD	Mean	SD
Pearson_{Δ} ↑	0.071317	0.008266	**0.158009**	0.017618
$\text{Pearson}_{\Delta,\text{DE}}$ ↑	0.147989	0.036863	**0.334136**	0.039319
$L2$ ↓	1.501639	0.007458	**1.455663**	0.022097
MSE ↓	0.001145	0.000010	**0.001083**	0.000039
MAE ↓	0.019312	0.000058	**0.018831**	0.000222
W1 ↓	**7.101689**	0.001898	7.105784	0.005875
Disc. Cos. ↑	0.073415	0.008620	**0.162550**	0.017430

(b) GSE target-LOO (19 targets).

Metric	CellFlow		ChromaPert	
	Mean	SD	Mean	SD
Pearson_{Δ} ↑	0.0881	0.0048	**0.1757**	0.0010
$\text{Pearson}_{\Delta,\text{DE}}$ ↑	0.0926	0.0156	**0.3251**	0.0016
$L2$ ↓	3.0498	0.0580	**1.4323**	0.0003
MSE ↓	0.007033	0.000395	**0.001101**	0.000001
MAE ↓	0.03312	0.00059	**0.01727**	0.00001
W1 ↓	7.4933	0.0304	**7.0825**	0.0002
Disc. Cos. ↑	0.0883	0.0049	**0.1737**	0.0011

(c) GSE cross-state-LOO (133 state–target conditions).

Table 7: External-prior ablation on GSE target-LOO.

Variant	Pearson_{Δ}	$\text{Pearson}_{\Delta,\text{DE}}$	$L2$	MSE	MAE	W1	Disc. Cos.
Full	0.173501	0.369660	1.4354	0.001054	0.018632	7.1014	0.177905
w/o cCRE	0.166821	0.373276	1.4458	0.001066	0.018758	7.1042	0.171148
w/o Multi-TSS	0.170940	0.351339	1.4328	0.001051	0.018580	7.0987	0.176367
w/o ESM-2	0.156857	0.326840	1.4650	0.001099	0.018886	7.1079	0.163678
w/o CLM-X	0.170435	0.361918	1.4661	0.001101	0.018953	7.1096	0.175327
Pretrained priors only	0.108649	0.248848	1.4674	0.001100	0.019008	7.1028	0.110360
Regulatory priors only	0.158469	0.316061	1.5222	0.001209	0.019461	7.1274	0.163237

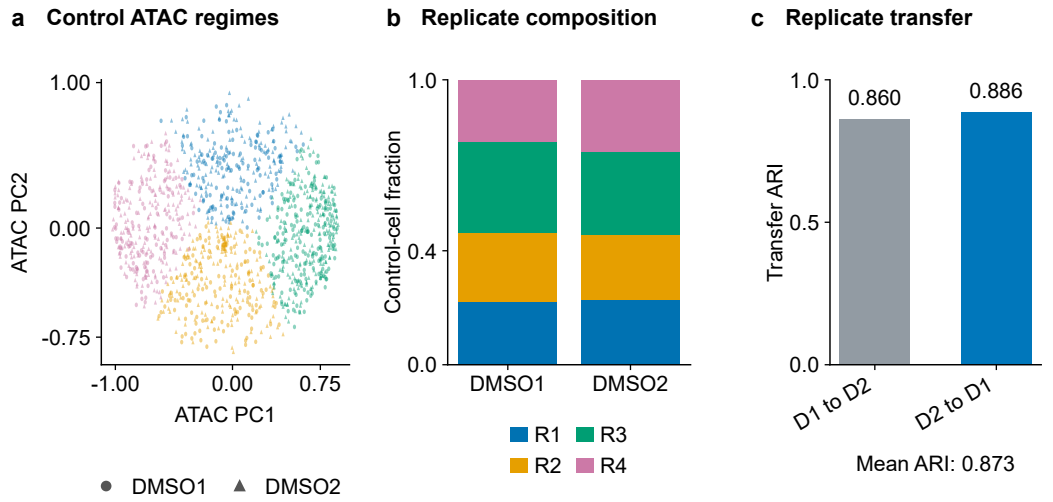

Figure A3: **Candidate chromatin regimes in control K562 cells.** **(a)** PCA display of the normalized 32-axis ATAC representation, colored by the four KMeans regimes. Circles and triangles distinguish DMSO1 and DMSO2. **(b)** Regime proportions within each control replicate. **(c)** Bidirectional replicate-transfer ARI. D1 and D2 denote DMSO1 and DMSO2; ARI measures partition agreement. Clustering is a post-hoc input audit with $k = 4$, while ChromaPert conditions on continuous chromatin features.

0.001054 to 0.001101 and MAE from 0.018632 to 0.018953. Its frozen RNA-token representation provides complementary target information for predicting the magnitude of expression changes.

