COHERE: COHERENT RESPONSE TRANSFER FOR CROSS-CONTEXT PERTURBATION PREDICTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Predicting perturbation responses across cellular contexts requires capturing more than average transcriptional changes. Populations with similar means can differ in expressing fractions and positive-expression distributions, yielding distinct differential-expression (DE) evidence. We introduce COHERE, a framework that jointly transfers response means and marginal shape while enforcing mean–shape compatibility. To predict unobserved perturbation–context responses, COHERE uses target controls and responses to other training perturbations in the target context to adapt source mean changes, expressing fractions, and positive-expression quantiles. A shared low-rank transformation updates these statistics while preserving source components outside the response subspace. Constrained generation constructs finite, nonnegative cell populations that preserve feasible predicted gene-wise means. Across six genetic, chemical, and cytokine perturbation benchmarks, COHERE demonstrates broad performance advantages over evaluated baselines in mean-response prediction, perturbation discrimination, and DE recovery. Under the benchmark DE scoring protocol, learned shape improves the area under the precision–recall curve (AUPRC) at fixed predicted means on all six datasets, including a relative 32.5% gain on Nadig. Matched retraining shows that preserving source components outside the response subspace improves DE recovery and marginal fidelity on PBMC and Nadig. These findings support coherent transfer of marginal response shape as a complement to accurate mean prediction.

1 INTRODUCTION

Predicting how cells respond to genetic, chemical, or cytokine perturbations is central to identifying gene function, understanding disease mechanisms, and prioritizing therapeutic interventions. Perturb-seq and sci-Plex measure genetic and chemical perturbation responses at single-cell resolution (Dixit et al., 2016; Srivatsan et al., 2020), with genome-scale screens revealing transcriptional effects and cellular heterogeneity (Replogle et al., 2022). Because conventional single-cell RNA sequencing destroys assayed cells, these screens yield unpaired control and perturbed populations rather than before-and-after measurements of the same cells (Bunne et al., 2023). *Perturbation response prediction* infers responses for unmeasured perturbation–context combinations.

Perturbation responses are context-dependent: the same intervention can elicit different transcriptional changes across cell types, donors, or cell lines. This motivates *cross-context perturbation prediction*, which combines responses from other contexts with information about the recipient environment (Lotfollahi et al., 2019; Adduri et al., 2026; Yuan et al., 2026). We study partial target-context coverage, where target basal cells and responses to a panel of other training perturbations are available, but the query response in that context is not. As Figure 1a illustrates, source responses to the query perturbation provide intervention-specific evidence, while the target training panel guides how that evidence is adapted to the recipient context.

Differential-expression (DE) patterns help connect perturbations to gene function and regulatory mechanisms (Morris et al., 2023; Nadig et al., 2025). Accurate mean-response prediction alone does not ensure recovery of these DE patterns: similar means can arise from different expressing fractions and positive-expression distributions. In Figure 1b, two response populations have equal means, yet one lies on both sides of a shared control and the other lies entirely above it. Wilcoxon-based DE

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

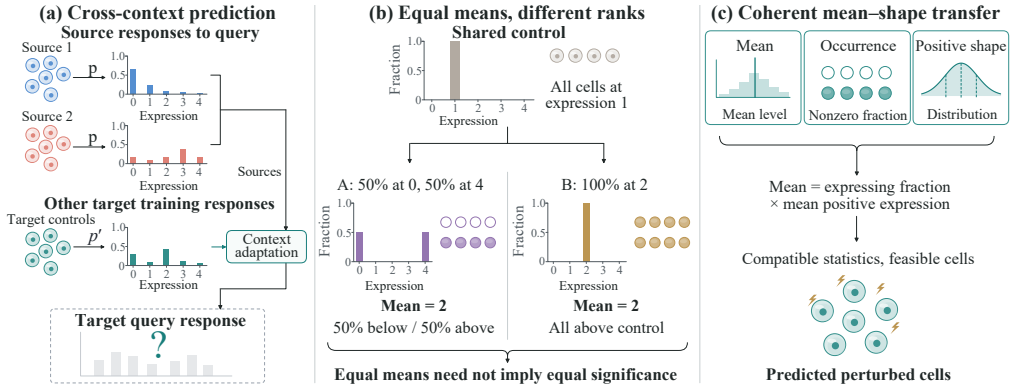


Figure 1: **Motivation for coherent cross-context response transfer.** (a) Source responses to query perturbation p provide intervention-specific evidence. Target basal cells and responses to $p' \neq p$ guide adaptation, while the target response to p is unobserved. (b) Responses A and B have mean 2 but different ranks against the shared control, so rank-test significance can differ. (c) The mean equals the expressing fraction times the mean positive expression. COHERE generates finite, nonnegative cell populations that reconcile transferred means and marginal shapes.

evaluation compares expression ranks between perturbed and control cells; equal means therefore need not yield equal significance or significance-based gene rankings (Adduri et al., 2026).

Existing methods transfer perturbation effects through latent shifts and disentangled representations (Lotfollahi et al., 2019; 2023; Piran et al., 2024), or model response populations using transport, alignment, flows, and diffusion (Bunne et al., 2023; Adduri et al., 2026; Klein et al., 2025; Yuan et al., 2026). Doloris separates sparsity masking from expression generation (Chi et al., 2026). These approaches go beyond averages, but distributional prediction alone does not reveal what shape adaptation contributes beyond mean transfer. We address this question by explicitly adapting mean and marginal shape as coupled response statistics. A gene's population mean equals its expressing fraction times its mean positive expression (Figure 1c), so independently specified statistics need not be jointly realizable. We reconcile the adapted statistics into finite populations that conserve feasible predicted means, allowing us to assess shape contributions to DE recovery at fixed means.

We propose COHERE, a framework connecting empirical response transfer to constrained cell generation. Using the target training panel, it jointly adapts source mean changes, expressing fractions, and positive-expression quantiles through a shared low-rank transformation that retains source components outside the response subspace. An auxiliary residual based on perturbation identity features supplements mean prediction. Constrained realization then reconciles the adapted statistics into nonnegative populations that conserve feasible predicted means.

Across six genetic, chemical, and cytokine perturbation benchmarks, COHERE performs strongly in mean-response prediction, perturbation discrimination, and DE recovery (Table 1 and Figure 3). Under the benchmark DE scoring protocol, learned shape improves AUPRC at fixed predicted means on all six datasets, from 0.251 to 0.332 on Nadig. These gains arise primarily from expression occurrence, while positive-expression variation improves marginal fidelity on PBMC and Nadig. Matched retraining shows that preserving source components outside the response subspace improves DE recovery and marginal fidelity on both datasets.

We contribute (i) source-preserving joint adaptation of empirical mean responses and marginal shape, (ii) finite-cell realization enforcing mean–shape compatibility and conserving feasible predicted means, and (iii) fixed-mean interventions and matched retraining that distinguish the effects of shape adaptation and source preservation on DE recovery and marginal fidelity.

2 Related Work

Cross-context perturbation prediction. scGen estimates a perturbation shift in a variational latent space and transfers it to another cell population (Lotfollahi et al., 2019). CPA combines basal representations with perturbation, dose, and covariate embeddings, while biolord disentangles anno-

tated attributes from unobserved cell-specific factors to support prediction of missing combinations (Lotfollahi et al., 2023; Piran et al., 2024). scPILOT combines a generative representation with query-conditioned latent optimal transport to transfer observed perturbation responses across contexts (Wang et al., 2026). COHERE instead adapts empirical response statistics, learning a target-context transformation from the training panel and applying it to source evidence for the query perturbation.

Population and occurrence-aware modeling. CellOT fits perturbation-specific neural optimal-transport maps and transfers learned responses to unseen patient samples (Bunne et al., 2023). STATE uses a distributional prediction objective (Adduri et al., 2026). CellFlow conditions flow matching on experimental information (Klein et al., 2025), while scDFM penalizes maximum mean discrepancy of predicted endpoint distributions (Yu et al., 2026). PerturbDiff models response distributions through functional diffusion of kernel embeddings (Yuan et al., 2026). Doloris separates zero from positive expression with a sparsity mask (Chi et al., 2026). COHERE jointly adapts mean and marginal-shape statistics and realizes finite, nonnegative populations conserving feasible predicted gene-wise means.

Response and biological representations. scDEBART pretrains response representations on log fold-change profiles (Sung & Kim, 2026). Molina & Zhang (2026) separate global, perturbation-specific, cell-line-specific, and interaction components to align biological information with response structure. Complementary approaches encode biological priors through molecular features in chemCPA (Hetzel et al., 2022), coexpression and Gene Ontology graphs in GEARS (Roohani et al., 2024), and multiple knowledge graphs in TxPert (Wenkel et al., 2026). At the cell-state level, scGPT provides pretrained representations (Cui et al., 2024). COHERE uses empirical response descriptors for transfer, with frozen perturbation-identity features supplying a bounded auxiliary mean residual.

Evaluation of perturbation predictions. Ahlmann-Eltze et al. (2025) show that several deep models do not outperform simple baselines in the single- and double-gene prediction settings they study, motivating strong baseline comparisons. PerturBench standardizes tasks and evaluates fit and perturbation ranking (Wu et al., 2025), while Cell-Eval measures response fidelity, perturbation discrimination, and differential expression (Adduri et al., 2026). Our fixed-mean factorial interventions separate occurrence and positive-quantile contributions to DE recovery.

3 Problem Setup

We study *cross-context prediction under partial target-context coverage* using unpaired control populations X_c^0 and perturbed populations X_{cp} in processed log1p expression over G genes. Here c indexes cellular contexts and p indexes perturbation identities. The set $\mathcal{I}_{\mathrm{tr}}$ contains the context–perturbation combinations available for fitting.

For target context c, define the fixed training panel and source set

$$\mathcal{P}_c^{\mathrm{tr}} = \{p : (c,p) \in \mathcal{I}_{\mathrm{tr}}\},$$
$$\mathcal{S}_{c,p} = \{(X_s^0, X_{sp}) : s \in \mathcal{C}_{\mathrm{src}}(c),\ (s,p) \in \mathcal{I}_{\mathrm{tr}}\},$$

where $\mathcal{C}_{\mathrm{src}}(c)$ contains eligible source contexts $s \neq c$. Available target-panel responses and their source sets supervise context adaptation.

In the query-disjoint formulation, a test query $p_* \notin \mathcal{P}_c^{\mathrm{tr}}$ has no target response available for fitting, selection, or prediction, including at other doses or in other technical batches. The same identity may nevertheless occur in the source training observations. Given X_c^0, $\mathcal{S}_{c,p_*}$ (possibly empty), and optional frozen identity features e_{p_*}, the fitted predictor generates $\widehat{X}_{cp_*} \in \mathbb{R}_{\geq 0}^{N \times G}$ for a requested cell count N. Benchmark-specific source rules and split exceptions are detailed in Appendix A.

4 Coherent Response Transfer

COHERE couples source-preserving response adaptation with mean-compatible cell realization (Figure 2). Using the target training panel, it jointly adapts source mean changes, expressing fractions, and positive-expression quantiles through shared low-rank mixing, with a bounded identity residual supplementing the mean prediction. The adapted statistics are then reconciled into a finite, nonnegative cell population that conserves the feasible predicted gene-wise means.

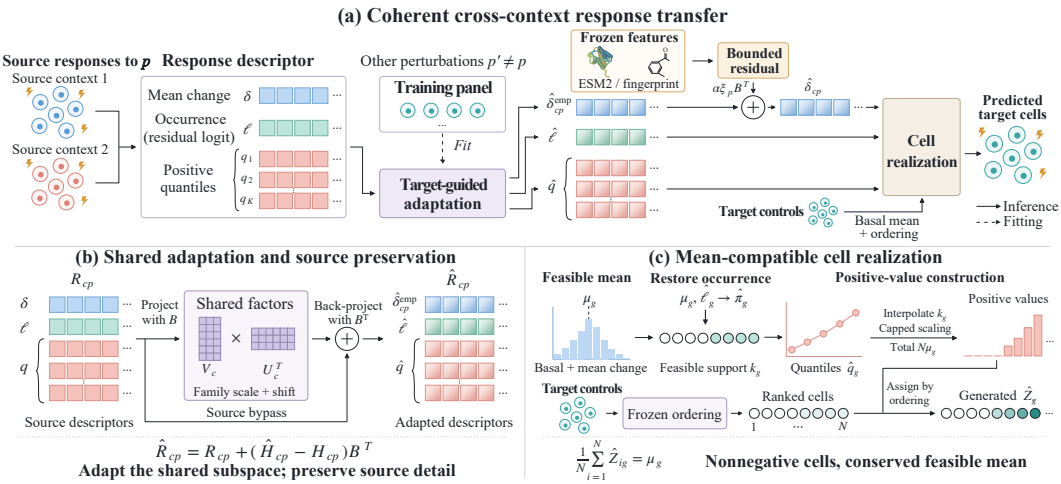


Figure 2: **Overview of COHERE.** (a) Source-response statistics are adapted using the target training panel, supplemented by an identity-based mean residual, and realized as a cell population. (b) Shared low-rank mixing and family-specific affine parameters update response-basis coordinates. Coordinate changes are back-projected and added to source descriptors, preserving components outside the basis. (c) Feasible means and restored expression probabilities determine nonzero support. Positive values interpolated from quantiles undergo capped scaling and assignment by control-derived cell orderings.

4.1 Structured Empirical Response

We compute expression statistics in model coordinates $z = L/\gamma$, where L is processed log1p expression and $\gamma > 0$ is a fixed scale. Throughout, expression vectors are rows, with nonlinearities, division, and clipping applied elementwise. For an observed pair (c, p), let $m_{cp}, m_c^0 \in \mathbb{R}^{1\times G}$ denote perturbed and control means, and let $\pi_{cp} \in [0,1]^{1\times G}$ denote expressing fractions. We represent the mean response by $\delta_{cp} = m_{cp} - m_c^0$ and marginal shape by a residual occurrence logit ℓ_{cp} and positive-expression quantile rows $q_{cp,1}, \ldots, q_{cp,K}$ at levels $0 < \tau_1 < \cdots < \tau_K < 1$. When a gene has no positive observations, its empirical positive-expression quantiles are zero-coded and the occurrence channel records the absence of positive support. For genes with positive expression, $m_{cp,g} = \pi_{cp,g}\mathbb{E}[z_g \mid z_g > 0]$. Let $C > 0$ be a reference cap. If $0 \le z_g \le C$, then $\pi_{cp,g} \ge m_{cp,g}/C$, motivating an occurrence coordinate relative to the mean-to-cap ratio:

$$\begin{aligned} \rho_{cp} &= \mathrm{clip}(m_{cp}/C, 0, 1-\epsilon), \\ \ell_{cp} &= \mathrm{logit}\left[\mathrm{clip}\left(\frac{\pi_{cp}-\rho_{cp}}{\max(1-\rho_{cp},\epsilon)}, \eta, 1-\eta\right)\right]. \end{aligned} \tag{1}$$

Observed expression may exceed C. The constants $\epsilon, \eta \in (0, 1/2)$ stabilize boundary values.

For each source in $\mathcal{S}_{c,p}$, we form rows $(\delta_{sp}, \ell_{sp}, q_{sp,1}, \ldots, q_{sp,K})$ using that context's controls. Coordinate-wise Huber aggregation gives $R_{cp} \in \mathbb{R}^{(K+2)\times G}$, with $R_{cp} = 0$ encoding missing source evidence when no sources are available. A shared orthonormal basis $B \in \mathbb{R}^{G\times d}$, with $B^\top B = I_d$, is estimated by truncated randomized SVD of centered training mean responses. The projection $H_{cp} = R_{cp}B \in \mathbb{R}^{(K+2)\times d}$ contains unstandardized response-basis coordinates. We standardize its rows using a scalar median and scaled median absolute deviation (MAD) for each family, with the basis and normalizers fitted on $\mathcal{I}_{\mathrm{tr}}$ and then fixed.

4.2 Source-Preserving Context Adaptation

The recipient map aligns panel-centered source and target coordinates through shared low-rank mixing and family-specific affine adjustments. For each recipient and descriptor row, we compute source and target panel centers from the standardized coordinates of the aggregated source descriptors and observed target statistics, respectively. Both centers are obtained by averaging over the same usable perturbations in $\mathcal{P}_c^{\mathrm{tr}}$. A training row is usable when its target-response and control statistics are

available, even if source observations are missing. Subtracting the corresponding source-panel center from each standardized row of H_{cp} gives $h \in \mathbb{R}^{1\times d}$. We use three descriptor families $f \in \{\delta, \pi, q\}$, representing mean response, occurrence encoded by ℓ, and positive quantiles, respectively. The recipient map for family f is

$$T_{c,f}(h) = \kappa_{c,f}\left[h + (hV_c)U_c^\top\right] + a_{c,f}\mathbf{1}, \tag{2}$$

where $U_c, V_c \in \mathbb{R}^{d\times r_c}$ are shared across families, while $\kappa_{c,f} > 0$ and $a_{c,f} \in \mathbb{R}$ are family-specific scalars. All K quantile rows share the same family parameters. After applying the map, we add the corresponding target-panel centers and reverse standardization to obtain $\widehat{H}_{cp}$ in unstandardized basis coordinates. For $n_c \geq n_{\min}$ usable training perturbations, we set $r_c = \min(r_{\max}, n_c - 1, d)$, where $r_{\max}$ caps the mixing rank and $n_{\min}$ is the minimum panel size for adaptation. Panels with fewer than $n_{\min}$ usable perturbations bypass the entire map, including centering, so $\widehat{H}_{cp} = H_{cp}$.

We back-project only the coordinate change and add it to the source descriptor:

$$\widehat{R}_{cp} = R_{cp} + (\widehat{H}_{cp} - H_{cp})B^\top. \tag{3}$$

Consequently, $\widehat{R}_{cp}(I_G - BB^\top) = R_{cp}(I_G - BB^\top)$: source components outside the response subspace are preserved at the descriptor stage. The first row of $\widehat{R}_{cp}$ is the empirical mean response $\widehat{\delta}^{\text{emp}}_{cp}$. The remaining rows are $\widehat{\ell}_{cp}$ and $\widehat{q}_{cp,1}, \ldots, \widehat{q}_{cp,K}$.

The recipient map is supervised by observed statistics of the target training panel. Let $\mathcal{L}_\delta$, $\mathcal{L}_\pi$, and $\mathcal{L}_q$ denote the mean, occurrence, and positive-quantile losses, and ν_f their fixed values under unadapted source transfer. We minimize

$$\mathcal{L}_{\text{emp}} = \frac{1}{3}\sum_{f\in\{\delta,\pi,q\}} \frac{\mathcal{L}_f}{\max(\nu_f, \epsilon_{\text{loss}})}. \tag{4}$$

The constant $\epsilon_{\text{loss}} > 0$ stabilizes the fixed reference normalization. The mean term averages SmoothL1 and cosine losses. For the occurrence term, predicted and target probabilities are restored separately from their respective feasible means and occurrence logits (Equation 6), without identity correction. Binary cross-entropy is then computed between these probabilities. Before cell realization, the positive-quantile term applies pinball loss at $\{\tau_k\}$ between monotone nonnegative predicted knots and the empirical positive-expression quantiles of the observed target training population. Appendix B specifies the boundary encodings and reductions.

4.3 Bounded Identity Residual

A residual branch supplements empirical transfer with frozen perturbation features $e_p \in \mathbb{R}^{1\times d_e}$: ESM2 embeddings for protein identities (Lin et al., 2023) and molecular fingerprints for chemicals. An MLP with support and sign heads h_{supp} and h_{sign} produces a correction in basis coordinates:

$$\begin{aligned} \xi_p &= \boldsymbol{\sigma}_r \odot \text{sigmoid}\big(h_{\text{supp}}(e_p)\big) \odot \tanh\big(h_{\text{sign}}(e_p)\big), \\ \widehat{\delta}_{cp} &= \widehat{\delta}^{\text{emp}}_{cp} + \alpha\xi_p B^\top. \end{aligned} \tag{5}$$

Here $\boldsymbol{\sigma}_r \in \mathbb{R}^{1\times d}_{>0}$ is a training residual scale and $\alpha = \text{sigmoid}(\beta)$ is one learned shrinkage scalar per dataset. The gates bound each correction by $|\xi_{p,j}| \leq \sigma_{r,j}$. The identity correction is shared across recipient contexts within each dataset. The branch directly changes only the mean row, with possible effects on realized occurrence and positive values through the feasible mean.

With the empirical model frozen, we fit the residual head at unit correction amplitude using identity-held-out predictions from auxiliary empirical fits on the training partition. These auxiliary fits share the fixed training representation. We construct residual targets r by subtracting predicted target mean responses from observed ones and projecting the differences into the fixed response basis. For coordinate j, the support and sign labels are $y_j^{\text{supp}} = \mathbf{1}\{|r_j| \geq \sigma_{r,j}\}$ and $y_j^{\text{sign}} = \mathbf{1}\{r_j > 0\}$. The training objective combines corrected-response SmoothL1/cosine loss with support and sign cross-entropy terms, weighted by $\lambda_{\text{supp}}, \lambda_{\text{sign}} \geq 0$, respectively. Sign supervision is applied only where $y_j^{\text{supp}} = 1$. We then freeze the residual head and fit the shrinkage parameter β using identity-held-out corrected predictions from auxiliary empirical and residual-head fits on the training partition. Appendix B details the auxiliary-fit dependencies and stage-wise optimization.