Regulatory-locus and promoter-sequence context. Removing cCRE decreases all-gene correlation from 0.173501 to 0.166821 and discriminative cosine from 0.177905 to 0.171148, while DE correlation rises slightly to 0.373276. cCRE annotations describe candidate regulatory elements (Moore et al., 2020). These results show that target-locus context contributes to gene-wide response prediction, with a different trade-off on the top-response subset. Multi-TSS removal lowers DE correlation to 0.351339 while slightly improving MSE and MAE. TSS-centered sequence windows represent alternative promoter contexts associated with a target (Abugessaisa et al., 2019). Their value may lie in distinguishing the relative responses of strongly affected genes rather than minimizing global expression error. These metric-specific effects motivate combining protein-level target information with locus-level regulatory context.

Complementarity across prior families. Pretrained-only and regulatory-only variants reach all-gene correlations of 0.108649 and 0.158469, respectively, below 0.173501 for Full. Their DE correlations and directional agreement are also lower. This advantage shows the benefit of combining molecular target descriptors with regulatory context. The individual ablations show which prediction criteria are most sensitive to each information source.

A.5 Controlled input and architecture baselines

Table 8: Controlled input and architecture baselines on GSE target-LOO. Lower $L2$, MSE, MAE, and W1 are better; higher correlations and discriminative cosine are better.

Method	Pearson$_\Delta$ ↑	Pearson$_{\Delta,\mathrm{DE}}$ ↑	$L2$ ↓	MSE ↓	MAE ↓	W1 ↓	Disc. Cos. ↑
Concat-Linear-ATAC	-0.018874	-0.053584	1.895157	0.001942	0.024209	7.305901	-0.013298
Concat-Ridge-ATAC	-0.014108	-0.041709	1.889544	0.001929	0.024143	7.304001	-0.008318
Concat-MLP-ATAC	0.015239	-0.017022	2.132850	0.002410	0.027351	8.050125	0.032644
RNA-only MLP	0.038374	0.026677	2.034526	0.002155	0.026057	7.776974	0.049617
Additive mean shift	0.006048	0.030176	1.715645	0.001614	0.022049	7.269627	0.015198
ChromaPert	**0.173501**	**0.369660**	**1.435411**	**0.001054**	**0.018632**	**7.101355**	**0.177905**

These controls compare structured chromatin conditioning with direct ATAC concatenation, RNA-only regression, and an additive mean-shift control. The unperturbed Control reference is reported

in the main benchmark (Table 2). ChromaPert leads on all six condition-mean metrics, combining higher response correlations with lower absolute errors. It also has the lowest supplementary W1 among these controls.

A.6 Numerical sensitivity and inference cost

Table 9: NFE accuracy–latency trade-off on GSE target-LOO.

NFE	Pearson$_\Delta$ ↑	Pearson$_{\Delta,\mathrm{DE}}$ ↑	MSE ↓	MAE ↓	Query (s) ↓
5	0.173494	0.369639	0.001054	0.018632	0.004764
8 (default)	**0.173493**	**0.369657**	**0.001054**	**0.018632**	**0.006947**
10	0.173497	0.369667	0.001054	0.018631	0.008818
20	0.173500	0.369691	0.001053	0.018631	0.016500
50	0.173500	0.369703	0.001053	0.018631	0.036394

ODE integration steps. Table 9 varies the number of function evaluations from 5 to 50. Predictive metrics change little, whereas mean branch-query time rises from 0.004764 to 0.036394 seconds. We therefore retain NFE$= 8$: within the tested range, additional ODE steps increase latency without a material accuracy gain.

Transport-mean blend. Table 10 reports blend sensitivity on GSE target-LOO. The primary configuration uses a blend of 0.18. Across the tested settings, absolute errors remain similar, whereas mean-response correlations vary more noticeably.

Table 10: Transport-mean blend on GSE target-LOO. The bold shaded row marks the frozen default configuration.

Blend	Pearson$_\Delta$ ↑	Pearson$_{\Delta,\mathrm{DE}}$ ↑	MSE↓	MAE↓
0.00	0.1573	0.3366	0.0011	0.0189
0.09	0.1581	0.3355	0.0011	0.0190
0.18 (default)	**0.1735**	**0.3697**	**0.0011**	**0.0186**
0.36	0.1588	0.3391	0.0011	0.0187

B Data and Evaluation Protocols

B.1 Dataset details and preprocessing

We use the dataset names from the original studies. *GSE274113 Perturb-Multiome* (Martin-Rufino et al., 2025) (GEO accession GSE274113) is a paired RNA–ATAC–guide assay of primary human hematopoietic cells undergoing in-vitro erythroid differentiation. The collection includes days 7, 9, 11, and 14 and 14 biological replicates. The benchmark retains the non-targeting (NT) control and 19 target labels: `ATF4`, `BCL11A`, `FOSL1`, `GATA1`, `GATA2`, `GFI1B`, `IRF1`, `IRF9`, `KLF1`, `LDB1`, `LMO2`, `MYB`, `NFE2`, `PRDM16`, `RUNX1`, `SNAI2`, `SOX12`, `SPI1`, and `TAL1`. Metadata contain 137,604 non-empty cells, with 32 fine-grained cell types and ten consolidated labels.