4.4 MEAN-COMPATIBLE CELL REALIZATION

For a query (c, p_*) and requested cell count N, let $\bar{C} = C(1 - 2\epsilon)$ be the generation cap and $\mu = \mathrm{clip}(m_c^0 + \widehat{\delta}_{cp_*}, 0, \bar{C})$ the feasible mean. Write $\widehat{\ell} = \widehat{\ell}_{cp_*}$ and suppress query indices on the subsequent realization variables. We first restore continuous occurrence and choose an integer nonzero count for each gene:

$$\widehat{\pi} = \mu/C + (1 - \mu/C)\,\mathrm{sigmoid}(\widehat{\ell}),$$
$$k_g = \begin{cases} 0, & \mu_g = 0, \\ \min\big(N, \max\{\mathrm{round}(N\widehat{\pi}_g), \lceil N\mu_g/\bar{C} \rceil\}\big), & \mu_g > 0. \end{cases} \tag{6}$$

The lower bound ensures $k_g \bar{C} \geq N\mu_g$, so the positive cells can carry the required expression total. The realized expressing fraction is k_g/N, which may differ from the continuous prediction because of rounding and the capacity constraint.

For $\mu_g > 0$, we make the adapted positive quantiles nonnegative and monotone, then interpolate them to k_g values. We replace an all-zero sequence with a constant positive sequence and otherwise apply a positive floor, obtaining $0 < v_{g1} \leq \cdots \leq v_{gk_g}$. Capped scaling chooses $\lambda_g \geq 0$ such that

$$\widetilde{z}_{gi} = \min(\lambda_g v_{gi}, \bar{C}), \qquad \sum_{i=1}^{k_g} \widetilde{z}_{gi} = N\mu_g. \tag{7}$$

Positive v_{gi} and sufficient capacity ensure a solution. The positive mean $N\mu_g/k_g$ links expression scale to predicted mean and realized occurrence. Scaling preserves ordering but can alter quantile magnitudes and introduce ties at the cap.

To assemble cells, a frozen encoder and allocator assign gene-wise scores using only N context-matched control anchors sampled from X_c^0. Each score combines log-transformed basal expression with a learned control-derived adjustment in the response basis. The top k_g anchors receive the ordered positive values, with larger scores receiving larger values. All remaining entries, including genes with zero feasible means, are zero. The encoder and allocator are jointly pretrained on training populations using sorted-expression and smooth occupancy-proxy supervision (Appendix B). The model-coordinate population $\widehat{Z} \in [0, \bar{C}]^{N \times G}$ satisfies, up to numerical tolerance,

$$\frac{1}{N}\sum_{i=1}^{N} \widehat{Z}_{ig} = \mu_g, \qquad \widehat{X}_{cp_*} = \gamma \widehat{Z}.$$

The final equality restores log1p units for evaluation. Within each generated group, allocation combines gene-wise values across cells while preserving their realized marginals and means. Algorithm 1 summarizes prediction, with numerical and interpolation details in Appendix B.

5 EXPERIMENTS

5.1 EVALUATION PROTOCOL

Benchmarks. We evaluate PBMC cytokine responses and Replogle–Nadig CRISPRi using the fixed STATE-style partitions adopted in the PerturbDiff comparison (Replogle et al., 2022; Nadig et al., 2025). We also fit models separately on SciPlex3 (highest dose), Frangieh21, OP3, and McFaline using fixed PerturBench-derived splits (Wu et al., 2025). Evaluation includes queries without source observations. The official OP3 split includes six test queries with one training cell each. We additionally report results on the remaining 138 queries (Appendix E.2). Data sizes, gene spaces, context partitions, and eligible source contexts are given in Appendix A.1.

Baselines and selection. We compare COHERE with four grouped-mean baselines, CPA, CellFlow, scGen, scPILOT, STATE, and PerturbDiff. TxPert is additionally included where applicable to genetic perturbations. PerturbDiff uses released scratch and fine-tuned checkpoints on PBMC and Nadig and is trained from scratch on the four additional datasets. We train STATE on all six datasets. All locally trained models use target-validation checkpoint selection, with model-specific budgets

Table 1: **Primary benchmark test results.** All 13 metrics are reported separately for PBMC and Nadig. Blue shades identify the top three distinct values for each metric within each dataset before rounding; the darkest shade and bold indicate the best. Arrows indicate favorable directions, and dashes denote undefined values. PDS_1/PDS_2 use L1/L2 distance. *: Mean/perturbation copies target controls for the eight source-free Nadig queries.

Method	DEOver↑	DEPrec↑	ES↑	DirAgr↑	LFCSpear↑	AUPRC↑	AUROC↑	PDCorr↑	MSE↓	MAE↓	PDS_1↑	PDS_2↑	PDS_{cos}↑
PBMC													
Mean / perturbation	0.561	0.542	0.019	0.741	0.478	0.542	0.505	0.642	8.5e-4	0.013	0.730	0.777	0.787
Mean / cell type	0.503	0.539	-0.095	0.635	0.245	0.539	0.499	0.398	6.5e-4	0.009	0.627	0.595	0.608
Mean / batch	0.560	0.539	-0.226	0.738	0.507	0.539	0.500	0.508	5.7e-4	0.007	0.621	0.593	0.630
Mean / overall	0.555	0.539	–	0.658	0.250	0.539	0.499	0.408	1.0e-3	0.013	0.508	0.508	0.508
CPA	0.384	0.539	0.213	0.617	0.141	0.539	0.499	0.422	5.7e-4	0.010	0.632	0.606	0.661
CellFlow	0.441	0.540	-0.023	0.596	0.270	0.540	0.501	0.588	3.9e-4	0.009	0.729	0.776	0.847
scGen	0.400	0.539	-0.316	0.629	0.297	0.539	0.500	0.675	2.2e-4	0.007	0.881	0.896	0.967
scPILOT	0.355	0.539	0.008	0.450	0.135	0.539	0.499	0.299	1.7e-3	0.031	0.706	0.765	0.750
PerturbDiff scratch	0.560	0.579	0.274	0.752	0.520	0.612	0.615	0.817	1.9e-4	0.006	0.973	0.974	0.986
PerturbDiff finetuned	0.527	0.585	0.429	0.701	0.431	**0.644**	**0.659**	0.705	3.2e-4	0.008	0.941	0.959	0.971
STATE	0.511	0.564	0.015	0.774	0.536	0.594	0.600	0.826	1.5e-4	0.005	0.960	0.968	0.969
COHERE	**0.597**	**0.591**	**0.854**	**0.857**	**0.716**	0.643	0.646	**0.896**	**9.2e-5**	**0.004**	**0.985**	**0.977**	**0.991**
Nadig													
Mean / perturbation*	0.161	0.103	0.352	0.558	0.063	0.118	0.486	0.048	9.8e-2	0.204	0.579	0.597	0.606
Mean / cell type	0.179	0.097	0.334	0.758	0.617	0.103	0.381	0.412	8.5e-3	0.057	0.502	0.505	0.502
Mean / batch	0.113	0.103	0.340	0.436	-0.134	0.113	0.447	0.002	7.1e-2	0.175	0.505	0.505	0.506
Mean / overall	0.114	0.103	0.334	0.438	-0.136	0.114	0.450	-0.001	7.2e-2	0.177	0.502	0.503	0.501
CPA	0.087	0.088	0.335	0.644	0.221	0.099	0.356	0.216	1.0e-2	0.063	0.505	0.508	0.509
CellFlow	0.166	0.094	0.669	0.775	0.548	0.102	0.339	0.387	6.5e-3	0.052	0.667	0.676	0.611
scGen	0.155	0.098	0.697	0.803	0.539	0.105	0.383	0.396	7.2e-3	0.055	0.680	0.698	0.672
scPILOT	0.090	0.096	0.510	0.670	0.400	0.107	0.439	0.222	1.5e-2	0.087	0.627	0.630	0.678
TxPert	0.102	0.089	0.600	0.715	0.332	0.108	0.432	0.268	9.9e-3	0.066	0.577	0.592	0.649
PerturbDiff scratch	0.180	0.166	0.565	0.759	0.451	0.247	0.647	0.347	1.5e-2	0.081	0.691	0.696	0.578
PerturbDiff finetuned	0.211	0.188	0.758	0.782	0.534	0.295	0.685	0.375	1.5e-2	0.079	0.754	0.762	0.635
STATE	0.183	0.177	**0.788**	0.872	0.702	0.321	0.683	0.460	**5.3e-3**	**0.050**	0.801	0.820	**0.833**
COHERE	**0.249**	**0.207**	0.782	**0.886**	**0.711**	**0.332**	**0.768**	**0.478**	6.6e-3	0.056	**0.840**	**0.862**	0.746

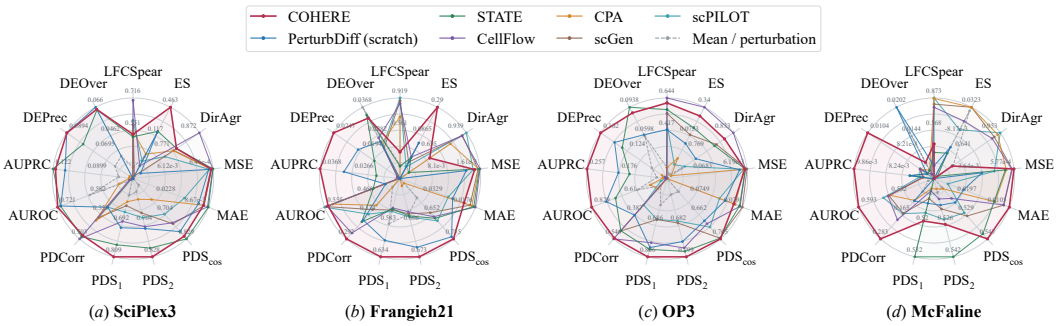
and selection criteria documented in Appendices B and H.3. For COHERE, checkpoint selection balances validation ranks equally across mean accuracy, perturbation discrimination, DE recovery, and effect-direction agreement. All methods within a dataset share the evaluation queries, populations, and gene space. Parameter counts and measured computational costs are reported in Appendix F.

Metrics. We report 13 Cell-Eval metrics. PDCorr compares control-relative means, MSE/MAE compare absolute means, PDS measures perturbation discrimination, and ES measures Spearman correlation of significant-DE gene counts. DEOver and DEPrec measure DE-list overlap and precision, while DirAgr and LFCSpear assess effect-direction and log-fold-change rank agreement. Real and predicted populations undergo separate Wilcoxon rank-sum tests against matched controls with Benjamini–Hochberg correction. AUROC and AUPRC (average precision) evaluate Cell-Eval significance scores against real-data DE labels at FDR < 0.05, with the score defined in Appendix A. PBMC DE tests pool cells by cytokine identity. Primary results aggregate by perturbation identity, while external results weight contexts equally (Appendix A). Reported 95% intervals use 2,000 paired identity-bootstrap draws conditional on the fitted models (Appendix G).

5.2 Primary And External Results

Among the methods in Table 1, COHERE achieves the best point estimates on 11 of 13 metrics in PBMC and nine in Nadig. It achieves the highest PDCorr on both, improving over STATE from 0.826 to 0.896 on PBMC and from 0.460 to 0.478 on Nadig. Mean errors are lowest for COHERE on PBMC and for STATE on Nadig. For DE recovery, fine-tuned PerturbDiff has slightly higher AUPRC and AUROC on PBMC, while COHERE leads most Nadig metrics. Its Nadig AUPRC margin over STATE is small, with a paired interval that includes zero (Appendix C.3).

Across the four additional benchmarks, COHERE achieves the highest AUPRC on Frangieh21, OP3, and McFaline among the compared methods (Figure 3). On OP3, COHERE achieves an AUPRC of



(*a*) **SciPlex3** (*b*) **Frangieh21** (*c*) **OP3** (*d*) **McFaline**

Figure 3: **Perturbation prediction on four external benchmarks.** Eight methods are shown. Metrics are min–max normalized across all methods in each benchmark table, with MSE and MAE reversed so larger radii indicate better performance. Ticks show raw values, while radii indicate relative performance within each dataset–metric pair. Full comparisons appear in Appendix C.

0.2568, compared with 0.1774 for PerturbDiff scratch, the highest-AUPRC baseline. The advantage persists on the 138-query subset excluding the six queries with one training cell each (Appendix E.2). The other metrics reveal distinct strengths: on SciPlex3, COHERE leads PDS$_1$, PDS$_2$, AUROC, and ES, while STATE has higher AUPRC and CellFlow has the highest PDCorr. On McFaline, COHERE leads PDCorr, MSE, MAE, and AUPRC (0.0099), although DE-count concordance is weak (ES $= -0.0688$). Complete results appear in Appendix C.

Direct transfer and recipient adaptation. The direct-transfer control retains source mean and shape descriptors with target-control restoration and the common realization procedure, distinguishing it from Mean/perturbation (Appendix Table 9). Relative to direct transfer, COHERE achieves higher PDCorr and lower MSE on both datasets. Benchmark AUPRC decreases from 0.72353 to 0.64345 on PBMC and increases from 0.21017 to 0.33193 on Nadig. The empirical backbone without identity correction shows the same pattern relative to direct transfer (Appendix D.2). A separate within-context PBMC analysis uses donor–cell-type-matched PBS controls and locally defined DE labels, yielding stable-score AP of 0.26676 for COHERE versus 0.23599 for direct transfer (Appendix H.2).

5.3 Mechanism and Sensitivity Analyses

Source-preserving reconstruction. To assess source preservation outside the response subspace, we retrain a subspace-only predictor using $\widehat{R}_{cp} = \widehat{H}_{cp}B^\top$ in place of Equation 3. It retains source coordinates within the basis and uses the same representation and realization procedure. All dependent adaptation and identity-residual stages are refitted with the same budgets and checkpoint-selection rule. Relative to this control, COHERE improves AUPRC by 0.11018 on PBMC and 0.12558 on Nadig (Appendix Table 9) and lowers marginal Wasserstein error on both (Table 3). On Nadig, these benefits accompany lower PDCorr and higher MSE, revealing a trade-off between preserving source detail and mean accuracy. Appendix C.1 reports the retraining protocol and paired intervals.

Marginal shape at fixed means. At the full model's feasible means μ, we independently choose source or learned occurrence-logit and positive-quantile descriptors while fixing controls, cell counts, allocator scores, and realization. These four inference-time interventions measure shape-channel effects within the fitted predictor without retraining it. The source-shape intervention combines source occurrence and quantiles with COHERE's feasible means, whereas direct transfer also retains source mean responses. Joint shape adaptation improves AUPRC on all six datasets under the benchmark DE scoring protocol (Table 2). Learned occurrence improves AUPRC under both quantile settings on all six datasets, whereas quantile effects are smaller and dataset-dependent, including negative effects on PBMC (Appendix C.8). Alternative DE rankings retain the Nadig gains, while the small PBMC joint-shape effect is sensitive to score clipping (Appendix C.3).

Marginal fidelity against distributional baselines. We compare predicted and observed gene-wise marginals using Wasserstein-1 (W1) distance in processed log1p space, including zero mass (Table 3).

Table 2: **Shape adaptation at fixed predicted means.** AUPRC after realizing source or learned occurrence-logit and positive-quantile descriptors at the same feasible means of the complete model. Controls, cell counts, ordering, and the realization procedure are fixed. Differences are learned minus source shape, computed before rounding. Paired intervals appear in Appendix Figure 6.

Shape input	PBMC	Nadig	SciPlex3	Frangieh21	OP3	McFaline
Source shape	0.64023	0.25057	0.11233	0.03242	0.23265	0.00941
Learned shape	**0.64345**	**0.33193**	**0.11979**	**0.03679**	**0.25679**	**0.00986**
ΔAUPRC	+0.00322	+0.08135	+0.00747	+0.00438	+0.02414	+0.00045

Table 3: **Marginal fidelity on primary benchmarks.** W1 uses processed log1p units; occurrence MAE is dimensionless. Perturbation identities receive equal weight; lower is better. Best values are **bold** and second-best underlined. Paired intervals and aggregation details appear in Appendix C.2.

	PBMC		Nadig	
Method	W1↓	Occ. MAE↓	W1↓	Occ. MAE↓
COHERE	**0.02945**	0.03921	0.10538	0.10053
STATE	0.03099	0.04673	**0.07386**	0.10492
PerturbDiff scratch	0.03317	**0.03193**	0.12872	0.10411
PerturbDiff finetuned	0.03518	0.04479	0.11350	**0.09671**
Subspace-only (retrained)	0.05882	0.25586	0.12857	0.14783

Occurrence MAE measures absolute error in expressing fractions. COHERE has lower W1 than STATE and both PerturbDiff variants on PBMC. On Nadig, it improves over both PerturbDiff variants, while STATE achieves lower W1. The subspace-only control has higher W1 and occurrence error than COHERE on both datasets.

Shape adaptation and within-positive variation. Relative to source shape at fixed means, learned shape reduces Wasserstein-1 distance by 3.8% on PBMC and 34.3% on Nadig, and lowers occurrence and normalized positive-quantile errors on both. At learned occurrence and fixed means, replacing source quantiles with learned quantiles further reduces Wasserstein error on both datasets, with paired intervals below zero (Appendix C.7). To isolate within-positive variation, the equal-positive control preserves feasible means and nonzero support while assigning identical positive values within each gene. This control increases Wasserstein error on both datasets relative to the full model. AUPRC decreases from 0.643 to 0.639 on PBMC but increases from 0.332 to 0.337 on Nadig (Appendix C.6). Thus, DE ranking and marginal fidelity can diverge.

Sensitivity to model design. Retraining without quantile adaptation and identity correction barely changes PBMC AUPRC and PDCorr but lowers both on Nadig (Appendix H.1). In matched fits of separate mean and shape adapters, a shape-aware basis increases test AUPRC over a mean-derived basis on both datasets but lowers PDCorr and raises MSE on Nadig (Appendix C.4).

6 Discussion and Conclusion

COHERE couples empirical mean-response transfer with marginal-shape adaptation and realizes finite cell populations that conserve feasible predicted means. Matched retraining supports the value of preserving source components for DE recovery and marginal fidelity on PBMC and Nadig. Fixed-mean interventions identify expression occurrence as the main contributor to DE-recovery gains from shape adaptation under the benchmark scoring protocol, while positive-expression variation improves marginal fidelity even when DE ranking does not improve. Mean-compatible realization provides a controlled link between prediction and analysis: it allows marginal shape to vary while keeping the predicted mean response fixed.

The present study evaluates gene-wise marginals under fixed target panels, with identity-bootstrap intervals conditional on single-seed fits. Evaluating joint cellular structure and robustness across training runs and target panels would extend this evidence. These findings motivate assessing mean accuracy, DE recovery, and marginal fidelity together when evaluating perturbation predictions.

AI USE STATEMENT

Generative AI tools were used for language editing and polishing. The authors reviewed all AI-assisted revisions and take responsibility for the final content of the submission.

ETHICS STATEMENT

This work analyzes existing public single-cell datasets and performs no new human or animal experiments. Perturbation predictions are computational hypotheses, not validated interventions or clinical recommendations. Generalization errors across cellular contexts can be consequential; the results should not be used to infer individual treatment effects or to bypass experimental validation. Any release of derived artifacts must respect the source datasets' access and licensing terms.

REPRODUCIBILITY STATEMENT

The appendix specifies task inputs, expression coordinates, data partitions, and evaluation units. It details model architectures, training objectives, optimization settings, checkpoint-selection rules, and bootstrap procedures, together with baseline configurations and training budgets.

REFERENCES

Abhinav K. Adduri, Dhruv Gautam, Beatrice Bevilacqua, et al. Predicting cellular responses to perturbation across diverse contexts with State. *Cell*, 189(19):5914–5931.e20, 2026. doi: 10.1016/j.cell.2026.07.052.

Constantin Ahlmann-Eltze, Wolfgang Huber, and Simon Anders. Deep-learning-based gene perturbation effect prediction does not yet outperform simple linear baselines. *Nature Methods*, 22(8):1657–1661, August 2025. ISSN 1548-7105. doi: 10.1038/s41592-025-02772-6. URL http://dx.doi.org/10.1038/s41592-025-02772-6.

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. ISSN 1548-7105. doi: 10.1038/s41592-023-01969-x. URL http://dx.doi.org/10.1038/s41592-023-01969-x.

Changxi Chi, Jun Xia, Yufei Huang, Zhuoli Ouyang, Cheng Tan, Yunfan Liu, Jingbo Zhou, Chang Yu, Liangyu Yuan, Siyuan Li, Zelin Zang, and Stan Z. Li. Doloris: Dual Conditional Diffusion Implicit Bridges with Sparsity Masking Strategy for Unpaired Single-Cell Perturbation Estimation. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/568b6cc71889ea0b2aa74152ef9c28db-Paper-Conference.pdf.

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, August 2024. ISSN 1548-7105. doi: 10.1038/s41592-024-02201-0. URL http://dx.doi.org/10.1038/s41592-024-02201-0.