GSE288996 CAT-ATAC K562 (Shevade et al., 2025) (GEO accession GSE288996) is a paired K562 CAT-ATAC perturbation assay. Although the raw accession contains iPSC, Dasatinib, and DMSO groups, the benchmark uses only the two biological replicates of K562 DMSO. The strict model subset contains approximately 11,928 cells (8,347 targeting and 3,581 non-targeting) and 18 targets: `ADNP`, `BRD2`, `CHD2`, `CNOT2`, `GPBP1L1`, `HIC2`, `KMT2B`, `MNT`, `PIAS1`, `PLAGL2`, `PQBP1`, `SIN3A`, `SLTM`, `TSC22D4`, `YEATS4`, `ZBED6`, `ZNF330`, and `ZNF669`. No cell-type label is used for this setting.

RNA and promoter ATAC counts are independently normalized to a library size of 10,000 and transformed with `log1p`. The model-ready files use a single normalization stage and retain raw library sizes, dataset-specific scalers, and the canonical gene order in the manifest. Gene-aligned promoter accessibility is stored with the `ATAC_PROM:` prefix, while RNA columns use the `RNA:` prefix. The benchmark uses 2,013 or 2,018 promoter coordinates. Extra genes are ignored by the adapter. Missing canonical genes must be resolved by the fixed mapping before a method is scored.

Table 11: Model-ready datasets and stored partitions. Protocol-specific leave-one-out masks are described separately below.

Dataset	Features			Fixed partitions (cells)			Protocols
	RNA	ATAC	Total	Train	Val.	Test	
Perturb-Multiome GSE274113	2,013	2,013	4,026	107,471	11,798	18,335	Target-LOO; cross-state-LOO
CAT-ATAC K562 GSE288996 (DMSO)	2,018	2,018	4,036	4,063	878	471	Target-LOO

The three evaluation settings use fixed masks. GSE target-LOO has 19 folds; each fold trains on NT controls and the other 18 targets and scores the held-out target. Across folds, evaluation covers 19 conditions and 2,432 observed test cells (19×128). K562 target-LOO has 18 folds; each fold trains on non-targeting controls and the other 17 targets. Evaluation covers 2,122 observed test cells in total. CNOT2, PLAGL2, and PQBP1 retain 53, 80, and 69 truth cells, respectively, because fewer cells pass the strict filter. GSE cross-state-LOO holds out seven erythroid states, producing $7 \times 19 = 133$ state–target conditions and 16,833 common-evaluation cells. The held-out states are `BFU-E`, `Basophilic Erythroblast`, `CFU-E`, `EoBasoMast Precursor`, `Orthochromatic Erythroblast`, `Polychromatic Erythroblast`, and `Pro-Erythroblast`. MEP, megakaryocyte, and myeloid precursor labels are excluded to define the fixed seven-state trajectory.

All methods use a common evaluation manifest specifying the same condition list, canonical gene order, truth cells, control reference, residual definition, metric implementation, and random seed. In GSE target-LOO, the prediction-time input contains a fixed pool of 1,024 control cells and the evaluator uses frozen cell-type-matched clean pools (up to 2,048 cells per pool) to compute residuals. Cross-state transfer may use clean controls from the held-out state only at prediction time. K562 uses one fixed pool of 1,024 global DMSO controls for both prediction and evaluation. These objects are described separately from the training split in Appendix B.2.

B.2 Fold-restricted protocol

The model is fitted separately within each fixed split. The promoter mapping M, cell-wise library-size transforms, canonical gene order, and fixed sequence, TSS, cCRE, and motif resources are defined before splitting. ATAC projections, prototypes, pooled control references, control-state offsets, source mixtures, and response scales are recomputed from the training portion of each fold. All trainable parameters are estimated from this portion alone.

For a state-indexed setting, the training coupling is

$$\pi_{s,p} = D_0(s) \otimes D_1(s,p), \qquad \pi_p = \sum_{s \in \mathcal{S}_p} \omega_{s|p} \pi_{s,p}, \qquad \omega_{s|p} = \frac{n_{s,p}}{\sum_{s'} n_{s',p}}, \tag{15}$$

where $D_0(s)$ and $D_1(s,p)$ are the training control and perturbed distributions in state s, and $n_{s,p}$ counts training perturbed cells. The mixture and its normalization use the same set of states with available training support. The pooled law applies to no-state settings and the pooled-only ablation. Product coupling describes independent sampling from the two populations, not observed cell pairs. The source-adjusted flow target is the condition-mean response in Equation 8. For a held-out state S, the transfer fit mask is $\mathcal{I}_{\text{fit}}(S) = \{i : s_i \neq S\}$; clean controls from S may be used as prediction-time source context. This protocol evaluates known-target transfer to a new cellular context.

The three frozen evaluation settings are GSE Target-LOO (19 held-out targets), K562 Target-LOO (18 held-out targets), and GSE Cross-State-LOO (7 states by 19 targets, or 133 held-out state–target conditions). Test rows are capped at 128 cells per condition. Table 12 summarizes the prediction controls and metric references for each protocol. For GSE, the canonical evaluator forms matched-control residuals $r^{\text{match}} = x_1 - \mu_{0,m}$ using frozen cell-type pools and the observed test composition. Prediction-time controls supply model context; test-group counts determine aggregation weights and the composition-weighted control reference. Held-out perturbed expression is used only to compute the observed response for scoring.

Table 12: Control pools and evaluation samples by protocol.

Setting	Prediction controls	Metric reference	Test rows
GSE target-LOO	Cell-type grouped clean pools	Frozen matched-control pools	Frozen target rows
GSE cross-state-LOO	Held-out-state clean pools at prediction time	Frozen matched-control pools	Frozen state–target rows
K562 target-LOO	Frozen global control pool	The same frozen global pool	Frozen target rows

B.3 Response residual and evaluation

For condition κ, let $\bar{x}_{\kappa,g}$ and $\widehat{\bar{x}}_{\kappa,g}$ denote the observed and predicted expression means. Let $b_{\kappa,g}$ be the frozen, composition-weighted matched-control mean in GSE, or the global DMSO mean in K562. The evaluator computes

$$\Delta_{\kappa,g} = \bar{x}_{\kappa,g} - b_{\kappa,g}, \qquad \widehat{\Delta}_{\kappa,g} = \widehat{\bar{x}}_{\kappa,g} - b_{\kappa,g}. \tag{16}$$

Both means use the same control reference. With the condition index omitted, these vectors are also denoted by $\bar{d}$ and $\widehat{\bar{d}}$ in the metric definitions below. Observed and predicted sets may have different sizes; their means are compared without cell-to-cell matching.

B.4 Inference and aggregation

For an evaluation condition κ, let $N_{\kappa,m}$ be the supplied number of test cells in cell-type group m, and define $\mathcal{G}_\kappa = \{m : N_{\kappa,m} > 0\}$. Evaluation requires a nonempty prediction-time source pool for every group. Pool coverage is checked before scoring, and missing groups are marked as uncovered. Each source item is a control multiome record containing an RNA–ATAC pair, $\big(x_{c,\kappa,m,i}, a_{\mathrm{PROM},\kappa,m,i}\big)$; RNA-only metric-reference pools are used for scoring. In the full-coverage case, let $\mathcal{C}_{\kappa,m} = \{(x_{c,\kappa,m,i}, a_{\mathrm{PROM},\kappa,m,i}) : i \in \mathcal{I}_{\kappa,m}\}$ and set

$$\nu_{\kappa,m} = \frac{N_{\kappa,m}}{\sum_{m'} N_{\kappa,m'}}. \tag{17}$$

The benchmark supplies these observed group counts as composition metadata. Prediction is conditional on this composition and does not include estimating perturbation-induced changes in cell-type abundance. Alternative aggregation rules use uniform weights $\nu^{\mathrm{unif}}_{\kappa,m} = 1/|\mathcal{G}_\kappa|$ or group frequencies estimated from training controls.

In the GSE target-LOO condition-mean readout, the deterministic model produces a response $\widehat{\Delta}_{\kappa,m,i}$ for each query control. Responses are aggregated before adding the frozen matched-control baseline:

$$\widehat{\Delta}_\kappa = \sum_{m \in \mathcal{G}_\kappa} \nu_{\kappa,m} \frac{1}{|\mathcal{I}_{\kappa,m}|} \sum_{i \in \mathcal{I}_{\kappa,m}} \widehat{\Delta}_{\kappa,m,i}, \qquad \widehat{\mu}^{\mathrm{pert}}_\kappa = b_\kappa + \widehat{\Delta}_\kappa. \tag{18}$$

GSE target-LOO uses matched clean groups, GSE cross-state uses held-out-state source controls under transfer mode, and K562 uses one global source pool. Metric-reference rows remain fixed. Appendix A.1 compares the three response-weighting rules with the same baseline and test cells and distinguishes prediction exports from observed cell counts.