Atray Dixit, Oren Parnas, Biyu Li, J. Chen, C. P. Fulco, L. Jerby-Arnon, N. D. Marjanovic, D. Dionne, T. Burks, R. Raychowdhury, B. Adamson, T. M. Norman, E. S. Lander, J. S. Weissman, N. Friedman, and A. Regev. Perturb-Seq: Dissecting Molecular Circuits with Scalable Single-Cell RNA Profiling of Pooled Genetic Screens. *Cell*, 167(7):1853–1866.e17, December 2016. ISSN 0092-8674. doi: 10.1016/j.cell.2016.11.038. URL http://dx.doi.org/10.1016/j.cell.2016.11.038.

Leon Hetzel, Simon Boehm, Niki Kilbertus, Stephan Günnemann, Mohammad Lotfollahi, and Fabian Theis. Predicting Cellular Responses to Novel Drug Perturbations at a Single-Cell Resolution. In *Advances in Neural Information Processing Systems*, volume 35, pp. 26711–26722, 2022. doi: 10.52202/068431-1937. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/aa933b5abc1be30baece1d230ec575a7-Paper-Conference.pdf.

Dominik Klein, Jonas Simon Fleck, Daniil Bobrovskiy, Lea Zimmermann, Sören Becker, Alessandro Palma, Leander Dony, Alejandro Tejada-Lapuerta, Guillaume Huguet, Hsiu-Chuan Lin, Nadezhda Azbukina, Fátima Sanchís-Calleja, Theo Uscidda, Artur Szalata, Manuel Gander, Aviv Regev, Barbara Treutlein, J. Gray Camp, and Fabian J. Theis. CellFlow enables generative single-cell phenotype modeling with flow matching. *bioRxiv*, 2025. doi: 10.1101/2025.04.11.648220.

Zeming Lin, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Nikita Smetanin, Robert Verkuil, Ori Kabeli, Yaniv Shmueli, Allan dos Santos Costa, Maryam Fazel-Zarandi, Tom Sercu, Salvatore Candido, and Alexander Rives. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379(6637):1123–1130, 2023. doi: 10.1126/science.ade2574.

Mohammad Lotfollahi, F. Alexander Wolf, and Fabian J. Theis. scGen predicts single-cell perturbation responses. *Nature Methods*, 16:715–721, 2019. doi: 10.1038/s41592-019-0494-8.

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, Jay Shendure, Jose L. McFaline-Figueroa, Pierre Boyeau, F. Alexander Wolf, Nafissa Yakubova, Stephan Günnemann, Cole Trapnell, David Lopez-Paz, and Fabian J. Theis. Predicting cellular responses to

complex perturbations in high-throughput screens. *Molecular Systems Biology*, 19(6):e11517, 2023. doi: 10.15252/msb.202211517.

Alexis Molina and Xinyi Zhang. Perturbation response decomposition enables biologically aligned generalization to unseen perturbations and cellular contexts. *bioRxiv*, 2026. doi: 10.64898/2026.07.24.740459.

John A. Morris, Christina Caragine, Zharko Daniloski, Júlia Domingo, Timothy Barry, Lu Lu, Kyrie Davis, Marcello Ziosi, Dafni A. Glinos, Stephanie Hao, E. P. Mimitou, P. Smibert, K. Roeder, E. Katsevich, T. Lappalainen, and N. E. Sanjana. Discovery of target genes and pathways at GWAS loci by pooled single-cell CRISPR screens. *Science*, 380(6646):eadh7699, 2023. doi: 10.1126/science.adh7699.

Ajay Nadig, Joseph M Replogle, Angela N Pogson, Mukundh Murthy, Steven A McCarroll, Jonathan S Weissman, Elise B Robinson, and Luke J O'Connor. Transcriptome-wide analysis of differential expression in perturbation atlases. *Nature Genetics*, 57(5):1228–1237, 2025. doi: 10.1038/s41588-025-02169-3.

Zoe Piran, Niv Cohen, Yedid Hoshen, and Mor Nitzan. Disentanglement of single-cell data with biolord. *Nature Biotechnology*, 42(11):1678–1683, November 2024. ISSN 1546-1696. doi: 10.1038/s41587-023-02079-x. URL http://dx.doi.org/10.1038/s41587-023-02079-x.

Joseph M. Replogle, Reuben A. Saunders, Angela N. Pogson, Jeffrey A. Hussmann, Alexander Lenail, Alina Guna, Lauren Mascibroda, Eric J. Wagner, Karen Adelman, Gila Lithwick-Yanai, Nika Iremadze, Florian Oberstrass, Doron Lipson, Jessica L. Bonnar, Marco Jost, Thomas M. Norman, and Jonathan S. Weissman. Mapping information-rich genotype-phenotype landscapes with genome-scale Perturb-seq. *Cell*, 185(14):2559–2575.e28, July 2022. ISSN 0092-8674. doi: 10.1016/j.cell.2022.05.013. URL http://dx.doi.org/10.1016/j.cell.2022.05.013.

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

Sanjay R. Srivatsan, José L. McFaline-Figueroa, Vijay Ramani, Lauren Saunders, Junyue Cao, Jonathan Packer, Hannah A. Pliner, Dana L. Jackson, Riza M. Daza, Lena Christiansen, Fan Zhang, Frank Steemers, Jay Shendure, and Cole Trapnell. Massively multiplex chemical transcriptomics at single-cell resolution. *Science*, 367(6473):45–51, January 2020. ISSN 1095-9203. doi: 10.1126/science.aax6234. URL http://dx.doi.org/10.1126/science.aax6234.

Jieun Sung and Wankyu Kim. scDEBART: Predicting in silico single-cell perturbation responses via large-scale differential expression learning. In *International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=pJyidZg93y.

Jialiang Wang, Ziqi Liu, Zhengqian Zhang, Yikun Cao, Junjun Ren, Peng Cheng, Jingjing Tian, Lingyun Xie, Xin Lu, Zhanwei Du, and Yongzhuang Liu. Predicting single-cell perturbation responses across biological contexts with a deep generative model integrating optimal transport. *Advanced Science*, pp. e77461, 2026. doi: 10.1002/advs.77461.

Frederik Wenkel, Wilson Tu, Cassandra Masschelein, Hamed Shirzad, Liam Hodgson, Ihab Bendidi, Cian Eastwood, Shawn T. Whitfield, Craig Russell, Yassir El Mesbahi, Jiarui Ding, Marta M. Fay, Berton Earnshaw, Emmanuel Noutahi, and Alisandra K. Denton. TxPert: using multiple knowledge graphs for prediction of transcriptomic perturbation effects. *Nature Biotechnology*, 2026. doi: 10.1038/s41587-026-03113-4.

Yan Wu, Esther Wershof, Sebastian Schmon, Marcel Nassar, Błażej Osiński, Ridvan Eksi, Zichao Yan, Rory Stark, Kun Zhang, and Thore Graepel. PerturBench: Benchmarking machine learning models for cellular perturbation analysis. In *Advances in Neural Information Processing Systems*, volume 38, pp. 106937–106977, 2025. doi: 10.52202/085713-3225. URL https://papers.nips.cc/paper_files/paper/2025/hash/8aee537279a66ced96319dfca3c00002-Abstract-Datasets_and_Benchmarks_Track.html.

Chenglei Yu, Chuanrui Wang, Bangyan Liao, and Tailin Wu. scDFM: Distributional Flow Matching Model for Robust Single-Cell Perturbation Prediction. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/75f029a1289a47aa99c86588239f0c12-Paper-Conference.pdf.

Xinyu Yuan, Xixian Liu, Ya Shi Zhang, Zuobai Zhang, Hongyu Guo, and Jian Tang. PerturbDiff: Functional diffusion for single-cell perturbation modeling. In *International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=yBAqonxCdp.

A Protocol and Metrics

A.1 Data Statistics

Table 4 summarizes the six processed benchmarks and their fixed cell-level partitions. The counts cover both source and target contexts; the training column therefore includes source observations as well as target-context training responses. The gene counts refer to the highly variable genes used for modeling and evaluation.

Table 4: **Dataset diversity, sizes, and cell-level partitions.** Perturbation counts exclude controls. Total cells include controls, counted once in the control column; splits count only perturbed cells.

Dataset	Type	Perturbations	Contexts	HVGs	Total cells	Controls	Perturbed cells		
							Training	Validation	Test
PBMC	Cytokine	90	216	2,000	9,697,974	629,701	6,657,336	150,484	2,260,453
Nadig	Genetic	2,023	4	2,000	643,413	39,165	572,545	4,825	26,878
SciPlex3	Chemical	188	3	4,000	178,213	11,935	114,255	24,060	27,963
Frangieh21	Genetic	248	3	4,000	218,331	18,334	162,311	18,530	19,156
OP3	Chemical	143	4	2,000	296,147	28,452	207,943	22,820	36,932
McFaline	Genetic	528	15	4,000	878,229	207,235	557,850	57,209	55,935

Contexts comprise 12 donors $\times$ 18 cell types in PBMC, four cell lines in Nadig, three cell lines in SciPlex3, three treatment environments in Frangieh21, four cell types in OP3, and three cell lines $\times$ five drug environments in McFaline. These counts describe the full benchmark, including source contexts, rather than only the held-out target contexts. SciPlex3 uses the highest-dose subset, and McFaline uses the full-covariate split across cell lines and drug environments. For OP3, 1,940 cells without an assigned partition are excluded from the counts and analysis. The control column reports all retained control cells; each evaluation uses the matched control population specified below.

Table 5 specifies the prediction targets and the available source contexts. Validation and test hold out perturbation responses within these target contexts, while other responses in the same contexts remain available for training. Thus, target contexts are partially observed rather than entirely excluded from training. For each query, empirical sources are restricted to observations of the same perturbation in eligible other contexts in the training partition. McFaline's three target combinations are T98G–lapatinib, A172–nintedanib, and U87MG–untreated.

The training counts in Table 5 refer to the target training panels for PBMC, Nadig, and Frangieh21; eligible source responses are additionally available for fitting. For SciPlex3, OP3, and McFaline, they count context–perturbation pairs across all contexts, with 131 training perturbations per cell line in SciPlex3. Validation and test counts refer to the listed target contexts. A context–perturbation pair counts a perturbation separately in each context.

Table 5: **Context and response partitions.** Counts exclude controls.

Dataset	Target contexts	Source contexts	Train	Val.	Test	Counting unit
PBMC	Donors 1, 4, 9, 12	Other donors, same cell type	24	4	62	Identities per target donor
Nadig	HepG2	K562, Jurkat, RPE1	900	60	380	Target identities
SciPlex3	A549, K562, MCF7	Other cell lines	393	85	86	Context–perturbation pairs
Frangieh21	Co-culture	Untreated, IFN-γ	74	87	87	Target identities
OP3	B cells, myeloid cells	Other cell types	331	89	144	Context–perturbation pairs
McFaline	Three combinations	Other cell-line–environment combinations	6,631	542	543	Context–perturbation pairs

A.2 Evaluation Protocol

Data provenance and comparison scope. PBMC and Nadig use the fixed PerturbDiff-style STATE partitions. PBMC uses the Parse Biosciences cytokine-stimulation panel as released with PerturbDiff (full transcriptome, 40,352 genes), from which the fixed 2,000-HVG log1p view used here is derived. Nadig uses the Replogle CRISPRi screen across four cell lines with HepG2 as the single recipient

context. The four external benchmarks use their fixed PerturBench-derived cell-level split tables (highest-dose SciPlex3; Frangieh21; OP3; McFaline GxE full-covariate). The processed files specify the highly variable gene sets and cell-level split membership. Within each dataset, the methods in the benchmark comparison tables use the same evaluation queries, populations, and gene space, with model-specific identity features and training budgets. OP3 preserves its published 144-query test population, including six queries with one training cell each, and is therefore not strictly query-disjoint. Appendix E.2 reports the corresponding overlap-exclusion analysis.

Split roles and source visibility. The index $\mathcal{I}_{\mathrm{tr}}$ identifies the training context–perturbation observations available for fitting, not a globally disjoint set of training perturbation identities. Under the query-disjoint formulation, training, validation, and test identities are disjoint within each target context, and the panel $\mathcal{P}_c^{\mathrm{tr}}$ is fixed across queries. Target training responses provide fitting supervision, target-validation responses support checkpoint selection, and target-test responses are reserved for evaluation. The same perturbation identity may nevertheless occur in the source training observations contributing to $\mathcal{S}_{c,p}$. For a query-disjoint target, responses to a test identity at other doses or in other technical batches are also excluded from fitting, checkpoint selection, and prediction.

Context, source, and identity definitions. Context and source definitions are summarized in Table 5. McFaline retains its full covariates and environmental dose strata and does not treat another dose in the same official context as an independent source context. The symbol p denotes perturbation identity across the retained dose and technical-batch strata. Dose and technical-batch indices are suppressed in the main-text notation. SciPlex3 and OP3 use 2,048-bit Morgan fingerprints with radius two, whereas genetic perturbations in Frangieh21 and McFaline use frozen mean-pooled ESM2 features. For McFaline, the retained 5,120-dimensional feature set covers all 528 identities after alias mapping. These identity features e_p are distinct from the context covariates.

Controls and populations. PBMC evaluates 2,260,453 test target cells across 4,452 conditions defined by donor, cell type, and perturbation, aggregated to 62 cytokine identities. The complete evaluation control background contains 235,478 cells. The Nadig evaluation includes 380 gene perturbations in HepG2, of which 372 have same-perturbation source evidence and eight do not. It comprises 26,878 test target cells and 4,976 evaluation control cells. The model's matched control anchors are sampled with replacement within the biological context. These anchors are distinct from the complete control background used by Cell-Eval, and individual baselines need not use them as model inputs. No observed pre/post pairing is assumed. Missing prior features are encoded as zero vectors; Frangieh21 includes 14 missing-prior identities among its 248 perturbation identities.

Metric definitions. All 13 metrics are computed with Cell-Eval 0.8.1. PDCorr is the Pearson correlation between predicted and real control-relative pseudobulk responses. MSE and MAE measure errors of absolute pseudobulk means. PDS evaluates discrimination of the correct perturbation against the fixed comparison pool using L1, L2 or cosine distance. DE significance is computed separately for real and predicted perturbations against their matched evaluation control background using cell-level Wilcoxon rank-sum tests, followed by Benjamini–Hochberg correction across genes at FDR < 0.05. DEOver and DEPrec are Cell-Eval's dynamic-length overlap and precision measures; the ranking length is set by real or predicted significant-DE counts, respectively, rather than by a single global top-k. DirAgr measures sign agreement on real-significant genes with predicted DE records; predicted significance is not required. LFCSpear measures log-fold-change rank agreement on real DE genes. AUROC and AUPRC use the real label $y_g = \mathbf{1}\{\mathrm{FDR}_g^{\mathrm{real}} < 0.05\}$ and the predicted score $s_g = -\log_{10}[\mathrm{clip}(\mathrm{FDR}_g^{\mathrm{pred}}, 10^{-10}, 1)]$, with missing predicted records assigned FDR one. AUPRC is computed as average precision; both metrics are undefined when the real labels contain only one class. Throughout the appendix, *native* refers to these benchmark significance scores and to evaluation under them, termed the benchmark DE scoring protocol in the main text.

ES is the Spearman correlation of predicted and real significant-DE gene counts, restricted to perturbations with positive real counts. Zero predicted counts remain included. By convention, the benchmark returns $\mathrm{ES} = 1$ when all real significant-DE gene counts are zero; one-element or constant-count correlations remain undefined. External ES is computed within each context and then macro-averaged, not from a pooled context table. Undefined quantities are displayed as dashes and

are not replaced with zero. Some Nadig DE metrics are defined for only 320 of the 380 identities; the source-free subset has only six identities with defined DE metrics.

Prediction conditions and metric aggregation. A PBMC prediction condition is indexed by donor, cell type, and perturbation, whereas the reported native DE evaluation pools cells by cytokine identity. For each cytokine, perturbed cells are pooled across the four target donors and their cell types, then tested against the complete control background. The 12 identity-level metrics are averaged arithmetically over cytokine identities. ES is instead computed across identities from the predicted and real significant-DE gene counts. Generation anchors are sampled with replacement from donor- and cell-type-matched controls and are distinct from this evaluation background. The local donor–cell-type error analysis in Appendix H.2 uses different control and aggregation definitions and is reported separately. Nadig results aggregate by gene perturbation identity in the single recipient context. External benchmarks first aggregate within each context and then apply equal-context macro averaging. Paired uncertainty analyses are reported in Appendix G.

Mean and distributional contributions to DE. For a gene, let Z and Z^0 denote independent draws from perturbed and control cells. The normalized Mann–Whitney statistic estimates $\Pr(Z > Z^0) + \frac{1}{2}\Pr(Z = Z^0)$, which is not determined by $\mathbb{E}[Z]$. Thus, even at fixed cell counts and control distributions, populations with equal means can yield different DE significance. Fold changes use a different summary: for positive group means under the evaluation's log1p input and geometric-mean convention, $\mathrm{LFC}_g = \log_2[\mathrm{expm1}(\bar{L}_g)/\,\mathrm{expm1}(\bar{L}^0_g)]$, where $L = 10z$ is log1p expression. A zero mean in only one group yields an infinite fold change; two zero means are assigned zero. With fixed controls and valid gene records, conserving $\bar{L}_g$ preserves these fold changes and their direction and rank metrics, while significance-based recovery can change. The fixed-mean interventions therefore assess distributional contributions to cell-level DE recovery; a metric depending only on the same pseudobulk means would remain unchanged.

Model provenance. The mean baselines implement training means grouped by perturbation, cell type, batch, or globally. No missing perturbation group is imputed from test responses. Nadig mean/perturbation* uses the training-group mean for 372 source-covered perturbations and copies target-context control cells for the other eight, using the fixed seed-42 with-replacement control bootstrap. All 380 identities and 26,878 target cells enter evaluation, including full-population PDS rankings. For source-free Nadig queries, scGen/scPILOT also use control-copy predictions; CPA uses zero-dose null interventions. Frangieh21 has no technical-batch field, so mean/batch is unavailable. CPA and CellFlow use their public implementations with the configurations and selection budgets specified in Appendix H.3. The released fine-tuned PerturbDiff generates in its unified 12,626-gene space before extracting the frozen 2,000 HVGs. STATE is trained from scratch with its official architecture adapted to the benchmark HVG space. Its PBMC training omits 90 unpairable training rows; evaluation populations are unchanged.

B Implementation Details

Notation and configuration. Table 6 summarizes the main quantities; Table 7 collects the shared settings and stage-specific optimization choices. Vectors of gene-wise statistics are rows, and expression quantities are in model coordinates unless stated otherwise.

Table 6: **Principal notation for COHERE.**

Symbol	Meaning and dimensions
$c, p;\ N, G, K$	Context and perturbation; numbers of generated cells, genes, and positive-quantile levels
$X_c^0, \mathcal{S}_{c,p}$	Target controls and eligible same-perturbation source populations
$z = L/\gamma$	Model coordinates; L is processed log1p expression
m_c^0, δ_{cp}	Control mean and perturbation-induced mean change, each $1 \times G$
$\ell_{cp}, q_{cp,k}$	Residual occurrence logit and positive-expression quantile at τ_k, each $1 \times G$
R_{cp}, H_{cp}	Aggregated source descriptor, $(K+2) \times G$, and projection $R_{cp}B$, $(K+2) \times d$
$B;\ U_c, V_c$	Orthonormal response basis, $G \times d$; shared recipient factors, $d \times r_c$
e_p, ξ_p, α	Frozen identity features, $1 \times d_e$; bounded basis correction, $1 \times d$; scalar shrinkage
$\mu, \widehat{\pi}$	Feasible mean and restored continuous expressing fraction, each $1 \times G$
$C, \bar{C};\ k_g$	Reference and generation caps; integer nonzero support for gene g
$\widehat{Z}, \widehat{X}$	Generated populations in model and log1p coordinates, each $N \times G$; $\widehat{X} = \gamma\widehat{Z}$

Table 7: **COHERE configuration summary.** Detailed objectives and dataset-specific fitting schedules are specified in the paragraphs below.

Component	Configuration
Response representation	$\gamma = 10$, basis dimension $d = 128$, maximum mixing rank 32, minimum panel size $n_{\min} = 3$; $K = 5$ positive-quantile knots
Empirical adapter	8,000 updates; AdamW, learning rate 3×10^{-4}, weight decay 10^{-4}; batch $\min(64, n_c)$; gradient clipping 1; EMA 0.995
Identity head	LayerNorm, widths 64 and 32, GELU, support/sign outputs; no dropout; same optimizer settings, clipping and EMA as the adapter
Identity-head budget	Primary full fits: 8,000 updates; external full fits and five auxiliary identity-held-out heads: 2,000 updates each
Auxiliary empirical fits	Ten identity folds; 2,000 updates per fit on PBMC and 4,000 on Nadig
Identity supervision	$\lambda_{\text{supp}} = \lambda_{\text{sign}} = 0.025$; residual-scale floor 10^{-4}; unit correction amplitude during head fitting
Global shrinkage	4,000 updates; Adam, learning rate 0.01; batch 64; clipping 1; initial $\alpha = 0.1$; no EMA
Encoder and allocator	Hidden width 256, token dropout 0.05, single state; 6,000 pretraining updates; AdamW, learning rate 2×10^{-4}, weight decay 0.01; batch 8; clipping 0.5
Pretraining cohorts	Primary datasets: available sizes in $\{64, 128, 256\}$; external datasets: 64 cells
Finite-cell realization	$C = 1.5$, $\epsilon = 10^{-6}$, $\eta = 10^{-4}$, $\bar{C} = C(1 - 2\epsilon)$; cumulative-maximum knots; capped scaling with 100 bisection iterations
Checkpoint selection	Adapter and primary full identity head: 2k/4k/6k/8k; shrinkage: 250/1k/2k/4k; equal-weight validation ranks across four metric families, earlier candidate on ties
Random seed	42 for local fitting and inference

Prediction procedure. Algorithm 1 summarizes prediction for one recipient–perturbation query using fixed fitted parameters. It distinguishes the adapted occurrence coordinate, its restoration using the identity-corrected feasible mean, and the integer support of the generated population. Training, missing-input conventions, and numerical details are specified below.

Missing inputs. When $\mathcal{S}_{c,p_*} = \varnothing$, COHERE uses a zero source descriptor in the same transformation and identity pathway, retaining the query in prediction and evaluation. This zero vector encodes unavailable source-response evidence rather than an observed biological null response. Missing frozen identity features e_p are encoded as zero vectors. Baseline-specific source-free predictions follow the conventions in Appendix A.

Representation and recipient training. We use $z = L/10$, where L is the processed log1p expression matrix used for the benchmark. This scaling changes numerical coordinates without applying an additional normalization. The empirical model uses a fixed 128-dimensional response basis, rank at most 32, and three response-family normalizers. We call a recipient–perturbation row usable when its observed response and recipient-control statistics are available in the training split; source coverage is not required, and source-free rows retain a zero source descriptor. Let n_c denote the number of such rows from $\mathcal{P}_c^{\text{tr}}$ in recipient c after any auxiliary identity holdout; n_c is 21–24 for the full PBMC fits and 900 for the full Nadig fit. We set $n_{\min} = 3$, so panels with fewer than three usable rows bypass both centering and adaptation. Empirical fits run 8,000 updates with checkpoints

Algorithm 1 COHERE prediction for a recipient–perturbation query

Require: Query (c, p_*), source populations $\mathcal{S}_{c,p_*}$, target controls X_c^0, identity features e_{p_*}, and cell count N
Require: Fitted basis B, family normalizers and panel centers, recipient maps, identity head and shrinkage α, frozen encoder and allocator
Ensure: Predicted population $\widehat{X}_{cp_*}$ in processed log1p units

Compute target-control mean m_c^0 and source descriptors $(\delta, \ell, q_1, \ldots, q_K)$ in $z = L/\gamma$
Huber-aggregate source descriptors into R_{cp_*}; set $R_{cp_*} = 0$ if the source set is empty
$H_{cp_*} \leftarrow R_{cp_*} B$
if $n_c \geq n_{\min}$ **then**
    Standardize H_{cp_*} and subtract the source-panel centers
    Apply the family-specific recipient maps in Equation 2
    Add target-panel centers and reverse standardization to obtain $\widehat{H}_{cp_*}$
else
    $\widehat{H}_{cp_*} \leftarrow H_{cp_*}$
end if
Reconstruct $\widehat{R}_{cp_*}$ with the source-preserving update in Equation 3
Extract $\widehat{\delta}^{\mathrm{emp}}_{cp_*}$, $\widehat{\ell}$, and $\widehat{q}$ from $\widehat{R}_{cp_*}$
Apply the identity correction to obtain $\widehat{\delta}_{cp_*}$ (Equation 5)
$\bar{C} \leftarrow C(1 - 2\epsilon)$; $\mu \leftarrow \mathrm{clip}(m_c^0 + \widehat{\delta}_{cp_*}, 0, \bar{C})$
Restore $\widehat{\pi}$ from $(\mu, \widehat{\ell})$ and compute k_g for each gene (Equation 6)
Sample N context-matched control anchors and compute frozen gene-wise ordering scores
Initialize $\widehat{Z} \leftarrow 0_{N \times G}$
for each gene g with $\mu_g > 0$ **do**
    Clip $\widehat{q}_{:,g}$ at zero and take its cumulative maximum in quantile order
    Interpolate k_g values; apply a positive floor, or use ones for an all-zero shape
    Obtain ordered capped values $\widetilde{z}_g$ with total $N\mu_g$ (Equation 7)
    Assign $\widetilde{z}_g$ to column g of $\widehat{Z}$ at the top k_g anchors, in ascending score order
end for
return $\widehat{X}_{cp_*} = \gamma \widehat{Z}$

at 2,000, 4,000, 6,000, and 8,000; the effective batch is $\min(64, n_c)$, without padding small panels. AdamW uses learning rate 3×10^{-4}, weight decay 10^{-4}, gradient clipping at one, and EMA decay 0.995. The variants in Appendix D.1 share these empirical training and selection budgets.

Centering and loss normalization. For each recipient, source and target panel centers are computed by averaging their corresponding normalized training rows. Each family normalizer consists of a scalar median and a scaled MAD, computed over identity-aggregated training rows, descriptor channels, and response-basis coordinates. Reference losses compare the unadapted source descriptor with the target training descriptor. The mean reference loss gives equal weight to SmoothL1 and cosine error, the occurrence reference loss uses BCE after target-control restoration, and the positive-quantile reference loss uses pinball loss. Each reference loss is floored at 10^{-8} and used to normalize the corresponding family loss; the three normalized losses are then averaged. Source aggregation uses a one-step coordinate-wise Huber location estimate with threshold 1.5, falling back to the arithmetic mean when the MAD is zero.

Empirical shape targets and losses. For M minibatch populations, q^{tar}_{bkg} is the empirical positive-expression quantile for population b and gene g at τ_k. Without positive observations, all K target knots are zero. Predicted knots are projected by clipping at zero and taking cumulative maxima,

$$\bar{q}_{bkg} = \max_{j \leq k} \max(\widehat{q}_{bjg}, 0), \qquad \mathcal{L}_q = \frac{1}{MKG} \sum_{b=1}^{M} \sum_{k=1}^{K} \sum_{g=1}^{G} \rho_{\tau_k}\left(q^{\mathrm{tar}}_{bkg} - \bar{q}_{bkg}\right),$$

where $\rho_\tau(u) = \max\{\tau u, (\tau - 1)u\}$. Zero-coded targets participate in this reduction without an additional positive-support mask, and zero-coded source knots remain in the complete descriptor arrays used by the Huber source aggregation. For occurrence supervision, predicted and target means are separately clipped to the feasible range and combined with their respective residual occurrence logits through the occurrence-restoration map $\pi = \mu/C + (1 - \mu/C)\,\mathrm{sigmoid}(\ell)$ before BCE is evaluated. Here μ denotes the respective feasible predicted or target mean. Each update first samples a recipient panel uniformly and then draws $\min(64, n_c)$ of its training rows without replacement.

Residual-head targets. Let r_{cp} be the identity-held-out empirical mean residual projected into the fixed response basis, and let $\bar{r}_p$ average these residuals over the training contexts available for identity p. The coordinate scale used both for the correction bound and the support label is

$$\sigma_{r,j} = \max\left\{1.4826 \ \mathrm{median}_p \left|\bar{r}_{p,j} - \mathrm{median}_{p'} \ \bar{r}_{p',j}\right|, 10^{-4}\right\}.$$

The residual head is trained with the full correction $\widehat{\delta}_{cp}^{\mathrm{emp}} + \xi_p B^\top$, corresponding to unit residual amplitude before shrinkage, and uses $\lambda_{\mathrm{supp}} = \lambda_{\mathrm{sign}} = 0.025$. Only after this head is frozen do we fit the dataset-level shrinkage α.

Identity-residual fitting and shrinkage. Frozen identity features pass through LayerNorm, a linear layer of width 64 followed by GELU, and a linear layer of width 32. Two zero-initialized linear heads produce support and sign logits, each with d coordinates. This branch uses no dropout and is trained with AdamW, using learning rate 3×10^{-4}, weight decay 10^{-4}, gradient clipping at one, and EMA decay 0.995.

On PBMC and Nadig, the full identity-residual head is trained for 8,000 updates and selected from checkpoints at 2,000, 4,000, 6,000, and 8,000 updates. External full heads use a fixed budget of 2,000 updates. Five auxiliary residual heads use identity holdouts, with 2,000 updates per head. The empirical pathway uses ten auxiliary empirical fits per dataset to construct held-out features, with 2,000 updates per fit on PBMC and 4,000 on Nadig. These auxiliary models support residual-head and shrinkage training and do not form an inference-time ensemble.

Empirical cross-fitting uses ten fixed identity folds per dataset, whereas the auxiliary residual heads use five identity folds assigned deterministically with fixed seed. The empirical fits share the response basis and cached descriptor representation, while each fold computes family normalizers and panel centers from its retained training identities. The full and auxiliary residual heads reuse the upstream empirical fits, cached identity-held-out predictions, and common residual scales.

Global shrinkage is fitted using 4,000 Adam updates with learning rate 0.01, batch size 64, gradient clipping at one, initial $\alpha = 0.1$, and no EMA. Checkpoints at 250, 1,000, 2,000, and 4,000 updates are candidates for validation-based selection. An identity-balanced sampler first selects an identity and then one of its rows. The loss combines normalized SmoothL1 and cosine terms after feasible mean restoration. The shrinkage schedule does not depend on source coverage.

Training-stage dependencies. Table 8 summarizes the training order, frozen inputs, and downstream roles of each component. Allocator pretraining and response adaptation share the fixed response representation but have separate objectives; the encoder and allocator are frozen before final cell realization. Their architecture and pretraining objectives are detailed below.

Table 8: **Training-stage overview.** Fits use training observations; target validation selects candidates.

Stage	Fitted quantities and supervision	Fixed inputs and downstream role
Response representation	Mean-response basis, family normalizers, and panel centers from training descriptors	Fixed representation for subsequent fits
Cell ordering	Encoder and allocator; sorted-expression and smooth occupancy objectives	Fixed basis; frozen ordering scores used during realization
Empirical adaptation	Recipient factors and family scalars; mean, occurrence, and quantile losses	Fixed representation; predicts empirical response descriptors
Auxiliary predictions	Identity-held-out empirical fits; residual targets and scales	Shared representation; supplies supervision for the identity branch
Identity correction	Full and auxiliary residual heads; corrected-response, support and sign losses	Frozen empirical pathway; full correction amplitude during head fitting
Global shrinkage	Dataset-level scalar α; identity-held-out corrected-mean loss	Frozen empirical and identity heads; sets final correction amplitude

Selection. For COHERE, the predefined candidate checkpoints are ranked using target-validation metrics in their favorable direction. Ranks are averaged within numerical (PDCorr/MSE/MAE), PDS (three distances), DE (DEOver/DEPrec/AUROC/AUPRC), and direction (DirAgr) families, then averaged equally across families. The candidate with the lowest average rank is selected; exact ties prefer the earlier update. LFCSpear and ES are reporting metrics, not selector terms.

Baselines retain their own configurations and candidate budgets, as detailed in Appendix H.3. For the SciPlex3/Frangieh21 shared empirical and global-shrinkage runs, validation uses identity-macro summaries as specified by their original protocol; the reported test tables use equal-context macro.

Realization. The reference expression cap is $C = 1.5$ in model units, equivalent to 15 in processed log1p units, with $\epsilon = 10^{-6}$, occurrence boundary $\eta = 10^{-4}$, and generation cap $\bar{C} = C(1 - 2\epsilon)$. It is a modeling and generation cap rather than an upper bound on every observed processed expression value. The residual occurrence coordinate is recombined with the feasible mean before determining support. The positive-quantile knots are clipped at zero and replaced by their cumulative maximum in quantile order, interpolated to the requested support size, then scaled under the cap to attain the predicted total. Ranking scores come from the single-responder-state encoder/allocator, pretrained for 6,000 updates with sorted-expression and smooth occupancy-proxy supervision. These scores assign the marginal values to control-derived cell representations. The encoder architecture and pretraining losses are specified below.

Finite-cell implementation. For each query (c, p_*), the requested count N equals its evaluation target-cell count, so the generated population $\widehat{X}_{cp_*}$ has the matched size. Generation uses a fixed context-matched set of control anchors sampled with replacement while preserving the required technical-batch strata. The support calculation in Equation 6 uses round-to-nearest with ties to even. The five knot levels are $(i + 0.5)/32$ for $i \in \{7, 15, 23, 28, 31\}$. Linear interpolation uses ranks $(j + 0.5)/k$, with $j = 0, \ldots, k - 1$, and constant endpoint extension. Before capped scaling, nonzero shapes receive the floor $\max(10^{-4} \min_{v>0} v, 10^{-8})$; an all-zero shape is replaced by ones. The capped scaling solves $\sum_j \min(\lambda v_j, \bar{C}) = N\mu_g$ by a fixed 100-iteration bisection on $\lambda \in [0, \bar{C}/\min_j v_j]$ with no tolerance-based early stopping, taking the final bracket midpoint as the scale. Shape scaling is computed in float64, and generated populations are stored in float32. We verify a maximum gene-wise mean deviation of at most 2×10^{-6} in model units; the external fixed-mean diagnostics use the tolerances specified in Appendix D.6. Allocator ties use stable ordering.

For $\mu_g > 0$, the restored probability satisfies $\widehat{\pi}_g \geq \mu_g/C$, giving $\mu_g/\bar{C} - \widehat{\pi}_g \leq 2\epsilon$. Nearest-integer rounding and the capacity ceiling in Equation 6 therefore yield

$$-\frac{1}{2N} \leq \frac{k_g}{N} - \widehat{\pi}_g \leq \frac{1}{N} + 2\epsilon.$$

For $\mu_g = 0$, nonnegativity and mean conservation require an all-zero gene.

Cell encoder and allocator. For a control anchor $x_i^0 \in \mathbb{R}^G$ in model units, the encoder token input is $[\log(x_i^0 + 10^{-4})B,\ \mathbf{1}\{x_i^0 > 0\}B] \in \mathbb{R}^{2d}$. The token MLP consists of Linear$(2d, 256)$, SiLU, dropout 0.05, Linear$(256, 256)$, and SiLU. We average the token representations across the cohort and pass the resulting summary through Linear$(256, 256)$, SiLU, and Linear$(256, 256)$. The transformed cohort summary is concatenated with each token representation and mapped by Linear$(512, 256)$, followed by LayerNorm and Linear$(256, d)$, to produce h_i. The reported model uses a single-state head with LayerNorm and a linear layer, so its state assignment is constant. The scoring function is

$$a_{ig} = \log(x_{ig}^0 + 10^{-4}) + [(h_i + v)B^\top]_g, \qquad v = u \odot w, \quad u, w \in \mathbb{R}^d,$$

where u and w are learned vectors initialized to zero and one, respectively. For a fixed gene and generated group, $(vB^\top)_g$ is constant across anchors and therefore does not affect their ordering. The equivalent ordering score is $\log(x_{ig}^0 + 10^{-4}) + (h_i B^\top)_g$. At inference, cell ordering depends on the control anchors and frozen encoder/allocator parameters, not on observed query responses or identity embeddings. Predicted mean and shape determine the marginal values and support through a separate realization procedure. At inference, cohort pooling is permutation invariant, and the per-cell encoder outputs and scores are permutation equivariant. Score ties are resolved by stable ordering in the input anchor order. Permuting the realized values within each gene and generated group preserves pseudobulk means, marginal distances, and univariate DE statistics. These metrics assess transferred marginals, whereas joint-distribution measures assess cross-gene assembly.

Encoder and allocator pretraining. During pretraining, the shared encoder is evaluated on both a training response cohort Y and its independently sampled control anchors X^0. A support-gated capped-softmax decoder produces P^{post} and P^{prior} during pretraining. The decoder is

replaced by the quantile-based realization procedure at inference. For a training cohort of n cells, let $\mu_g \in [0, \bar{C}]$ and o_g denote its mean and occurrence. For $\mu_g > 0$, the support size is $k_g = \max(\mathrm{round}(n\, o_g), \lceil n\mu_g/(C(1-\epsilon))\rceil)$. With logits a_{ig} and the k_g-th largest tie-broken logit θ_g, define $g_{ig} = \mathrm{clip}(\mathrm{sigmoid}((a_{ig} - \theta_g)/0.25), 10^{-4}, 1 - 10^{-4})$. The soft logits $a_{ig} + \log g_{ig}$ supply a straight-through gradient surrogate. The forward pass applies capped softmax to the original logits on the hard top-k_g support, with probability cap $c_g = C(1-\epsilon)/(n\mu_g)$ and zero mass outside the support. Its output $q_{\cdot g}$ defines $P_{ig} = n\,\mu_g\, q_{ig}$. Genes with $\mu_g = 0$ are exactly zero. For cohorts of n cells, define

$$L_{\mathrm{sort}}(P, Y) = \frac{1}{nG} \sum_{g,i} |P_{(i)g} - Y_{(i)g}|, \quad t_g(Y) = \frac{1}{n} \sum_i \frac{Y_{ig}}{Y_{ig} + 0.05},$$

where sorting is separate within each gene. With $\bar{t} = \mathrm{clip}(t, 10^{-4}, 1 - 10^{-4})$,

$$L_{\mathrm{occ}}(P, Y) = -\frac{1}{G} \sum_g [\bar{t}_g(Y) \log \bar{t}_g(P) + (1 - \bar{t}_g(Y)) \log(1 - \bar{t}_g(P))].$$

The objective is $\sum_{b \in \{\mathrm{post}, \mathrm{prior}\}} [0.5 L_{\mathrm{sort}}(P^b, Y) + 0.25 L_{\mathrm{occ}}(P^b, Y)]$. The occurrence pretraining objective is a smooth positive-mass proxy rather than BCE on exact zero fractions. Pretraining uses 6,000 updates, batch eight, AdamW learning rate 2×10^{-4}, weight decay 0.01, and gradient clipping 0.5. The main-data cohort size is sampled from available sizes in $\{64, 128, 256\}$; external datasets use 64-cell cohorts. The pretraining split is drawn within training episodes. After pretraining, the encoder and allocator are frozen and only their ordering scores are reused. COHERE's quantile-based mean-conserving realization procedure replaces the pretraining decoder.

C Additional Experimental Results

In unshaded tables, boldface marks the best values for each metric and dataset, including ties at the reported precision. For Cell-Eval comparisons, MSE and MAE are minimized and the other metrics maximized. Wasserstein, occurrence, and quantile errors are minimized.

C.1 Source-Preserving Reconstruction

The subspace-only control reconstructs all descriptor channels as $\widehat{R}_{cp} = \widehat{H}_{cp} B^\top$, replacing the source-preserving update in Equation 3. It uses the same source coordinates $H_{cp} = R_{cp} B$ as inputs, but omits the source component $R_{cp}(I_G - BB^\top)$ from the reconstructed descriptors. The fixed response basis, family normalizers, panel centers, control anchors, allocator, and realization procedure are shared with COHERE. We refit the empirical adapter and rebuild its dependent held-out residual targets and scales, then refit the full and auxiliary identity heads and global shrinkage. This is a matched retraining comparison of reconstruction choices; the fixed-mean interventions in Appendix C.8 instead change shape inputs within one fitted predictor.

Table 9: **Response-transfer controls.** Direct transfer retains source descriptors; the empirical backbone adapts them without identity correction. Subspace-only is retrained without source components outside the response basis. All settings share target-control restoration and realization. Best values are bold. Paired intervals appear in Appendices D.2 and C.1.

	PBMC			Nadig		
Method	PDCorr↑	MSE↓	AUPRC↑	PDCorr↑	MSE↓	AUPRC↑
Direct transfer	0.85218	1.221e-04	**0.72353**	0.33724	9.388e-03	0.21017
Empirical backbone	0.89570	9.232e-05	0.64340	0.47468	6.614e-03	0.32927
COHERE	**0.89572**	**9.221e-05**	0.64345	0.47792	6.612e-03	**0.33193**
Subspace-only (retrained)	0.85958	1.168e-04	0.53327	**0.48923**	**5.473e-03**	0.20634

Training and selection. All stages use seed 42 and the same formal training partition as the complete predictor. Table 10 lists 18 fitting trajectories per dataset, including auxiliary fits used to construct training targets. We evaluate four checkpoints from each full empirical, identity, and

shrinkage trajectory and select one per stage using the four-family validation criterion in Appendix B, giving 12 candidate evaluations per dataset. The selected empirical/identity/shrinkage updates are 2,000/2,000/4,000 on PBMC and 4,000/2,000/1,000 on Nadig.

Table 10: **Matched subspace-only retraining budget.** Counts are per dataset; updates are per fit. The response representation and allocator are reused.