B.5 Metric definitions

The evaluator first scores each non-empty condition and then averages over conditions. Main comparison tables report seven metrics; ablations use selected subsets. Auxiliary diagnostics are defined below. Higher is better for correlations, sign accuracy, PDS, and cosine; lower is better for error and transport distances.

Mean-response metrics. For a condition with G canonical genes and matched-control mean b, let $\bar{x}$ and $\widehat{\bar{x}}$ be the observed and predicted expression means. Their mean responses are $\bar{d} = \bar{x} - b$

and $\overline{\widehat{d}} = \overline{\widehat{x}} - b$, consistent with Equation 16. Pearson$_\Delta$ compares these two vectors across genes after averaging cells within the condition. For auxiliary sign accuracy, define $a_g = \mathbf{1}[\bar{d}_g > 0]$ and $\widehat{a}_g = \mathbf{1}[\overline{\widehat{d}}_g > 0]$:

$$\text{Pearson}_\Delta = \text{Pearson}_g(\bar{d}, \overline{\widehat{d}}), \tag{19}$$

$$\text{Pearson}_{\Delta,\text{DE}} = \text{Pearson}_{g \in S_{20}}(\bar{d}_g, \overline{\widehat{d}}_g), \quad S_{20} = \text{Top20}_g\, |\bar{d}_g|, \tag{20}$$

$$\text{acc}_{\text{all}} = G^{-1} \sum_g \mathbf{1}[a_g = \widehat{a}_g], \tag{21}$$

$$\text{acc}_{\text{de}} = |S_{20}|^{-1} \sum_{g \in S_{20}} \mathbf{1}[a_g = \widehat{a}_g], \tag{22}$$

$$L2_{\text{mean}} = \|\bar{x} - \overline{\widehat{x}}\|_2, \tag{23}$$

$$\text{MSE} = G^{-1} \sum_g (\bar{x}_g - \overline{\widehat{x}}_g)^2, \qquad \text{MAE} = G^{-1} \sum_g |\bar{x}_g - \overline{\widehat{x}}_g|. \tag{24}$$

Pearson$_\Delta$ and Pearson$_{\Delta,\text{DE}}$ use these mean-response definitions throughout the paper. The DE subset is selected from the observed mean response and is shared by all methods for that condition. The auxiliary sign-accuracy implementation treats zero as the non-positive sign class.

Auxiliary differential and identification metrics. For each condition, `DE_Spearman_Sig` first applies a Wilcoxon rank-sum test between perturbed and matched-control cells for every gene. The implemented Bonferroni criterion retains genes with $pG < 0.05$; a condition with fewer than five retained genes contributes NaN. On the retained genes, Spearman correlation compares the observed and predicted log-ratio scores $\log_2((\bar{x}_g + \epsilon)/(b_g + \epsilon))$ and $\log_2((\overline{\widehat{x}}_g + \epsilon)/(b_g + \epsilon))$. These scores use the evaluator's normalized expression scale, with $\epsilon > 0$ for numerical stability.

For perturbation identity, `PDS` ranks the L1 distance between a predicted residual mean and every true condition residual mean, excludes the target gene from the distance, and converts the rank to

$$\text{PDS} = 1 - \frac{\text{rank} - 1}{N - 1}, \tag{25}$$

where N is the number of conditions in the setting.

Transport and directional metrics. Observed and exported predicted residual rows are projected into one PCA space fitted on all evaluable true residual cells in the setting (at most 50 dimensions, random state 0). Each condition contributes at most 64 cells to transport calculations. `W1_Sinkhorn` uses uniform weights, Euclidean PCA cost, and Sinkhorn regularization $\text{reg} = 1.0$. The resulting regularized transport cost is reported as W1 and evaluates the exported output representation specified in Appendix B.6. `Discriminative_Cosine` measures response direction as the cosine similarity between the gene-space mean responses $\bar{d}$ and $\overline{\widehat{d}}$, without PCA projection.

B.6 Comparator implementations and prediction interfaces

All comparators receive the same canonical condition identifiers and gene adapter. RNA and promoter-ATAC coordinates are mapped before scoring; extra genes are dropped and missing canonical genes are handled by the method's fixed adapter while the evaluator dimension remains fixed.

Comparators use method-specific preprocessing and optimization schedules from their public implementations or run configurations, with adaptations specified below. Fold-specific fits are independent. Released foundation checkpoints remain frozen, and downstream adapters are trained on fit-fold expression.

Prediction outputs and scoring. All point-estimation metrics use condition means. GEARS, the frozen-foundation adapters, CPA, and the cross-state CellFlow+ATAC export provide condition means; for W1, these means are repeated to the required number of scoring rows. For methods that export cell samples, W1 uses those samples, while point-estimation metrics use their means. Thus, W1 compares method-specific output representations, whereas the primary metrics compare condition means.