Stage	Fits	PBMC updates	Nadig updates
Full empirical adapter	1	8,000	8,000
Auxiliary empirical adapters	10	2,000	4,000
Full identity head	1	8,000	8,000
Auxiliary identity heads	5	2,000	2,000
Global shrinkage	1	4,000	4,000
Total optimization updates	18	50,000	70,000

Matched evaluation and uncertainty. The selected predictors use the full registered test populations, 2,000 genes, and matched controls: 62 PBMC cytokines across 4,452 generation groups, and 380 Nadig perturbations. Tables 11 and 12 report native Cell-Eval metrics and 95% intervals from 2,000 paired identity-bootstrap draws, conditional on the fitted models. Each draw resamples all identities with replacement and retains their associated records. For ES, we recompute the real-versus-predicted DE-count Spearman correlation for each model within every draw, restricting to positive real-DE counts and retaining zero predicted counts. The point-estimate ES support is 62 identities on PBMC and 320 on Nadig. The MSE ratio is computed from the resampled mean errors: COHERE/subspace-only is 0.78959 $[0.73716, 0.83403]$ on PBMC and 1.20801 $[1.17975, 1.23693]$ on Nadig. The positive DE and marginal-fidelity contrasts coexist with a Nadig mean-accuracy advantage for subspace-only reconstruction.

Table 11: **Source-preservation comparison on PBMC.** Differences are COHERE minus subspace-only, with paired 95% intervals. Lower MSE/MAE and higher other metrics are favorable.

Metric	COHERE	Subspace-only	Difference	95% interval
DEOver	**0.59725**	0.49055	0.10670	[0.09786, 0.11471]
DEPrec	**0.59087**	0.53740	0.05347	[0.04987, 0.05691]
ES	**0.85364**	0.08813	0.76551	[0.51088, 1.01011]
DirAgr	**0.85700**	0.80670	0.05030	[0.04387, 0.05725]
LFCSpear	**0.71584**	0.59651	0.11933	[0.10699, 0.13241]
AUPRC	**0.64345**	0.53327	0.11018	[0.10319, 0.11731]
AUROC	**0.64563**	0.48708	0.15855	[0.14831, 0.16892]
PDCorr	**0.89572**	0.85958	0.03614	[0.02741, 0.04742]
MSE	**9.221e-05**	1.168e-04	-2.457×10^{-5}	$[-3.252 \times 10^{-5}, -1.843 \times 10^{-5}]$
MAE	**0.00382**	0.00465	-8.246×10^{-4}	$[-0.00101, -6.813 \times 10^{-4}]$
PDS_1	**0.98543**	0.97451	0.01093	[0.00520, 0.01795]
PDS_2	**0.97711**	0.96592	0.01119	[0.00468, 0.02081]
PDS_{cos}	**0.99063**	0.97503	0.01561	[0.00442, 0.03018]

Table 12: **Source-preservation comparison on Nadig.** Differences are COHERE minus subspace-only, with paired 95% intervals. Lower MSE/MAE and higher other metrics are favorable.

Metric	COHERE	Subspace-only	Difference	95% interval
DEOver	**0.24852**	0.20076	0.04776	[0.03164, 0.06553]
DEPrec	**0.20733**	0.15555	0.05178	[0.04566, 0.05827]
ES	**0.78167**	0.69837	0.08330	[0.05159, 0.11557]
DirAgr	**0.88607**	0.88102	0.00505	[−0.01064, 0.02042]
LFCSpear	**0.71128**	0.70833	0.00295	[−0.02947, 0.03327]
AUPRC	**0.33193**	0.20634	0.12558	[0.11491, 0.13606]
AUROC	**0.76812**	0.64499	0.12313	[0.10708, 0.14050]
PDCorr	0.47792	**0.48923**	−0.01131	[−0.01776, −0.00483]
MSE	0.00661	**0.00547**	0.00114	$[9.985 \times 10^{-4}, 0.00129]$
MAE	0.05598	**0.04960**	0.00639	[0.00587, 0.00690]
PDS_1	**0.84019**	0.80251	0.03768	[0.03064, 0.04564]
PDS_2	**0.86208**	0.82323	0.03885	[0.03170, 0.04672]
PDS_{cos}	**0.74600**	0.70231	0.04368	[0.03223, 0.05552]

C.2 Marginal Fidelity Against External Baselines

We compare COHERE, STATE, PerturbDiff scratch and fine-tuned predictors, and the retrained subspace-only control on the primary datasets. The evaluations share 2,000 genes in processed log1p coordinates and match each generation group's observed and predicted cell counts. PBMC comprises 4,452 groups and 2,260,453 target cells; Nadig comprises 380 groups and 26,878 target cells. For each gene and group, W1 integrates the difference between the full empirical cumulative distribution functions, including zero mass. Occurrence MAE is the absolute difference in fractions with expression strictly above zero. We average over genes, then over groups within each perturbation identity, and finally weight the 62 PBMC or 380 Nadig identities equally. Thus W1 retains processed log1p units and occurrence error is dimensionless. Table 3 gives all method-level values; Table 13 gives paired differences from 2,000 seed-42 identity-bootstrap draws.

COHERE has lower W1 than all three external predictors on PBMC and both PerturbDiff variants on Nadig, whereas STATE has lower Nadig W1. Occurrence MAE favors PerturbDiff scratch on PBMC and the fine-tuned model on Nadig. Subspace-only reconstruction increases both errors on both datasets, complementing the DE comparisons in Appendix C.1. These external comparisons and the fixed-mean shape interventions in Appendix C.7 answer distinct questions: relative predictive fidelity and the effect of adapted shape at a shared mean, respectively.

Table 13: **Paired marginal-error contrasts.** COHERE minus comparator, with 95% identity-bootstrap intervals; negative values favor COHERE.

Dataset	Comparator	W1 difference	Occ. MAE difference
PBMC	STATE	−0.00154 [−0.00169, −0.00139]	−0.00751 [−0.00862, −0.00643]
PBMC	PerturbDiff scratch	−0.00373 [−0.00408, −0.00338]	+0.00728 [+0.00637, +0.00819]
PBMC	PerturbDiff finetuned	−0.00573 [−0.00592, −0.00554]	−0.00558 [−0.00676, −0.00437]
PBMC	Subspace-only (retrained)	−0.02937 [−0.02959, −0.02917]	−0.21665 [−0.21825, −0.21476]
Nadig	STATE	+0.03152 [+0.03058, +0.03249]	−0.00439 [−0.00644, −0.00240]
Nadig	PerturbDiff scratch	−0.02335 [−0.02475, −0.02194]	−0.00358 [−0.00465, −0.00245]
Nadig	PerturbDiff finetuned	−0.00812 [−0.00991, −0.00637]	+0.00382 [+0.00224, +0.00542]
Nadig	Subspace-only (retrained)	−0.02319 [−0.02454, −0.02195]	−0.04730 [−0.04826, −0.04635]

C.3 DE Score Sensitivity

Stable scores from expression populations. We compute significance directly in log space for the four PBMC fixed-mean shape interventions, direct transfer, STATE, and both PerturbDiff variants. The four interventions are denoted 00, 10, 01, and 11, with the first digit selecting source (0) or learned (1) occurrence and the second selecting source or learned positive quantiles; 11 is COHERE. All comparisons use the same 62 identities, 2,000 genes, 235,478 real control cells, real-data BH

labels, and registered target-cell counts. For sample sizes n_1, n_2, pooled tie multiplicities t_j, and U_g the first-sample statistic, define

$$v_g = \frac{n_1 n_2}{12}\left[n_1 + n_2 + 1 - \frac{\sum_j (t_j^3 - t_j)}{(n_1+n_2)(n_1+n_2-1)}\right],$$
$$a_g = \frac{\max(U_g, n_1 n_2 - U_g) - n_1 n_2/2 - 1/2}{\sqrt{\max(v_g, 10^{-300})}}.$$

We evaluate $s_g^{\text{stable}} = \max\{0, -[\log 2 + \log \Phi(-a_g)]/\log 10\}$ in float64, where Φ is the standard normal CDF. Direct log-CDF evaluation avoids underflow and the native significance-score ceiling.

To retain multiple-testing adjustment without numerical underflow, we also compute BH-adjusted $\log q_g$ directly from $\log p_g$. We compare three scores: stable unadjusted p ($-\log_{10} p_g$), stable uncapped q ($-\log_{10} q_g$), and stable capped q ($\min\{10, -\log_{10} q_g\}$). The capped and uncapped q scores use the same gene statistics and differ only in the score ceiling. AP and AUROC retain score ties and use fixed real-data DE labels at FDR < 0.05; all 62 identities have both label classes. Intervals use 2,000 paired identity-bootstrap draws at seed 42. Baseline populations are generated from frozen predictors using the benchmark queries, genes, controls, and cell counts. Because baseline cell realizations and numerical backends can differ from the native evaluation, clipping effects are identified by comparing capped and uncapped scores within each population.

Table 14: **Stable PBMC significance rankings.** Each score column uses the same underlying population for a given predictor. AUPRC and AUROC are identity means; 00–11 denote the fixed-mean shape interventions.

	Capped q		Uncapped q		Unadjusted p	
Predictor	AUPRC	AUROC	AUPRC	AUROC	AUPRC	AUROC
00	0.64023	0.64162	0.69474	0.66071	0.69474	0.66071
10	0.64373	0.64602	0.69495	0.66482	0.69496	0.66482
01	0.63999	0.64133	0.69420	0.66030	0.69421	0.66030
11 (COHERE)	0.64345	0.64563	0.69409	0.66423	0.69409	0.66423
Direct transfer	**0.72353**	**0.71787**	**0.75911**	**0.72800**	**0.75912**	**0.72800**
STATE	0.59405	0.60048	0.63422	0.63168	0.63422	0.63168
PerturbDiff scratch	0.61087	0.61367	0.64730	0.63062	0.64731	0.63062
PerturbDiff finetuned	0.64433	0.65938	0.68428	0.68122	0.68428	0.68122

Table 15: **Stable PBMC AUPRC contrasts.** Paired differences with 95% identity-bootstrap intervals. Shape contrasts fix the complete model's feasible means; predictors retain their own fitted means.

Contrast	Capped q	Uncapped q	Unadjusted p
Occ. / source Q	$+0.00350$ [$+0.00046$, $+0.00636$]	$+0.00022$ [-0.00386, $+0.00418$]	$+0.00022$ [-0.00385, $+0.00418$]
Occ. / learned Q	$+0.00346$ [$+0.00041$, $+0.00634$]	-0.00012 [-0.00420, $+0.00386$]	-0.00011 [-0.00420, $+0.00386$]
Q / source Occ.	-0.00024 [-0.00038, -0.00011]	-0.00053 [-0.00064, -0.00042]	-0.00053 [-0.00064, -0.00042]
Q / learned Occ.	-0.00028 [-0.00044, -0.00012]	-0.00086 [-0.00105, -0.00068]	-0.00086 [-0.00105, -0.00068]
Joint shape	$+0.00322$ [$+0.00019$, $+0.00605$]	-0.00065 [-0.00470, $+0.00330$]	-0.00065 [-0.00469, $+0.00330$]
Interaction	-0.00004 [-0.00021, $+0.00014$]	-0.00033 [-0.00043, -0.00023]	-0.00033 [-0.00044, -0.00023]
COHERE − Direct	-0.08008 [-0.08422, -0.07593]	-0.06502 [-0.07138, -0.05850]	-0.06502 [-0.07138, -0.05850]
COHERE − STATE	$+0.04939$ [$+0.04218$, $+0.05689$]	$+0.05987$ [$+0.04605$, $+0.07401$]	$+0.05987$ [$+0.04606$, $+0.07401$]
COHERE − Scratch	$+0.03257$ [$+0.02386$, $+0.04124$]	$+0.04679$ [$+0.03382$, $+0.06029$]	$+0.04679$ [$+0.03383$, $+0.06029$]
COHERE − Finetuned	-0.00088 [-0.00866, $+0.00738$]	$+0.00981$ [-0.00375, $+0.02340$]	$+0.00981$ [-0.00374, $+0.02340$]

Table 16: **Stable PBMC AUROC contrasts.** Paired differences with 95% identity-bootstrap intervals. Shape contrasts fix the complete model's feasible means; predictors retain their own fitted means.

Contrast	Capped q	Uncapped q	Unadjusted p
Occ. / source Q	$+0.00440$ $[+0.00075, +0.00794]$	$+0.00412$ $[+0.00008, +0.00799]$	$+0.00412$ $[+0.00008, +0.00799]$
Occ. / learned Q	$+0.00430$ $[+0.00070, +0.00780]$	$+0.00393$ $[-0.00008, +0.00783]$	$+0.00393$ $[-0.00007, +0.00783]$
Q / source Occ.	-0.00029 $[-0.00040, -0.00017]$	-0.00041 $[-0.00050, -0.00031]$	-0.00041 $[-0.00050, -0.00031]$
Q / learned Occ.	-0.00039 $[-0.00056, -0.00023]$	-0.00060 $[-0.00076, -0.00043]$	-0.00060 $[-0.00076, -0.00043]$
Joint shape	$+0.00401$ $[+0.00050, +0.00751]$	$+0.00352$ $[-0.00048, +0.00738]$	$+0.00352$ $[-0.00048, +0.00738]$
Interaction	-0.00010 $[-0.00023, +0.00004]$	-0.00019 $[-0.00030, -0.00007]$	-0.00019 $[-0.00030, -0.00007]$
COHERE − Direct	-0.07223 $[-0.07847, -0.06620]$	-0.06377 $[-0.06843, -0.05893]$	-0.06377 $[-0.06843, -0.05893]$
COHERE − STATE	$+0.04515$ $[+0.03466, +0.05617]$	$+0.03255$ $[+0.01723, +0.04782]$	$+0.03255$ $[+0.01723, +0.04782]$
COHERE − Scratch	$+0.03197$ $[+0.02080, +0.04294]$	$+0.03360$ $[+0.02039, +0.04692]$	$+0.03360$ $[+0.02039, +0.04691]$
COHERE − Finetuned	-0.01375 $[-0.02376, -0.00333]$	-0.01699 $[-0.02999, -0.00373]$	-0.01700 $[-0.02999, -0.00374]$

The joint-shape AUPRC contrast changes from $+0.003218$ $[0.000192, 0.006048]$ under capped q to -0.000648 $[-0.004698, 0.003295]$ under uncapped q. The paired change in the contrast, uncapped minus capped, is -0.003866 $[-0.005908, -0.001927]$. Direct transfer remains ahead of COHERE under both rankings, with COHERE-minus-Direct AUPRC differences of -0.08008 and -0.06502. For AUPRC, COHERE exceeds STATE and PerturbDiff scratch under all three stable rankings, whereas its intervals against fine-tuned PerturbDiff include zero. The small PBMC shape gain is score-dependent, whereas the complete-predictor trade-off against direct transfer persists.

Rankings from benchmark DE records. For the original benchmark predictions, we compare the native Cell-Eval significance scores (Appendix A) with rankings derived from the retained DE records. Raw-q ranks the retained BH-adjusted p-values directly, removing the 10^{-10} score-clipping threshold while retaining ties from multiple testing and numerical underflow. The p-rank alternative uses retained unadjusted p-values. These comparisons hold the original prediction populations and real-DE labels fixed for COHERE, STATE, PerturbDiff scratch, direct source transfer, and the four fixed-mean shape interventions. The empirical adapter is reported separately from these paired contrasts. Tables 17 and 18 report paired AUPRC differences with 95% identity-bootstrap intervals under each ranking. Retained unadjusted p-values resolve the Nadig rankings, whereas numerical underflow prevents this comparison on PBMC, as indicated by dashes. The population-based calculations in Tables 14–16 provide stable PBMC p rankings. The primary benchmark evaluation uses the native Float32 FDR-based scores. In the fixed-mean comparison, Occ. denotes the occurrence channel and Q the positive-quantile channel; the channel after the slash is held fixed to its source or learned setting. For baselines, PerturbDiff denotes the main-text scratch model.

Table 17: **Fixed-mean shape effects under alternative DE rankings.** Paired AUPRC differences [95% identity-bootstrap intervals].

Contrast	Native	Raw-q	p-rank
		PBMC	
Occ. / source Q	+0.00350 [0.00052, 0.00643]	+0.00018 [−0.00398, 0.00416]	—
Occ. / learned Q	+0.00346 [0.00056, 0.00635]	−0.00016 [−0.00430, 0.00376]	—
Q / source Occ.	−0.00024 [−0.00038, −0.00011]	−0.00051 [−0.00062, −0.00039]	—
Q / learned Occ.	−0.00028 [−0.00045, −0.00010]	−0.00084 [−0.00102, −0.00066]	—
Joint	+0.00322 [0.00035, 0.00612]	−0.00067 [−0.00478, 0.00323]	—
Interaction	−0.00004 [−0.00022, 0.00014]	−0.00034 [−0.00045, −0.00022]	—
		Nadig	
Occ. / source Q	+0.07727 [0.06965, 0.08551]	+0.08264 [0.07401, 0.09208]	+0.08325 [0.07475, 0.09266]
Occ. / learned Q	+0.07911 [0.07108, 0.08787]	+0.08650 [0.07760, 0.09667]	+0.08732 [0.07826, 0.09752]
Q / source Occ.	+0.00224 [0.00150, 0.00307]	+0.00273 [0.00171, 0.00387]	+0.00270 [0.00167, 0.00382]
Q / learned Occ.	+0.00409 [0.00220, 0.00632]	+0.00659 [0.00387, 0.01002]	+0.00677 [0.00399, 0.01023]
Joint	+0.08135 [0.07316, 0.09001]	+0.08924 [0.08008, 0.09965]	+0.09002 [0.08077, 0.10043]
Interaction	+0.00184 [0.00036, 0.00366]	+0.00386 [0.00169, 0.00672]	+0.00407 [0.00183, 0.00698]

Table 18: **DE-ranking sensitivity against matched baselines.** COHERE minus baseline AUPRC [95% identity-bootstrap intervals] on matched populations and real-DE labels.

Baseline	Native	Raw-q	p-rank
	PBMC		
STATE	+0.04939 [0.04184, 0.05663]	+0.06876 [0.05524, 0.08289]	—
PerturbDiff	+0.03166 [0.02289, 0.04024]	+0.04792 [0.03512, 0.06181]	—
Direct transfer	−0.08008 [−0.08404, −0.07602]	−0.06509 [−0.07144, −0.05850]	—
	Nadig		
STATE	+0.01087 [−0.00172, 0.02431]	+0.02375 [0.00866, 0.04030]	+0.02070 [0.00555, 0.03707]
PerturbDiff	+0.08470 [0.07206, 0.09792]	+0.09914 [0.08420, 0.11602]	+0.09973 [0.08481, 0.11652]
Direct transfer	+0.12176 [0.10960, 0.13448]	+0.13062 [0.11734, 0.14509]	+0.13138 [0.11818, 0.14576]

The matched comparisons in Table 18 preserve the principal dataset-level contrast: direct source transfer retains higher AUPRC than COHERE on PBMC, while COHERE has the highest Nadig AUPRC among these matched methods. Nadig's native COHERE–STATE interval crosses zero, whereas its raw-q and p-rank intervals are positive. On Nadig, COHERE also retains the highest AUROC among these matched methods, but STATE and direct transfer exchange order under the p-value ranking. Thus the score transformation affects some margins and secondary ordering without accounting for the large Nadig occurrence effect.

PBMC baseline score saturation. Table 19 examines the retained per-gene DE records behind the PBMC baseline rows whose native AUPRC values cluster near the real-label prevalence of 0.53935. All methods in this diagnostic have complete predicted DE-record coverage. The 10^{-10} FDR floor maps nearly all genes to the maximum native score, leaving few distinct scores and, for some methods, fully tied queries. This score saturation accounts for near-prevalence AUPRC.

Table 19: **PBMC baseline significance-score saturation.** Coverage, ceiling fraction and unique-score counts are averaged over 62 cytokine identities. Coverage is the fraction of genes with predicted DE records; ceiling is the fraction with predicted FDR at or below 10^{-10}. All-tied reports identities with one native score across all 2,000 genes. AUPRC and AUROC recomputed from retained records reproduce the reported values.

Method	Coverage (%)	Ceiling (%)	Unique scores	All-tied	AUPRC / AUROC
Mean / perturbation	100	99.42	10.21	0/62	0.54159 / 0.50489
Mean / cell type	100	99.90	3.10	0/62	0.53894 / 0.49912
Mean / batch	100	99.97	1.55	41/62	0.53926 / 0.49981
Mean / overall	100	99.90	2.94	1/62	0.53894 / 0.49913
CPA	100	99.92	2.65	2/62	0.53902 / 0.49930
CellFlow	100	99.78	5.31	0/62	0.53980 / 0.50099
scGen	100	99.98	1.47	37/62	0.53925 / 0.49979
scPILOT	100	99.85	4.05	3/62	0.53876 / 0.49874

C.4 Alternative Response Bases

Controlled response-basis comparison. The two empirical adapters use the same training budget, cell ordering, and realization procedure, with identity correction disabled. Figure 4a shows that the shape-aware basis improves DE ranking on both datasets while slightly reducing PDCorr; both aggregate MSE-ratio intervals include one.