Control. *Control* uses unperturbed expression without a learned perturbation adjustment. Its reported scores provide an empirical control-reference comparison and are excluded from best/second-best learned-model highlighting.

GEARS. GEARS (Roohani et al., 2024) uses its published gene-relation graph and perturbation MLP to predict post-perturbation expression. We followed the official optimizer, preprocessing, and checkpoint rule with the published architecture, and mapped the result to the canonical panel.

CellOracle and GraphVCI. CellOracle (Kamimoto et al., 2023) uses the published regulatory-network and vector-field simulation workflow. GraphVCI (Wu et al., 2022) uses its graph-based variational causal-inference predictor. We used the public implementations with their default network construction, hidden dimensions, and regularization, and fixed graph resources before each split. Both methods produce condition-level gene responses after the supplied gene mapping, with graph construction restricted to fit-fold resources.

scGPT. scGPT (Cui et al., 2024) is evaluated with its native `TransformerGenerator` perturbation decoder, following the official repository protocol, default preprocessing, and checkpoint rule. Generated outputs are mapped to the canonical RNA panel before residualization.

scFoundation, GeneCompass, CellFM, and CellVQ. scFoundation (Hao et al., 2024), GeneCompass (Yang et al., 2024), CellFM (Zeng et al., 2025), and CellVQ (Wang et al., 2026) provide frozen gene representations. A canonical GEARS-style downstream adapter converts each representation to a condition mean, followed by the same gene mapping and residual evaluator. We trained the downstream adapter for 15 epochs for scFoundation (learning rate 10^{-2}), 15 epochs for CellFM (learning rate 5×10^{-3}), and 5 epochs for CellVQ (learning rate 5×10^{-4}). GeneCompass uses the same frozen-embedding bridge with the optimizer settings of its public GEARS integration.

CPA. CPA (Lotfollahi et al., 2023) uses its context-aware perturbation autoencoder and latent intervention predictor, trained with the default architecture and optimization settings in the public implementation, then decodes the predicted expression into the canonical panel. The exported prediction is the decoded condition mean.

STATE. STATE (Adduri et al., 2026) uses its published latent transition model and basal-sampled cell-set prediction interface. We trained it from scratch with the public implementation because the released transition checkpoints target different dosage scenarios. The generated set is mapped to the canonical panel. GSE cross-state-LOO reports STATE as unavailable under the frozen export.

CellFlow, scDFM, PerturbDiff, and DeSCOPE. CellFlow (Klein et al., 2025) is evaluated with the reference flow-matching implementation, including its 50-dimensional PCA latent and ODE integration. scDFM (Yu et al., 2026), PerturbDiff (Yuan et al., 2026), and DeSCOPE (Wu et al., 2026) use their published inference interfaces and released checkpoints/configurations, with adaptation restricted to the fit fold. These methods export generated cell samples, which are mapped to the canonical panel.

CellFlow+ATAC ablation. The ATAC-augmented CellFlow control applies train-only PCA_{50} to the RNA input and train-only SVD_{64} to the gene-aligned ATAC_PROM input, concatenates the two representations, and passes the result to the native CellFlow predictor. This design tests direct ATAC concatenation while retaining native CellFlow response generation. The GSE cross-state export contains the condition-level prediction.

C Method and Implementation Details

C.1 Prior feature interface and provenance

The prior features comprise frozen ESM-2 protein representations (Lin et al., 2023), frozen CLM-X RNA-token representations (Li et al., 2026), TSS-centered sequence features from GENCODE/RefTSS annotations on GRCh38 (Frankish et al., 2021; Abugessaisa et al., 2019; Schneider

et al., 2017), and locus descriptors derived from the dataset guide library, ENCODE cCRE annotations, and lightweight JASPAR proxies (Martin-Rufino et al., 2025; Moore et al., 2020; Castro-Mondragon et al., 2022). Missing descriptors are zero-filled, and their availability is recorded for audit. If target p has $m_p > 0$ guide descriptors $\{\psi_{p,j}\}$, the pooled guide feature is

$$\bar{\psi}_p = \frac{1}{m_p} \sum_{j=1}^{m_p} \psi_{p,j}; \tag{26}$$

otherwise the guide block is zero. Target and state labels define the splits. Control ATAC state summaries use a control-only TF–IDF/LSI transform (Stuart et al., 2021). Any control-derived accessibility summary or response-normalization quantity is recomputed within the active training fold, including in held-out-state transfer. Only frozen embeddings and static annotations specified before splitting are shared across folds.

The target-locus condition is

$$\mathbf{p}_i = \mathrm{concat}\left(e_p^{\mathrm{ESM2}}, e_p^{\mathrm{CLMX}}, f_p^{\mathrm{seq}}, f_p^{\mathrm{reg}}\right), \tag{27}$$

where the two target embeddings, 96-dimensional sequence summary, and 128-dimensional regulatory summary enter the target block in Equation 2. The stored 265-coordinate regulatory descriptor includes repeated summary and composite coordinates and is reduced as a single feature bank. Static locus annotations and fold-derived control summaries remain separate, and the input hashes and projection scope are recorded in the release audit. The resulting pooled and state-aware flow conditions have 576 and 736 dimensions, respectively.

Each external-prior ablation in Table 7 replaces the named ESM-2, CLM-X, TSS-sequence, or regulatory block with zeros when constructing c_i, retaining the same network capacity.

The CGRR source bank is rebuilt for each fold. Its projection map, ATAC prototypes P_s, control means μ_s^0, and control-state offsets B_s are rebuilt from training-fold controls. In Target-LOO, target/locus features retrieve permitted training perturbations whose observed compositions define the query mixture. In Cross-State, the clean query ATAC selects among the fixed prototypes. Equation 7 combines the control-state offsets using the weights returned by the route for that protocol. If a protocol specifies a fixed support or feature-distance gate, the gate is applied before normalizing the routing weights; the source bank is unchanged. Training mixtures use the observed perturbation composition across source groups, with control frequencies as a fallback when support is insufficient. These observed weights define the source-adjusted training endpoint. At inference, Equation 7 instead uses query-dependent weights from target/locus similarity for Target-LOO or clean-control ATAC similarity for Cross-State. Training and inference therefore use different weighting rules. Cross-State reuses the ATAC weights to combine state-centered responses $A_{p,s}$; it does not add the routed source offset to the final prediction (Equation 13).

C.2 Flow target and numerical details

For the pooled source-on candidate, CGRR first constructs d_i^{source} using training-fold controls. Writing T_ϕ for the complete PCA-anchored transport operator, the flow target is the source-adjusted response and its decoded displacement is added back to the routed source:

$$\widetilde{\Delta}_i = \Delta_i - d_i^{\mathrm{source}}, \qquad a_i = T_\phi(\widetilde{\Delta}_i \mid c_i), \qquad \Delta_i^{\mathrm{base}} = d_i^{\mathrm{source}} + a_i. \tag{28}$$

The flow uses the training-control mean as the PCA anchor:

$$z_0 = E_{\mathrm{PCA}}(\bar{b}_{\mathrm{tr}}), \qquad z_{1,i} = E_{\mathrm{PCA}}(\bar{b}_{\mathrm{tr}} + \widetilde{\Delta}_i). \tag{29}$$

The linear conditional-flow path is

$$z_{t,i} = z_0 + t(z_{1,i} - z_0), \qquad u_i^* = z_{1,i} - z_0, \qquad t \sim \mathcal{U}(0,1), \tag{30}$$

and optimization minimizes the squared velocity error in Equation 12. The velocity field is conditioned on the pooled or state-aware target–locus condition. CGRR uses target/locus or clean-control ATAC queries according to the protocol. At test time, a fixed-step solver integrates from the deterministic control-RNA anchor. For K Euler steps, $h = 1/K$ and

$$z^{(k+1)} = z^{(k)} + h\, v_\phi(z^{(k)}, kh, c_i), \qquad z^{(0)} = z_0, \quad k = 0, \ldots, K-1, \tag{31}$$

Each step uses one velocity evaluation, so K is the NFE, with default $K = 8$. The decoded displacement and stabilized source-on candidate are

$$a_i = D_{\mathrm{PCA}}(z^{(K)}) - D_{\mathrm{PCA}}(z_0), \qquad \widehat{\Delta}_i = \mathcal{C}(\Delta_i^{\mathrm{base}}). \tag{32}$$

The transport flow applies the configured mean blend $\rho = 0.18$ to its decoded displacement; the state-expanded branch uses state-centered rows and the same blend. The reliability gate is fitted from training rows and evaluated without held-out perturbation responses. Writing the affine PCA reconstruction as

$$D_{\mathrm{PCA}}(z) = \mu_{\mathrm{PCA}} + Bz, \qquad B \in \mathbb{R}^{G\times 128}, \qquad a_i = B\big(z^{(K)} - z_0\big), \tag{33}$$

shows that the PCA offset cancels in the decoded displacement. The candidate preserving operator then acts on the combined source and flow base. Let $\mathcal{M}_{\mathrm{tr}}$ be the union of candidate genes and genes protected by training-fold DE support, and let $\bar{\Delta}_{\mathrm{tr}}$ be the mean training response. The canonical candidate-preserving operator is

$$\widehat{\Delta}_{i,g} = \mathcal{C}(\Delta_i^{\mathrm{base}})_g = \begin{cases} \Delta_{i,g}^{\mathrm{base}}, & g \in \mathcal{M}_{\mathrm{tr}}, \\ \alpha\Delta_{i,g}^{\mathrm{base}} + (1-\alpha)\bar{\Delta}_{\mathrm{tr},g}, & g \notin \mathcal{M}_{\mathrm{tr}}, \end{cases} \qquad \alpha = 0.10. \tag{34}$$