Both bases have dimension 128 and use the same eligible training responses. The Mean Basis is obtained from centered mean responses. For the Shape Basis, centered mean, occurrence-logit, and positive-quantile blocks are normalized by family-specific RMS scales. The K quantile blocks receive a joint factor $1/\sqrt{K}$, and the leading right singular vectors of the stacked matrix define the basis. This balances the three response families without increasing adapter capacity. For the PBMC test-set differences, switching from the Mean Basis to the Shape Basis increases the energy-weighted coverage of occurrence variation from 0.139 to 0.255 and that of positive-quantile variation from 0.224 to 0.549. The corresponding increases on Nadig are from 0.589 to 0.831 for occurrence and from

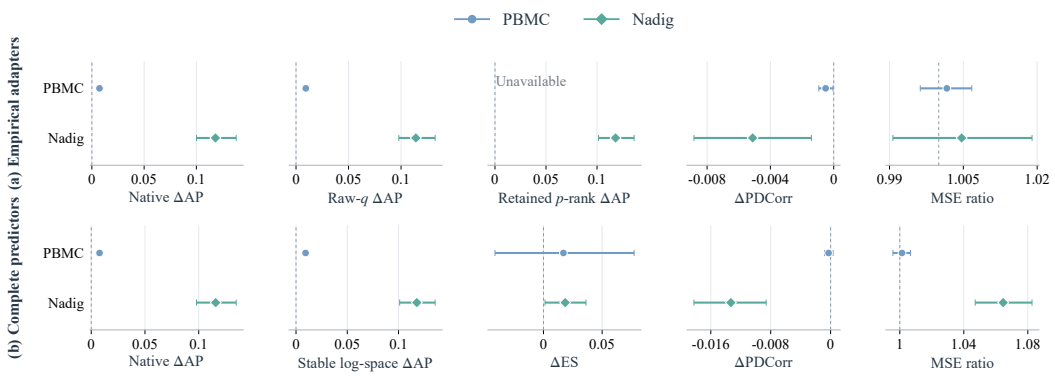


Figure 4: **Response-basis comparisons.** (a) Empirical adapters without identity correction. (b) Complete predictors refitted with separate mean and shape adapters. Points and bars show estimates and paired 95% identity-bootstrap intervals. MSE is the ratio of aggregate MSEs (shape-aware/mean-derived); all other effects are shape-aware minus mean-derived. Dashed lines mark one for ratios and zero for differences. Unadjusted-p rankings use retained values in (a) and direct log-space evaluation in (b); PBMC lacks the retained-p comparison in (a). ES denotes significant-DE-count correlation.

0.636 to 0.839 for positive quantiles. Mean-response coverage decreases slightly. These coverage values describe test-set geometry; performance evidence comes from the retrained comparison in Figure 4a. The mean of per-identity MSE ratios is 1.0062 on PBMC and 1.0142 on Nadig. This estimand differs from the ratio of aggregate MSEs shown in the figure.

Complete-predictor basis comparison. To test whether the representation effect persists through the complete predictor, we fit both the mean-derived and shape-aware 128-dimensional bases with the same training data, losses, budgets, candidate updates, and validation selector. Within each basis, mean and shape predictions use separately fitted adapters. The shape adapters learn all three descriptor families jointly; their mean output is replaced by that of a mean-only adapter before identity correction. Descriptor centers, normalizers, residual targets and scales, identity heads, and shrinkage are fitted for the corresponding basis. Auxiliary fits use the same identity partitions as the original model. The common frozen allocator retains its original ordering representation, isolating the change to the response basis and its dependent predictors. Empirical adapters and identity heads each train for 8,000 updates, with candidates at 2,000, 4,000, 6,000, and 8,000; shrinkage trains for 4,000 updates with candidates at 250, 1,000, 2,000, and 4,000. Checkpoints are selected independently within each basis and stage using equal-weight ranks across the four metric families defined in Appendix B, and both selected configurations are evaluated on the test split. Both variants use separate mean and shape paths. For queries without source observations, they bypass the learned adapters and identity correction and pass the zero source descriptor to the common realization procedure. The paired contrast tests basis choice under a common architecture and policy; original shared-adapter COHERE is a separate reference.

Figure 4b shows that the shape-aware basis improves native and stable-score test AP on both datasets. On Nadig, the gain accompanies lower PDCorr and higher MSE; PBMC's PDCorr interval includes zero and its MSE-ratio interval includes one. These results show how basis choice changes the balance between mean accuracy and DE ranking.

Table 20: **Complete-predictor basis comparison under benchmark scoring.** Both bases use separately refitted mean and shape adapters; original shared-adapter COHERE is shown for reference.

Method	DEOver↑	DEPrec↑	ES↑	DirAgr↑	LFCSpear↑	AUPRC↑	AUROC↑	PDCorr↑	MSE↓	MAE↓	PDS_1↑	PDS_2↑	PDS_{cos}↑
PBMC													
Mean basis	**0.59785**	0.59073	0.85260	**0.85779**	**0.71704**	0.64305	0.64519	0.89552	**9.184e-05**	**0.00382**	**0.98543**	**0.97763**	0.98959
Shape-aware basis	0.58634	**0.59561**	**0.86961**	0.84085	0.67215	**0.65075**	**0.65338**	0.89525	9.197e-05	0.00389	**0.98543**	**0.97763**	0.99011
COHERE	0.59725	0.59087	0.85364	0.85700	0.71584	0.64345	0.64563	**0.89572**	9.221e-05	**0.00382**	**0.98543**	0.97711	**0.99063**
Nadig													
Mean basis	0.25555	0.21251	0.78721	0.88229	0.70391	0.33225	0.77039	0.47043	**0.00660**	**0.05595**	**0.84232**	**0.86392**	0.74679
Shape-aware basis	**0.28797**	**0.25972**	**0.80575**	0.87562	0.69966	**0.44803**	**0.78938**	0.45711	0.00702	0.05731	0.83445	0.85639	**0.75409**
COHERE	0.24852	0.20733	0.78167	**0.88607**	**0.71128**	0.33193	0.76812	**0.47792**	0.00661	0.05598	0.84019	0.86208	0.74600

Table 21 reports validation performance for the selected checkpoint combinations. On PBMC, the mean-derived basis has higher AUPRC and PDCorr and lower MSE. On Nadig, it improves mean accuracy and direction agreement, whereas the shape-aware basis improves DE ranking and perturbation discrimination. The trade-off between mean accuracy and DE ranking on Nadig therefore appears in both validation and test results.

Table 21: **Validation performance of the complete-predictor basis variants.** Both variants use separate mean and shape adapters. Values are from the checkpoint combination selected within each basis. Best values within each dataset are bold.

Basis	DEOver↑	DEPrec↑	DirAgr↑	AUPRC↑	AUROC↑	PDCorr↑	MSE↓	MAE↓	PDS_1↑	PDS_2↑	PDS_{cos}↑
PBMC											
Mean-derived	**0.70367**	0.67943	**0.85903**	**0.74212**	**0.66886**	**0.90251**	**1.565e-04**	**0.00465**	**1.00000**	**1.00000**	**1.00000**
Shape-aware	0.69130	**0.68263**	0.84146	0.73912	0.66614	0.90225	1.602e-04	0.00481	**1.00000**	**1.00000**	**1.00000**
Nadig											
Mean-derived	0.24869	0.19760	**0.84571**	0.28611	0.73826	**0.46928**	**0.00645**	**0.05568**	0.83556	0.85750	0.77667
Shape-aware	**0.26533**	**0.23031**	0.84416	**0.37121**	**0.75257**	0.46390	0.00676	0.05690	**0.84111**	**0.86500**	**0.78250**

Stable ranking for basis comparisons. The complete-predictor refits and COHERE are additionally compared using the stable unadjusted-p ranking defined in Appendix C.3. Controls, genes, populations and real-data BH labels are matched within this basis comparison: 235,478 controls on PBMC and 4,976 on Nadig, with 2,000 genes each. All 62 PBMC identities and 320 of 380 Nadig identities have defined AP; paired intervals resample the full identity universe and average finite values. Stable AP for the mean-basis refit, shape-aware refit, and COHERE is 0.69367, 0.70300, and 0.69409 on PBMC, and 0.35166, 0.46952, and 0.34894 on Nadig.

C.5 Recipient-Adaptation Control

This comparison isolates recipient adaptation in the empirical pathway: both settings omit the identity residual and use the same source descriptors, target-control restoration, and realization procedure.

Table 22: **Recipient adaptation with the identity residual disabled.** Both predictions share source descriptors, target-control restoration, and the realization procedure. Mean accuracy improves on both datasets, while DE ranking improves only on Nadig. Paired intervals appear in Appendix G.

Dataset	Prediction	PDCorr↑	MSE↓	AUPRC↑
PBMC	Direct source transfer	0.85218	1.221e-04	0.72353
PBMC	Empirical backbone	0.89570	9.232e-05	0.64340
Nadig	Direct source transfer	0.33724	9.388e-03	0.21017
Nadig	Empirical backbone	0.47468	6.614e-03	0.32927

C.6 EQUAL-POSITIVE REALIZATION

The equal-positive intervention preserves COHERE's feasible means, nonzero support, cell ordering, and controls, while setting each positive value for gene g to $N\mu_g/k_g$ when $k_g > 0$; genes with zero feasible mean remain zero. The generated populations preserve the intended nonzero support and remain zero for genes with zero feasible mean across 8,904,000 query–gene pairs on PBMC and 760,000 on Nadig. The source-shape comparator uses source occurrence and source quantiles at the same feasible mean of the complete model. DE comparisons use the retained per-gene test records.

Table 23: **Equal-positive realization and DE recovery.** AUPRC and AUROC are averaged over perturbation identities. Native and raw-q use the definitions in Appendix C.3. All rows share the complete model's feasible means; equal-positive additionally preserves its nonzero support.

Dataset	Shape configuration	Native		Raw-q	
		AUPRC	AUROC	AUPRC	AUROC
PBMC	Source shape	0.64023	0.64162	0.69472	0.66071
	COHERE	0.64345	0.64563	0.69406	0.66423
	Equal positive	0.63881	0.64012	0.68650	0.65731
Nadig	Source shape	0.25057	0.73842	0.25873	0.73868
	COHERE	0.33193	0.76812	0.34797	0.76833
	Equal positive	0.33735	0.77636	0.36046	0.77676

Equal-positive realization increases the full marginal Wasserstein error from 0.02945 to 0.04108 on PBMC and from 0.10538 to 0.15576 on Nadig. Paired error increases are 0.01164 $[0.01146, 0.01182]$ and 0.05039 $[0.04988, 0.05090]$. The Nadig AP gain thus accompanies a loss of distributional fidelity: better significance rankings need not imply more accurate marginals.

C.7 MARGINAL-FIDELITY DIAGNOSTICS

We evaluate all 4,452 PBMC and 380 Nadig test queries, keeping the complete model's feasible means, cell counts, ordering, and the common mean-conserving realization algorithm fixed across the four shape interventions and equal-positive realization. For each gene, Wasserstein-1 distance is the integral of the absolute difference between the complete empirical CDFs, including zero mass, in log1p units. It handles unequal real and generated sample counts without truncating either population. Occurrence MAE includes every gene. Positive-quantile MAE uses 32 midpoint probabilities $(j + 1/2)/32$, $j = 0, \ldots, 31$, on genes with positive support in the real population and all five predictions. For normalized error, each quantile vector is divided by the larger of its own mean and 10^{-8}. Errors are averaged first over genes, then over contexts within each perturbation identity, and finally over identities. Intervals use 2,000 paired identity-bootstrap draws. Common support covers 91.88% of real positive-support genes on PBMC and 99.44% on Nadig.

Figure 5 connects the fixed-mean DE contrasts to realized marginal fidelity. Learned shape lowers occurrence and normalized quantile errors on both datasets; all four paired error intervals lie below zero. Table 24 extends the comparison to full marginals. Joint shape adaptation reduces Wasserstein error by 3.8% on PBMC and 34.3% on Nadig. Learned quantiles further reduce this error at learned occurrence by 0.00019 and 0.00245, respectively, with negative paired intervals. On PBMC, replacing source quantiles with learned quantiles at fixed learned occurrence lowers Wasserstein error but slightly reduces DE AUPRC, separating positive-expression fidelity from significance ranking. For joint learned-minus-source shape, the unnormalized quantile improvement on PBMC has an interval including zero, whereas the normalized shape error decreases. Support omission and inflation are unchanged across the five settings: 0.08123 and 0.34151 on PBMC, and 0.00564 and 0.00117 on Nadig. Omission is measured on real-positive genes and inflation on predicted-positive genes.

Table 24: **Realized marginal fidelity at fixed means.** Lower is better. Wasserstein-1 distance (W_1) and unnormalized quantile MAE use log1p units; normalized quantile MAE is dimensionless. Quantile errors use positive support shared by the real population and all five predictions.

Dataset	Shape	W_1	Occurrence MAE	Quantile MAE	Norm. quantile MAE
PBMC	Source occurrence / source quantiles	0.03062	0.04027	0.50321	0.10339
PBMC	Learned occurrence / source quantiles	0.02964	0.03921	0.50257	0.10335
PBMC	Source occurrence / learned quantiles	0.03054	0.04027	0.50291	0.10185
PBMC	Learned occurrence / learned quantiles	0.02945	0.03921	0.50215	0.10186
PBMC	Equal positive	0.04108	0.03921	0.55790	0.21290
Nadig	Source occurrence / source quantiles	0.16039	0.15071	0.33893	0.11367
Nadig	Learned occurrence / source quantiles	0.10782	0.10053	0.20913	0.11346
Nadig	Source occurrence / learned quantiles	0.15787	0.15071	0.33701	0.10688
Nadig	Learned occurrence / learned quantiles	0.10538	0.10053	0.20549	0.10614
Nadig	Equal positive	0.15576	0.10053	0.28405	0.28232

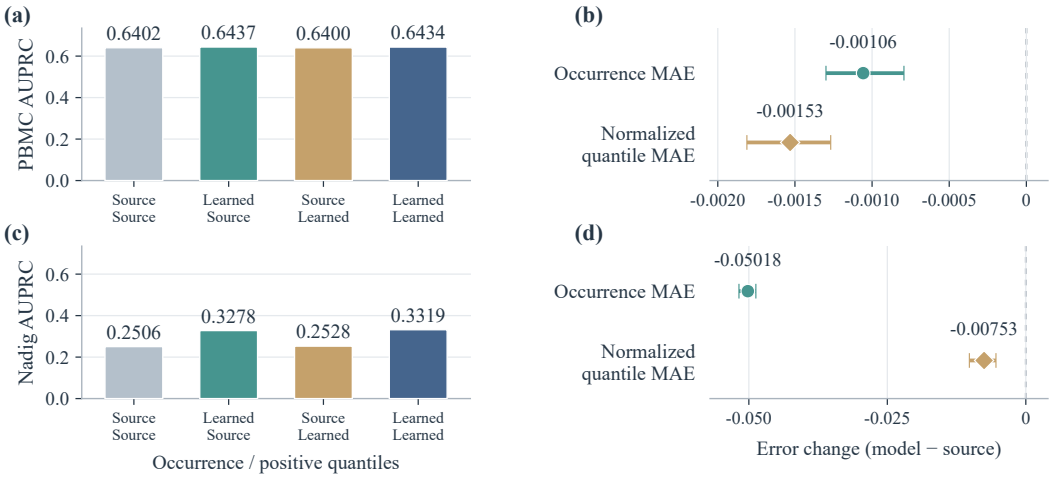


Figure 5: **DE recovery and marginal fidelity at fixed means.** (a,c) AUPRC for the four combinations of source or learned occurrence-logit and positive-quantile descriptors. (b,d) Learned-minus-source changes in realized occurrence MAE and normalized positive-quantile MAE, with paired 95% identity-bootstrap intervals. Panels (a,b) use 62 PBMC cytokines; Nadig uses 320 identities with defined AUPRC in (c) and all 380 identities in (d). Occurrence uses all genes; quantile errors use positive support common to truth and all five predictions, including equal-positive realization. Feasible means and the realization procedure are fixed.

C.8 Joint and Channel-Wise Shape Effects

For metric M, write M_{ab} for source (0) or learned (1) occurrence (a) and positive quantiles (b). The joint contrast is $M_{11} - M_{00}$; the interaction is $M_{11} - M_{10} - M_{01} + M_{00}$. These are fixed-model inference interventions, with the feasible mean, controls, cell count, ordering, and realization procedure held fixed.

Table 25 reports the joint and channel-wise shape effects across all six datasets: source and learned AUPRC, DEPrec, and AUROC, together with individual-channel AUPRC and DEPrec differences. Complete primary-dataset four-corner metrics are reported in Appendix E.3; the external comparison protocol is described in Appendix D.6. Paired uncertainty is reported in Appendix G.

Table 25: **Shape contributions to DE recovery at fixed predicted means.** (a) Source and learned shape (COHERE), with the higher point estimate in each metric pair in bold. (b,c) Learned-minus-source channel effects on AUPRC and DEPrec, with the indicated counterpart fixed (Occ.: occurrence logits; Q: positive quantiles). All interventions retain COHERE's feasible means and realization procedure. Paired AUPRC intervals appear in Appendix G.

(a) Joint shape adaptation across six datasets

	AUPRC↑		DEPrec↑		AUROC↑	
Dataset	Source	Learned	Source	Learned	Source	Learned
PBMC	0.64023	**0.64345**	0.58986	**0.59087**	0.64162	**0.64563**
Nadig	0.25057	**0.33193**	0.17357	**0.20733**	0.73842	**0.76812**
SciPlex3	0.11233	**0.11979**	0.08548	**0.08933**	**0.73434**	0.72148
Frangieh21	0.03242	**0.03679**	0.02383	**0.02471**	0.54189	**0.54522**
OP3	0.23265	**0.25679**	0.15866	**0.16204**	0.82011	**0.82881**
McFaline	0.00941	**0.00986**	0.00978	**0.01038**	0.58445	**0.59291**

(b) Individual channel contributions: ΔAUPRC

Channel / fixed counterpart	PBMC	Nadig	SciPlex3	Frangieh21	OP3	McFaline
Occ. / learned Q	+0.00346	+0.07911	+0.00737	+0.00431	+0.02224	+0.00046
Occ. / source Q	+0.00350	+0.07727	+0.00736	+0.00404	+0.02339	+0.00045
Quant. / learned Occ.	-0.00028	+0.00409	+0.00011	+0.00033	+0.00074	-6.5×10^{-7}
Quant. / source Occ.	-0.00024	+0.00224	+0.00009	+0.00007	+0.00190	-1.9×10^{-6}

(c) Individual channel contributions: ΔDEPrec

Channel / fixed counterpart	PBMC	Nadig	SciPlex3	Frangieh21	OP3	McFaline
Occ. / learned Q	+0.00113	+0.03389	+0.00391	+0.00092	+0.00290	+0.00061
Occ. / source Q	+0.00107	+0.03311	+0.00356	+0.00099	+0.00338	+0.00060
Quant. / learned Occ.	-0.00006	+0.00065	+0.00029	-0.00010	-0.00001	-9.5×10^{-7}
Quant. / source Occ.	-0.00012	-0.00013	-0.00006	-0.00003	+0.00047	-7.1×10^{-6}

C.9 External Benchmark Results

Table 1 contains the primary-dataset comparisons. The following dataset-specific tables report all 13 metrics in each row and use equal-context macro aggregation. As in Table 1, blue shades identify the top three distinct values in each metric before rounding; the darkest shade and boldface indicate the best. Ties share the same rank, and missing values are excluded. MSE and MAE are minimized; all other metrics are maximized. The COHERE method name is bolded for identification. Rank shading in the structural-control tables also uses unrounded values. All results follow the dataset-specific protocols in Appendix A.

Radar visualization. Figure 3 displays eight methods across all 13 metrics, with separate training per dataset. Each axis is min–max normalized over all tabulated methods in that dataset; MSE and MAE are reversed so outward is better. Axis ticks show raw metric values. Missing values remain missing. Radii are not comparable across metrics or datasets, and polygon area is not an aggregate score. The tables retain raw values and all available methods, including the other mean baselines.

C.9.1 SCIPLEX3

Table 26: **SciPlex3 benchmark test results.** Results use equal-context macro averaging.

Method	DEOver↑	DEPrec↑	ES↑	DirAgr↑	LFCSpear↑	AUPRC↑	AUROC↑	PDCorr↑	MSE↓	MAE↓	PDS_1↑	PDS_2↑	PDS_{cos}↑
Mean / perturbation	0.0321	0.0530	-0.1344	0.6202	0.3627	0.0558	0.5728	0.1293	1.43e-02	4.39e-02	0.5800	0.5702	0.5504
Mean / cell type	0.0185	0.0395	-0.4009	0.7916	0.6519	0.0414	0.3720	0.3419	1.06e-03	9.95e-03	0.5174	0.5174	0.5174
Mean / batch	0.0191	0.0414	-0.0809	0.8401	**0.7155**	0.0431	0.4100	0.4710	9.50e-04	9.24e-03	0.6174	0.6072	0.6511
Mean / overall	0.0169	0.0402	-0.3768	0.6445	0.4658	0.0421	0.4032	0.1227	6.63e-03	3.07e-02	0.5174	0.5174	0.5174
CPA	0.0174	0.0420	-0.1044	0.7731	0.5459	0.0451	0.4360	0.3867	1.01e-03	1.10e-02	0.6038	0.5998	0.6254
CellFlow	0.0167	0.0399	-0.2475	**0.8722**	0.7043	0.0416	0.3897	**0.5029**	6.85e-04	9.09e-03	0.6769	0.7083	0.7460
PerturbDiff scratch	**0.0660**	**0.0894**	0.1531	0.6489	0.3053	0.1101	0.7074	0.3032	1.37e-03	1.32e-02	0.7006	0.7157	0.7912
scGen	0.0166	0.0402	-0.3677	0.7751	0.5801	0.0423	0.3839	0.3695	8.80e-04	1.07e-02	0.6185	0.6535	0.7510
scPILOT	0.0171	0.0399	-0.1892	0.7414	0.6020	0.0417	0.3814	0.3031	1.29e-03	1.99e-02	0.6419	0.6727	0.7009
STATE	0.0641	0.0771	0.1718	0.7859	0.5181	**0.1223**	0.7080	0.4882	**6.41e-04**	**8.67e-03**	0.7644	0.7933	**0.8289**
COHERE	0.0645	0.0893	**0.4626**	0.7854	0.5306	0.1198	**0.7215**	0.4807	9.55e-04	1.10e-02	**0.8092**	**0.8279**	0.8154

C.9.2 FRANGIEH21

Table 27: **Frangieh21 benchmark test results.** Results use equal-context macro averaging. Mean/batch is unavailable because this dataset has no technical-batch field, as indicated by dashes.