The operator uses no held-out response values. This source-on candidate is one input to the Target-LOO reliability gate. The state-centered branch instead uses $r_{i,s}$ from Equation 10 as its flow target, without source addback. Protocol-specific reconstruction follows Equation 13, adding the matched-control baseline once after candidate selection or ATAC-weighted state transfer. The velocity-network architecture and default training configuration are specified in Appendix C.3 and Table 13.

C.3 Canonical implementation and training configuration

Configuration and fold construction. Table 13 gives the primary ChromaPert configuration for the GSE protocols. Each fold holds out one target or state and fits the model using the remaining targets and permitted controls. PCA is fitted once on outer-training control RNA and then fixed. Target embeddings are frozen; condition normalization, source summaries, and CGRR routes are fold-specific. Preprocessing and controls follow Appendices B.1–B.2.

Evaluation-setting instantiation. ChromaPert uses the same pooled-reference decomposition and CGRR source bank in every setting. A conditional transport flow is fit on source-adjusted deltas, a state-expanded transport branch is available when state summaries are used, and Target-LOO applies a train-only reliability gate to the decoded candidate bases; Cross-State uses the statebase branch directly. Unseen-target settings use the target/locus context route to infer a training-state mixture. Held-out-state transfer uses clean-control ATAC to match the query to the training prototypes. When population composition is supplied, state-specific predictions are aggregated with those counts; state-transfer evaluation returns the selected state prediction.

Velocity network and optimization. The velocity MLP concatenates z_t, t, and the pooled or state-aware condition c (576 or 736 dimensions, respectively). LayerNorm precedes two width-288 hidden linear layers, each with GELU and dropout, followed by a linear 128-dimensional velocity output. AdamW minimizes Equation 12 for the setting-specific epoch budgets at a constant learning rate. Direction/norm and supervised ATAC loss weights are zero. Only the transport velocity MLPs are trained; PCA and the source bank remain fixed.

Primary settings. Table 13 specifies the primary configuration. We report temperature and blend sensitivity (Tables 4c and 10) and three-seed variation (Appendix A.3) separately. These analyses characterize performance under the tested configurations and the selected primary setting.

D Discussion and Limitations

ChromaPert connects perturbation targets to cellular regulatory context through chromatin-aware conditioning and source-routed conditional transport flow modeling. Across the three evaluation settings, it achieves the highest all-gene response correlation across the benchmark methods. Ablations

show the complementary value of molecular target priors and regulatory context, while biological analyses show that predictions retain chromatin-associated and state-resolved response structure.

The main limitation is the requirement for paired scRNA-seq and scATAC-seq control profiles with a compatible promoter-accessibility mapping, restricting direct application when matched multimodal controls are unavailable. The framework predicts post-perturbation expression, with population readouts conditional on supplied state proportions. Future work will extend ChromaPert to unpaired or single-modality inputs and incorporate distal regulatory information.

Table 13: Default configuration for ChromaPert on GSE target-LOO.

Parameter	**Default**
Data and representations	
Held-out targets / canonical RNA genes	19 / 2,013
Random seed	114
ESM-2 / CLM-X dimensions	64 / 64
TSS-sequence / regulatory dimensions	96 / 128
Target–locus feature dimension	352
Static locus block dimension	224
Control RNA / ATAC state dimensions	96 / 64
Pooled / state-aware flow condition	576 / 736
PCA latent / velocity-output dimensions	128 / 128
Velocity architecture and optimization	
Hidden layers / width	2 / 288
Input normalization / activation	LayerNorm / GELU
Dropout probability (flow)	0.11
Optimizer	AdamW
Training epochs (flow)	360
Batch size	64
Learning rate	8×10^{-4}
Weight decay	10^{-4}
Learning-rate schedule	Constant (no scheduler)
Routing, integration, and evaluation	
CGRR feature temperature / ATAC temperature	0.20 / 0.15
ATAC query parameter for Cross-State (`atac_top_k`)	50
Transport-mean blend	0.18
Candidate-preserving prior α	0.10
ODE solver / function evaluations	Euler / 8
Condition-response aggregation	Official/test composition
Genes for Pearson$_{\Delta,\mathrm{DE}}$	Top 20 absolute mean responses