Method	DEOver↑	DEPrec↑	ES↑	DirAgr↑	LFCSpear↑	AUPRC↑	AUROC↑	PDCorr↑	MSE↓	MAE↓	PDS_1↑	PDS_2↑	PDS_{cos}↑
Mean / perturbation	**0.0368**	0.0115	-0.1325	0.3231	0.1754	0.0124	0.3363	0.0374	1.78e-02	5.59e-02	0.5800	0.5813	0.5787
Mean / cell type	0.0271	0.0121	-0.0832	0.8152	0.8060	0.0120	0.5443	-0.0040	1.31e-02	4.77e-02	0.5077	0.5054	0.4970
Mean / batch	–	–	–	–	–	–	–	–	–	–	–	–	–
Mean / overall	0.0271	0.0121	-0.0832	0.8152	0.8060	0.0120	0.5443	-0.0040	1.31e-02	4.77e-02	0.5077	0.5054	0.4970
CPA	0.0259	0.0118	-0.2189	0.7694	0.7096	0.0113	0.5319	0.1268	1.87e-03	1.94e-02	0.5061	0.5212	0.5158
CellFlow	0.0276	0.0120	-0.0651	0.8920	0.8555	0.0117	0.5483	0.2634	**1.61e-03**	**1.76e-02**	0.5540	0.5763	0.6598
PerturbDiff scratch	0.0265	0.0198	0.0359	0.2281	0.0235	0.0292	0.3782	0.2000	4.41e-03	2.92e-02	0.6068	0.6524	0.7459
scGen	0.0245	0.0120	-0.1012	0.8968	0.8874	0.0117	0.5509	0.1483	2.01e-03	2.04e-02	0.5574	0.5807	0.6430
scPILOT	0.0252	0.0119	-0.1359	**0.9392**	**0.9187**	0.0116	**0.5552**	0.1705	3.01e-03	3.14e-02	0.5595	0.5878	0.6776
TxPert	0.0153	0.0117	-0.0274	0.5186	0.2619	0.0118	0.4610	0.1252	2.78e-03	2.35e-02	0.5151	0.5315	0.5644
STATE	0.0345	0.0166	-0.0704	0.3232	0.1618	0.0188	0.4143	0.2081	1.75e-03	1.84e-02	0.5556	0.5851	0.6669
COHERE	0.0335	**0.0247**	**0.2901**	0.5467	0.3191	**0.0368**	0.5452	**0.2923**	2.70e-03	2.34e-02	**0.6338**	**0.6727**	**0.7549**

C.9.3 OP3

Table 28: **OP3 benchmark test results.** Results use equal-context macro averaging. The official 144-query test set includes six queries with one training cell each; the 138-query sensitivity analysis is in Appendix E.2.

Method	DEOver↑	DEPrec↑	ES↑	DirAgr↑	LFCSpear↑	AUPRC↑	AUROC↑	PDCorr↑	MSE↓	MAE↓	PDS_1↑	PDS_2↑	PDS_{cos}↑
Mean / perturbation	0.0590	0.0878	-0.2837	0.7095	0.1552	0.0807	0.4994	0.1962	1.61e-01	1.44e-01	0.6069	0.5658	0.5631
Mean / cell type	0.0119	0.0694	-0.3152	0.7151	0.4222	0.0586	0.3331	0.2646	8.18e-03	2.91e-02	0.5070	0.5070	0.5070
Mean / batch	0.0119	0.0752	-0.2118	0.6747	0.0811	0.0683	0.4009	0.1509	1.37e-01	1.34e-01	0.5133	0.5092	0.5088
Mean / overall	0.0095	0.0735	-0.3224	0.6735	0.0777	0.0667	0.3850	0.1489	1.38e-01	1.35e-01	0.5070	0.5070	0.5070
CPA	0.0091	0.0733	-0.1343	0.6837	0.1552	0.0655	0.4031	0.2157	1.26e-02	3.97e-02	0.5054	0.5057	0.5069
scGen	0.0088	0.0659	0.0848	0.7926	0.5357	0.0548	0.2821	0.4943	**6.19e-03**	**2.90e-02**	0.6629	0.6723	0.7619
scPILOT	0.0096	0.0694	-0.1917	0.7157	0.4581	0.0588	0.3262	0.3227	1.27e-02	6.39e-02	0.6623	0.6647	0.7147
CellFlow	0.0094	0.0688	**0.3404**	**0.8333**	**0.6437**	0.0576	0.2891	**0.5458**	9.04e-03	3.23e-02	0.7516	0.7656	0.7001
PerturbDiff scratch	0.0611	0.1419	0.0042	0.7429	0.4211	0.1774	0.7230	0.3895	2.05e-02	5.52e-02	0.7703	0.7418	0.6877
STATE	**0.0938**	0.1308	0.1494	0.7881	0.5626	0.1495	0.6556	0.4920	9.39e-03	3.34e-02	0.7775	0.7785	0.7623
COHERE	0.0871	**0.1620**	0.2592	0.8122	0.6092	**0.2568**	**0.8288**	0.5406	1.04e-02	4.20e-02	**0.8061**	**0.7990**	**0.7652**

C.9.4 McFaline

Table 29: **McFaline benchmark test results.** Results use equal-context macro averaging.

Method	DEOver↑	DEPrec↑	ES↑	DirAgr↑	LFCSpear↑	AUPRC↑	AUROC↑	PDCorr↑	MSE↓	MAE↓	PDS_1↑	PDS_2↑	PDS_{cos}↑
Mean / perturbation	0.0151	0.0050	0.0272	0.4000	0.1097	0.0058	0.4521	0.0525	8.24e-03	3.35e-02	0.5088	0.5098	0.5091
Mean / cell type	0.0062	0.0062	0.0298	0.7733	0.7848	0.0064	0.5376	0.0359	1.39e-03	1.43e-02	0.5031	0.5033	0.5025
Mean / batch	0.0066	0.0061	0.0317	0.7962	0.7439	0.0063	0.5295	0.0461	7.64e-03	3.23e-02	0.5032	0.5032	0.5028
Mean / overall	0.0066	0.0061	0.0258	0.7966	0.7437	0.0063	0.5295	0.0460	7.65e-03	3.24e-02	0.5030	0.5029	0.5027
CPA	0.0074	0.0061	**0.0323**	0.9224	0.8688	0.0063	0.5296	0.0950	1.31e-03	1.48e-02	0.5067	0.5081	0.5099
scGen	0.0065	0.0061	0.0323	0.8827	0.8155	0.0064	0.5316	0.1581	7.40e-04	1.21e-02	0.5104	0.5181	0.5299
scPILOT	0.0057	0.0061	0.0003	**0.9532**	**0.8735**	0.0065	0.5429	0.1596	2.47e-03	2.93e-02	0.5157	0.5233	0.5276
TxPert	0.0075	0.0071	0.0276	0.3406	0.3524	0.0066	0.4935	0.1025	1.09e-03	1.43e-02	0.5119	0.5138	0.5256
CellFlow	0.0066	0.0061	0.0125	0.8572	0.7856	0.0064	0.5337	0.1897	7.92e-04	1.22e-02	0.5076	0.5130	0.5170
PerturbDiff scratch	**0.0202**	0.0055	-0.0242	0.1732	0.2307	0.0070	0.4405	0.0974	3.63e-03	2.45e-02	0.5122	0.5161	0.5200
STATE	0.0074	0.0055	-0.0328	0.2132	0.3582	0.0070	0.4427	-0.0112	1.41e-03	1.51e-02	**0.5320**	**0.5420**	0.5429
COHERE	0.0090	**0.0104**	-0.0688	0.3873	0.4380	**0.0099**	**0.5929**	**0.2827**	**5.77e-04**	**1.05e-02**	0.5188	0.5257	**0.5467**

D Complete Component Comparisons

Table 30: **Component interventions on PBMC and Nadig.** Source occurrence, Source quantiles, and Source shape replace their learned counterparts at COHERE's fixed feasible means without retraining. The empirical backbone omits identity correction; direct transfer also bypasses recipient adaptation. All share the realization procedure. Paired intervals appear in Appendix G.

Method	DEOver↑	DEPrec↑	ES↑	DirAgr↑	LFCSpear↑	AUPRC↑	AUROC↑	PDCorr↑	MSE↓	MAE↓	PDS_1↑	PDS_2↑	PDS_{cos}↑
PBMC													
Direct source transfer	**0.638**	**0.619**	0.847	0.856	**0.740**	**0.724**	**0.718**	0.852	1.2e-4	**0.004**	0.964	0.955	**0.991**
Empirical backbone	0.597	0.591	0.854	**0.857**	0.716	0.643	0.646	**0.896**	**9.2e-5**	**0.004**	**0.985**	**0.977**	**0.991**
Source occurrence	0.596	0.590	**0.872**	**0.857**	0.716	0.640	0.641	**0.896**	**9.2e-5**	**0.004**	**0.985**	**0.977**	**0.991**
Source quantiles	0.597	0.591	0.856	**0.857**	0.716	0.644	0.646	**0.896**	**9.2e-5**	**0.004**	**0.985**	**0.977**	**0.991**
Source shape	0.596	0.590	0.870	**0.857**	0.716	0.640	0.642	**0.896**	**9.2e-5**	**0.004**	**0.985**	**0.977**	**0.991**
COHERE	0.597	0.591	0.854	**0.857**	0.716	0.643	0.646	**0.896**	**9.2e-5**	**0.004**	**0.985**	**0.977**	**0.991**
Nadig													
Direct source transfer	0.184	0.146	0.711	0.802	0.564	0.210	0.687	0.337	9.4e-3	0.063	0.686	0.697	**0.775**
Empirical backbone	0.245	0.206	**0.782**	0.884	0.709	0.329	0.766	0.475	**6.6e-3**	**0.056**	0.839	**0.862**	0.751
Source occurrence	0.221	0.173	0.726	**0.886**	**0.711**	0.253	0.744	**0.478**	**6.6e-3**	**0.056**	**0.840**	**0.862**	0.746
Source quantiles	**0.254**	**0.207**	**0.782**	**0.886**	**0.711**	0.328	0.767	**0.478**	**6.6e-3**	**0.056**	**0.840**	**0.862**	0.746
Source shape	0.226	0.174	0.728	**0.886**	**0.711**	0.251	0.738	**0.478**	**6.6e-3**	**0.056**	**0.840**	**0.862**	0.746
COHERE	0.249	**0.207**	**0.782**	**0.886**	**0.711**	**0.332**	**0.768**	**0.478**	**6.6e-3**	**0.056**	**0.840**	**0.862**	0.746

D.1 Empirical Adaptation Variants

We examine the empirical adapter by varying the response families updated and the structure of the recipient map, using the shared empirical backbone as the reference. Mean adaptation with source shape updates only the mean descriptor, while center-plus-affine adaptation omits low-rank mixing. Independent empirical adaptation uses separate factor pairs for mean, occurrence, and positive quantiles, with the five quantiles sharing their family's factors. Table 31 reports all 13 metrics. Direct-transfer and complete-model results for the recipient-adaptation and identity-residual comparisons below appear in Table 30.

On PBMC, mean-response accuracy is similar across these adapted configurations. On Nadig, the shared empirical backbone achieves an AUPRC of 0.329, compared with 0.250 for mean adaptation with source shape and 0.253 for center-plus-affine adaptation. Independent empirical adaptation achieves an AUPRC of 0.332.

Table 31: **Empirical adaptation variants on PBMC and Nadig.** All variants omit identity correction and use the common realization procedure. The shared empirical backbone serves as the reference.

Variant	DEOver↑	DEPrec↑	ES↑	DirAgr↑	LFCSpear↑	AUPRC↑	AUROC↑	PDCorr↑	MSE↓	MAE↓	PDS_1↑	PDS_2↑	PDS_{cos}↑
PBMC													
Empirical backbone	0.597	0.591	0.854	0.857	0.716	0.643	0.646	0.896	9.2e-5	0.004	0.985	0.977	0.991
Mean adaptation with source shape	0.596	0.590	0.874	0.857	0.716	0.640	0.642	0.896	9.2e-5	0.004	0.985	0.977	0.991
Center + affine	0.598	0.591	0.840	0.860	0.720	0.643	0.646	0.895	9.3e-5	0.004	0.985	0.977	0.991
Independent empirical	0.597	0.591	0.865	0.857	0.716	0.643	0.645	0.896	9.2e-5	0.004	0.985	0.977	0.991
Nadig													
Empirical backbone	0.245	0.206	0.782	0.884	0.709	0.329	0.766	0.475	6.6e-3	0.056	0.839	0.862	0.751
Mean adaptation with source shape	0.227	0.171	0.720	0.881	0.707	0.250	0.736	0.471	7.0e-3	0.057	0.842	0.864	0.759
Center + affine	0.175	0.148	0.656	0.833	0.645	0.253	0.707	0.407	8.0e-3	0.060	0.627	0.637	0.601
Independent empirical	0.248	0.211	0.788	0.884	0.713	0.332	0.769	0.476	6.6e-3	0.056	0.840	0.862	0.746

D.2 Recipient Adaptation

Differences are empirical backbone minus direct source transfer, with the identity residual disabled in both settings. Negative favors adaptation for MSE/MAE, positive for the other metrics. Core uncertainty estimates are summarized in Appendix G.

Table 32: **Recipient-adaptation differences.** Empirical backbone minus direct source transfer, with identity correction disabled in both configurations. Lower is better for MSE/MAE and higher for all other metrics.

Dataset	PDCorr	MSE	MAE	PDS-L1	PDS-L2	PDS-cos	DEOver	DEPrec	DirAgr	LFCSpear	AUROC	AUPRC	ES
PBMC	+0.04351	-2.977e-05	-0.0003324	+0.02133	+0.02159	-0.0007804	-0.04089	-0.02851	+0.001116	-0.02401	-0.07228	-0.08012	+0.007157
Nadig	+0.1374	-0.002774	-0.006998	+0.1539	+0.165	-0.02413	+0.06144	+0.06077	+0.08278	+0.1446	+0.07912	+0.1191	+0.07029

D.3 Identity Residual Contribution

The auxiliary identity residual updates the mean-response channel of a frozen empirical backbone. Feasible-mean restoration and realization can consequently change support and positive values. We assess its net contribution by comparing the empirical prediction ($\alpha = 0$) with the complete COHERE model (Table 30). For PBMC and Nadig, complete-minus-empirical AUPRC changes are 4.5×10^{-5} and 2.7×10^{-3}, and PDCorr changes are 2.5×10^{-5} and 3.2×10^{-3}, respectively. The identity-shuffle comparison below separately evaluates identity correspondence.

D.4 Identity Correspondence

The identity shuffle applies a fixed seed-42 permutation per dataset to the identity-to-feature assignment: every eligible identity feature vector is reassigned while the vector multiset, zero and missing entries, response labels, folds, and all frozen empirical inputs are preserved. The permuted assignment is used consistently during training, validation, and test, with the control and missing-prior conventions unchanged. Full and auxiliary identity-residual heads and the shrinkage scalar are retrained with the empirical backbone frozen. This comparison assesses the contribution of identity–feature correspondence after retraining.

The correct-identity reference retains the original identity-to-feature assignment and shares the shuffled model's frozen empirical backbone and evaluation population. For PBMC and Nadig, this reference is the final COHERE model. For all four external datasets, it uses a different empirical-backbone checkpoint from the final benchmark model. The external pairs therefore assess identity correspondence within these fixed configurations, rather than the final external benchmark models. Appendix D.5 reports the empirical and independent-family comparisons for the final models.

Table 33: **PBMC identity-correspondence comparison.** The identity-shuffle branch is retrained with the empirical backbone frozen.

Method	DEOver↑	DEPrec↑	ES↑	DirAgr↑	LFCSpear↑	AUPRC↑	AUROC↑	PDCorr↑	MSE↓	MAE↓	PDS_1↑	PDS_2↑	PDS_{cos}↑
COHERE	0.5973	**0.5909**	0.8536	0.8570	**0.7158**	**0.6434**	**0.6456**	**0.8957**	**9.22e-05**	**3.82e-03**	0.9854	**0.9771**	**0.9906**
Identity shuffle	0.5973	0.5908	0.8539	**0.8571**	**0.7158**	**0.6434**	**0.6456**	**0.8957**	9.23e-05	**3.82e-03**	**0.9857**	0.9768	**0.9906**
Empirical backbone	**0.5974**	0.5908	**0.8542**	**0.8571**	**0.7158**	**0.6434**	**0.6456**	**0.8957**	9.23e-05	**3.82e-03**	0.9854	0.9768	**0.9906**

Table 34: **Nadig identity-correspondence comparison.** The identity-shuffle branch is retrained with the empirical backbone frozen.

Method	DEOver↑	DEPrec↑	ES↑	DirAgr↑	LFCSpear↑	AUPRC↑	AUROC↑	PDCorr↑	MSE↓	MAE↓	PDS_1↑	PDS_2↑	PDS_{cos}↑
COHERE	**0.2485**	**0.2073**	**0.7817**	**0.8861**	**0.7113**	**0.3319**	**0.7681**	**0.4779**	**6.61e-03**	**5.60e-02**	**0.8402**	0.8621	0.7460
Identity shuffle	0.2484	0.2066	0.7805	0.8846	0.7100	0.3307	**0.7681**	0.4765	6.62e-03	**5.60e-02**	0.8398	**0.8622**	0.7450
Empirical backbone	0.2452	0.2063	**0.7817**	0.8844	0.7087	0.3293	0.7663	0.4747	**6.61e-03**	**5.60e-02**	0.8395	0.8619	**0.7507**

Table 35: **SciPlex3 identity-correspondence comparison.** The identity-shuffle branch is retrained with the empirical backbone frozen. Matched reference uses the correct identity assignments on the same frozen backbone, which differs from the final benchmark backbone.

Method	DEOver↑	DEPrec↑	ES↑	DirAgr↑	LFCSpear↑	AUPRC↑	AUROC↑	PDCorr↑	MSE↓	MAE↓	PDS_1↑	PDS_2↑	PDS_{cos}↑
Matched reference	**0.0645**	**0.0893**	0.4626	0.7854	0.5306	**0.1198**	**0.7215**	**0.4807**	**9.55e-04**	**1.10e-02**	0.8092	**0.8279**	0.8154
Identity shuffle	0.0634	0.0890	**0.4742**	**0.7869**	**0.5360**	0.1197	0.7212	0.4795	9.60e-04	**1.10e-02**	**0.8095**	0.8271	**0.8162**

Table 36: **Frangieh21 identity-correspondence comparison.** The identity-shuffle branch is retrained with the empirical backbone frozen. Matched reference uses the correct identity assignments on the same frozen backbone, which differs from the final benchmark backbone.

Method	DEOver↑	DEPrec↑	ES↑	DirAgr↑	LFCSpear↑	AUPRC↑	AUROC↑	PDCorr↑	MSE↓	MAE↓	PDS_1↑	PDS_2↑	PDS_{cos}↑
Matched reference	0.0362	**0.0225**	0.3988	**0.5700**	**0.3464**	0.0344	**0.5487**	**0.2302**	**5.20e-03**	3.09e-02	**0.6302**	**0.6566**	**0.7113**
Identity shuffle	**0.0393**	**0.0225**	**0.3989**	0.5694	0.3456	**0.0345**	0.5484	0.2296	**5.20e-03**	**3.08e-02**	0.6301	**0.6566**	0.7112

Table 37: **OP3 identity-correspondence comparison.** The identity-shuffle branch is retrained with the empirical backbone frozen. Matched reference uses the correct identity assignments on the same frozen backbone, which differs from the final benchmark backbone.

Method	DEOver↑	DEPrec↑	ES↑	DirAgr↑	LFCSpear↑	AUPRC↑	AUROC↑	PDCorr↑	MSE↓	MAE↓	PDS_1↑	PDS_2↑	PDS_{cos}↑
Matched reference	0.0867	**0.1620**	**0.2554**	0.8116	0.6088	0.2565	0.8283	0.5392	**1.04e-02**	**4.19e-02**	**0.8070**	**0.8000**	**0.7656**
Identity shuffle	**0.0876**	**0.1620**	**0.2554**	**0.8123**	**0.6100**	**0.2569**	**0.8292**	**0.5411**	**1.04e-02**	4.20e-02	0.8059	0.7992	0.7647

Table 38: **McFaline identity-correspondence comparison.** The identity-shuffle branch is retrained with the empirical backbone frozen. Matched reference uses the correct identity assignments on the same frozen backbone, which differs from the final benchmark backbone.

Method	DEOver↑	DEPrec↑	ES↑	DirAgr↑	LFCSpear↑	AUPRC↑	AUROC↑	PDCorr↑	MSE↓	MAE↓	PDS_1↑	PDS_2↑	PDS_{cos}↑
Matched reference	**0.0090**	**0.0104**	-0.0688	**0.3873**	0.4380	**0.0099**	0.5929	**0.2827**	**5.77e-04**	**1.05e-02**	**0.5188**	**0.5257**	**0.5467**
Identity shuffle	0.0086	**0.0104**	**-0.0623**	0.3868	**0.4385**	**0.0099**	**0.5930**	0.2826	**5.77e-04**	**1.05e-02**	**0.5188**	0.5255	0.5465

D.5 Shared Factors and Additional Controls

Shared factors provide a compact parameterization for adapting mean, occurrence, and positive-quantile descriptors (Table 51). Independent variants use one low-rank factor pair per family, with the five quantiles sharing their family's factors. Both parameterizations use the same factor rank, source reconstruction, priors, and realization. Each independent complete model refits its backbone, full and five auxiliary identity-residual heads, and global shrinkage coefficient. This backbone supplies new residual labels and scales through the stage-wise identity-holdout procedure (Appendix B).

Independent backbone auxiliary fits use ten identity folds, terminating at 2,000 updates on PBMC, Nadig, Frangieh21, and OP3, and 4,000 on SciPlex3 and McFaline. Nadig compares complete pipelines with different auxiliary empirical fitting budgets: 4,000 updates per fit for shared factors and

2,000 for independent factors. Full PBMC and Nadig identity-residual heads use 8,000 updates with four-candidate validation. External full and all auxiliary identity-residual heads use 2,000 updates. Each independent global scalar uses 4,000 updates and the four-candidate selector described above. All fits use seed 42.

Independent parameterizations produce dataset- and metric-dependent changes across the six benchmarks (Table 39). Differences are independent full minus shared COHERE, with MSE expressed as a relative percentage change. Negative MSE changes favor the independent model. Complete 13-metric results follow.

Table 39: **Independent versus shared complete models.** Differences are independent minus shared, with MSE expressed as a relative percentage change. Equal-rank parameterizations use the auxiliary fitting budgets specified above.

Dataset	ΔPDCorr	MSE change (%)	ΔAUPRC
PBMC	-0.000009	+0.098	-0.000092
Nadig	+0.001275	-0.301	+0.002481
SciPlex3	-0.004871	+12.451	-0.001681
Frangieh21	-0.000770	+3.053	+0.000573
OP3	+0.010738	-2.356	-0.018296
McFaline	-0.000126	+0.016	-0.000001

Center-plus-affine removes low-rank factors from the empirical backbone and has neither identity correction nor global shrinkage. External fits run 8,000 updates and select among checkpoints at 2,000, 4,000, 6,000, and 8,000 updates. Direct source transfer bypasses the recipient operator while retaining raw source detail, target basal restoration and the realization procedure. These are empirical-backbone comparisons, not single-component ablations of the complete predictor.

The unshrunk control freezes the selected shared backbone and identity-residual head and sets $\alpha = 1$ before feasible mean projection. The fixed-mean/source-shape control freezes the full feasible mean, anchors and allocator scores, replacing occurrence and positive quantiles jointly with their source descriptors. Neither inference control refits or selects a new checkpoint.

Table 40: **Shared and independent factors on PBMC.** Empirical predictors omit identity correction; complete predictors include it. Independent full refits the backbone and its identity correction.

Method	DEOver↑	DEPrec↑	ES↑	DirAgr↑	LFCSpear↑	AUPRC↑	AUROC↑	PDCorr↑	MSE↓	MAE↓	PDS_1↑	PDS_2↑	PDS_{cos}↑
Empirical backbone	**0.5974**	0.5908	0.8542	**0.8571**	**0.7158**	**0.6434**	**0.6456**	**0.8957**	9.23e-05	**3.82e-03**	**0.9854**	0.9768	**0.9906**
Independent empirical	0.5972	0.5908	0.8648	**0.8571**	**0.7158**	**0.6434**	0.6454	**0.8957**	9.23e-05	**3.82e-03**	**0.9854**	0.9768	**0.9906**
COHERE	0.5973	**0.5909**	0.8536	0.8570	**0.7158**	**0.6434**	**0.6456**	**0.8957**	**9.22e-05**	**3.82e-03**	**0.9854**	**0.9771**	**0.9906**
Independent full	0.5972	0.5908	**0.8649**	**0.8571**	**0.7158**	**0.6434**	0.6455	**0.8957**	9.23e-05	**3.82e-03**	**0.9854**	0.9768	**0.9906**

Table 41: **Shared and independent factors on Nadig.** Empirical predictors omit identity correction; complete predictors include it. Independent full refits the backbone and its identity correction. Auxiliary fitting budgets differ between the shared and independent pipelines (Appendix D.5).

Method	DEOver↑	DEPrec↑	ES↑	DirAgr↑	LFCSpear↑	AUPRC↑	AUROC↑	PDCorr↑	MSE↓	MAE↓	PDS_1↑	PDS_2↑	PDS_{cos}↑
Empirical backbone	0.2452	0.2063	0.7817	0.8844	0.7087	0.3293	0.7663	0.4747	6.61e-03	5.60e-02	0.8395	0.8619	**0.7507**
Independent empirical	0.2479	0.2110	0.7879	0.8837	0.7132	0.3323	0.7692	0.4760	**6.59e-03**	**5.59e-02**	0.8403	0.8622	0.7462
COHERE	0.2485	0.2073	0.7817	**0.8861**	0.7113	0.3319	0.7681	0.4779	6.61e-03	5.60e-02	0.8402	0.8621	0.7460
Independent full	**0.2555**	**0.2124**	**0.7885**	0.8852	**0.7187**	**0.3344**	**0.7710**	**0.4792**	**6.59e-03**	5.60e-02	**0.8413**	**0.8625**	0.7414

Table 42: **Model variants and inference-time controls on SciPlex3.** Independent-factor and center-plus-affine models are retrained. Unshrunk sets $\alpha = 1$ in the fitted COHERE model; Source shape replaces shape descriptors at COHERE's fixed feasible means.

Method	DEOver↑	DEPrec↑	ES↑	DirAgr↑	LFCSpear↑	AUPRC↑	AUROC↑	PDCorr↑	MSE↓	MAE↓	PDS_1↑	PDS_2↑	PDS_{cos}↑
Empirical backbone	0.0638	0.0892	0.4495	0.7839	0.5250	0.1193	0.7212	**0.4830**	**9.38e-04**	**1.09e-02**	0.8017	0.8301	0.8235
Independent empirical	**0.0661**	0.0885	0.4353	0.7757	0.5236	0.1175	0.7228	0.4786	1.05e-03	1.13e-02	0.8183	**0.8426**	**0.8280**
COHERE	0.0645	**0.0893**	0.4626	0.7854	0.5306	0.1198	0.7215	0.4807	9.55e-04	1.10e-02	0.8092	0.8279	0.8154
Independent full	0.0651	0.0886	0.4387	0.7737	0.5303	0.1181	0.7242	0.4759	1.07e-03	1.15e-02	**0.8235**	0.8415	0.8237
Center + affine	0.0549	0.0818	0.3712	**0.7871**	0.5477	0.0950	0.6685	0.3255	1.02e-03	1.11e-02	0.5464	0.5464	0.6001
Direct source transfer	0.0483	0.0849	0.3157	0.6674	0.3796	0.1085	0.6963	0.1999	1.29e-03	1.19e-02	0.5561	0.5928	0.7027
Unshrunk ($\alpha = 1$)	0.0635	0.0892	**0.4733**	0.7870	**0.5536**	**0.1216**	0.7197	0.4540	1.09e-03	1.17e-02	0.8173	0.8221	0.8019
Source shape	0.0581	0.0855	0.3483	0.7854	0.5306	0.1123	**0.7343**	0.4807	9.55e-04	1.10e-02	0.8092	0.8279	0.8154

Table 43: **Model variants and inference-time controls on Frangieh21.** Independent-factor and center-plus-affine models are retrained. Unshrunk sets $\alpha = 1$ in the fitted COHERE model; Source shape replaces shape descriptors at COHERE's fixed feasible means.

Method	DEOver↑	DEPrec↑	ES↑	DirAgr↑	LFCSpear↑	AUPRC↑	AUROC↑	PDCorr↑	MSE↓	MAE↓	PDS_1↑	PDS_2↑	PDS_{cos}↑
Empirical backbone	0.0339	0.0249	0.2823	**0.5502**	0.3246	0.0369	**0.5483**	0.2886	2.67e-03	2.33e-02	0.6310	0.6734	0.7577
Independent empirical	0.0311	0.0251	**0.2969**	0.5482	0.3178	0.0372	0.5455	0.2907	2.77e-03	2.38e-02	0.6274	0.6704	0.7590
COHERE	0.0335	0.0247	0.2901	0.5467	0.3191	0.0368	0.5452	0.2923	2.70e-03	2.34e-02	0.6338	0.6727	0.7549
Independent full	0.0358	0.0251	0.2942	0.5475	0.3177	**0.0374**	0.5452	0.2915	2.78e-03	2.38e-02	0.6273	0.6705	0.7591
Center + affine	**0.0404**	**0.0277**	0.0398	0.4833	0.2885	0.0351	0.5401	0.2845	2.20e-03	2.06e-02	0.6326	0.6675	**0.8006**
Direct source transfer	0.0315	0.0257	-0.0549	0.4078	0.2282	0.0291	0.4837	0.2974	**1.97e-03**	**1.99e-02**	0.5959	0.6306	0.7599
Unshrunk ($\alpha = 1$)	0.0337	0.0245	0.2433	0.5391	**0.3263**	0.0370	0.5403	**0.3029**	2.83e-03	2.41e-02	**0.6364**	**0.6745**	0.7470
Source shape	0.0350	0.0238	0.2351	0.5467	0.3191	0.0324	0.5419	0.2923	2.70e-03	2.34e-02	0.6338	0.6727	0.7549

Table 44: **Model variants and inference-time controls on OP3.** Independent-factor and center-plus-affine models are retrained. Unshrunk sets $\alpha = 1$ in the fitted COHERE model; Source shape replaces shape descriptors at COHERE's fixed feasible means.

Method	DEOver↑	DEPrec↑	ES↑	DirAgr↑	LFCSpear↑	AUPRC↑	AUROC↑	PDCorr↑	MSE↓	MAE↓	PDS_1↑	PDS_2↑	PDS_{cos}↑
Empirical backbone	0.0791	0.1616	**0.2743**	0.8076	0.6023	0.2548	0.8259	0.5310	1.01e-02	4.14e-02	0.8100	0.8039	0.7648
Independent empirical	0.0834	0.1606	0.2597	0.8098	0.5985	0.2365	0.8215	0.5427	9.80e-03	4.00e-02	**0.8244**	**0.8127**	0.7794
COHERE	0.0871	0.1620	0.2592	0.8122	0.6092	0.2568	0.8288	0.5406	1.04e-02	4.20e-02	0.8061	0.7990	0.7652
Independent full	**0.0908**	0.1608	0.2026	**0.8130**	0.6057	0.2385	0.8250	**0.5513**	1.02e-02	4.07e-02	0.8215	0.8091	0.7793
Center + affine	0.0727	0.1664	0.1379	0.7573	0.4864	**0.2585**	0.8272	0.3936	1.22e-02	4.58e-02	0.7412	0.7158	0.6762
Direct source transfer	0.0747	**0.1708**	-0.0218	0.7748	0.5624	0.2265	0.8035	0.5125	**6.40e-03**	**2.82e-02**	0.6624	0.6823	**0.8496**
Unshrunk ($\alpha = 1$)	0.0899	0.1621	0.2505	0.8125	**0.6132**	0.2581	**0.8306**	0.5459	1.07e-02	4.26e-02	0.8020	0.7938	0.7657
Source shape	0.0760	0.1587	0.2146	0.8122	0.6092	0.2327	0.8201	0.5406	1.04e-02	4.20e-02	0.8061	0.7990	0.7652

Table 45: **Model variants and inference-time controls on McFaline.** Independent-factor and center-plus-affine models are retrained. Unshrunk sets $\alpha = 1$ in the fitted COHERE model; Source shape replaces shape descriptors at COHERE's fixed feasible means.

Method	DEOver↑	DEPrec↑	ES↑	DirAgr↑	LFCSpear↑	AUPRC↑	AUROC↑	PDCorr↑	MSE↓	MAE↓	PDS_1↑	PDS_2↑	PDS_{cos}↑
Empirical backbone	**0.0098**	0.0104	-0.0634	**0.3893**	0.4372	0.0099	0.5936	0.2766	**5.77e-04**	**1.05e-02**	0.5186	0.5248	0.5500
Independent empirical	0.0091	0.0104	-0.0660	0.3890	0.4372	0.0099	0.5935	0.2766	**5.77e-04**	**1.05e-02**	0.5187	0.5247	**0.5501**
COHERE	0.0090	0.0104	-0.0688	0.3873	0.4380	0.0099	0.5929	0.2827	**5.77e-04**	**1.05e-02**	0.5188	0.5257	0.5467
Independent full	0.0096	0.0104	-0.0729	0.3868	0.4367	0.0099	0.5926	0.2825	**5.77e-04**	**1.05e-02**	0.5188	0.5257	0.5468
Center + affine	0.0096	0.0105	-0.0697	0.3885	0.4361	0.0100	**0.5952**	0.2760	**5.77e-04**	**1.05e-02**	0.5184	0.5236	0.5462
Direct source transfer	0.0095	**0.0107**	-0.0670	0.3395	**0.4504**	**0.0102**	0.5878	0.2584	6.06e-04	1.06e-02	**0.5198**	0.5241	0.5395
Unshrunk ($\alpha = 1$)	0.0092	0.0104	-0.0756	0.3867	0.4377	0.0099	0.5921	**0.2853**	**5.77e-04**	**1.05e-02**	0.5190	**0.5260**	0.5451
Source shape	0.0092	0.0098	**-0.0631**	0.3873	0.4380	0.0094	0.5845	0.2827	**5.77e-04**	**1.05e-02**	0.5188	0.5257	0.5467

D.6 External Fixed-Mean Shape Interventions

The external four-corner interventions fix COHERE's predicted mean and allocator scores while replacing occurrence and positive quantiles independently with their source counterparts. Table 25 reports joint shape effects and individual-channel AUPRC/DEPrec contrasts across all six datasets. Paired AUPRC and DEPrec contrasts use context–perturbation pairs with defined values under all four interventions, with the included evaluation contexts held fixed. Mean equality is verified in model units with absolute tolerance 10^{-6} and relative tolerance 10^{-4}. Occurrence improves AUPRC under both quantile choices on all four external datasets; quantile effects are smaller and depend on

the metric and occurrence setting. McFaline's small negative quantile effects and OP3's negative AUPRC interaction remain present. Appendix G summarizes joint AUPRC uncertainty.

E Diagnostic Interpretation and Uncertainty

E.1 Nadig Source Coverage

All 380 identities remain in the main comparison. Table 46 separates 372 source-covered and eight source-free queries; only six source-free queries have defined DE-ranking metrics. Paired intervals resample the identities with defined metrics within each fixed stratum.

Table 46: **Nadig results by source coverage.** Counts distinguish all queries from those with defined DE-ranking metrics. Only six of the eight source-free identities have defined DE metrics.

Source	Method	Identities	Defined DE rankings	PDCorr	AUPRC	ES
covered	Empirical backbone	372	314	0.4774	0.3339	0.7869
covered	COHERE	372	314	0.4801	0.3366	0.7868
covered	Identity shuffle	372	314	0.4794	0.3354	0.7855
source free	Empirical backbone	8	6	0.3476	0.0846	0.2125
source free	COHERE	8	6	0.3766	0.0848	0.0857
source free	Identity shuffle	8	6	0.3408	0.0849	0.2571

COHERE minus empirical AUPRC is $+0.0027$ (95% paired interval $[+0.0021, +0.0033]$; 314 identities with defined paired AUPRC) on source-covered identities and $+0.00015$ ($[-0.00017, +0.00049]$; six such identities) on source-free identities. The source-free interval spans zero.

E.2 OP3 Overlap-Exclusion Sensitivity

The official OP3 test set has 144 queries, six with one cell in the training partition. With training data and fitted models unchanged, we exclude these six queries and compare all four methods on the remaining 138 queries (70 B-cell and 68 myeloid queries). Metrics are aggregated within each context and then averaged equally across contexts; ES is recomputed from the retained real and predicted significant-gene counts. PDS retains the original within-context competitor pool.

Table 47: **OP3 overlap-exclusion sensitivity for selected methods.** The original fitted models are evaluated on the same 138-query subset.

Method	PDCorr↑	MSE↓	AUPRC↑	DirAgr↑	ES↑
COHERE	0.5421	1.01e-02	0.2577	0.8132	0.2653
CellFlow	0.5537	7.32e-03	0.0576	0.8353	0.3994
STATE	0.4957	8.52e-03	0.1490	0.7879	0.1658
PerturbDiff scratch	0.3934	1.73e-02	0.1772	0.7435	0.0023

Table 47 shows that COHERE retains higher AUPRC than PerturbDiff scratch after overlap exclusion (0.2577 versus 0.1772), consistent with the full test-set comparison. CellFlow retains the highest PDCorr and lowest MSE among these four methods, so the trade-off between mean prediction and DE recovery also persists. The five additional controls below use the identical 138-query restriction and retain the full competitor pools for PDS.

Table 48: **OP3 control results after overlap exclusion.** All configurations use the same 138-query restriction. PDS retains its original full within-context competitor pool.

Method	DEOver↑	DEPrec↑	ES↑	DirAgr↑	LFCSpear↑	AUPRC↑	AUROC↑	PDCorr↑	MSE↓	MAE↓	PDS_1↑	PDS_2↑	PDS_{cos}↑
Independent full	**0.0918**	0.1614	0.2019	0.8136	0.6096	0.2390	0.8247	**0.5532**	9.92e-03	4.04e-02	**0.8184**	**0.8058**	0.7759
Center + affine	0.0716	0.1666	0.1423	0.7544	0.4811	0.2586	0.8275	0.3881	1.23e-02	4.60e-02	0.7346	0.7088	0.6685
Direct source transfer	0.0741	**0.1709**	-0.0200	0.7743	0.5619	0.2269	0.8025	0.5096	**6.45e-03**	**2.83e-02**	0.6617	0.6832	**0.8464**
Unshrunk ($\alpha = 1$)	0.0910	0.1628	**0.2580**	**0.8139**	**0.6183**	**0.2591**	**0.8311**	0.5480	1.04e-02	4.21e-02	0.7982	0.7897	0.7613
Source shape	0.0771	0.1592	0.2101	0.8132	0.6134	0.2331	0.8196	0.5421	1.01e-02	4.16e-02	0.8026	0.7951	0.7608

E.3 Fixed-Mean Intervention

Rows list occurrence before positive quantiles; Source and Learned denote unadapted source and learned descriptors, respectively. See Appendix C.7 for marginal-error definitions and the figure.

All four corners retain the complete model's feasible means, controls, cell counts, allocator scores, and realization procedure while changing only the occurrence-logit and positive-quantile inputs, thereby measuring shape-channel effects within the fitted predictor.

Table 49: **PBMC fixed-mean four-corner results.** Rows specify occurrence followed by positive quantiles. Source and Learned denote source and learned descriptors. All configurations retain the complete model's feasible mean, controls, cell count, allocator scores, and realization procedure.

Method	DEOver↑	DEPrec↑	ES↑	DirAgr↑	LFCSpear↑	AUPRC↑	AUROC↑	PDCorr↑	MSE↓	MAE↓	PDS_1↑	PDS_2↑	PDS_{cos}↑
Source / source	0.5955	0.5899	0.8703	**0.8570**	**0.7158**	0.6402	0.6416	**0.8957**	**9.22e-05**	**3.82e-03**	**0.9854**	**0.9771**	**0.9906**
Learned / source	**0.5973**	**0.5909**	0.8561	**0.8570**	**0.7158**	**0.6437**	**0.6460**	**0.8957**	**9.22e-05**	**3.82e-03**	**0.9854**	**0.9771**	**0.9906**
Source / learned	0.5955	0.5897	**0.8725**	**0.8570**	**0.7158**	0.6400	0.6413	**0.8957**	**9.22e-05**	**3.82e-03**	**0.9854**	**0.9771**	**0.9906**
COHERE	**0.5973**	**0.5909**	0.8536	**0.8570**	**0.7158**	0.6434	0.6456	**0.8957**	**9.22e-05**	**3.82e-03**	**0.9854**	**0.9771**	**0.9906**

Table 50: **Nadig fixed-mean four-corner results.** Rows specify occurrence followed by positive quantiles. Source and Learned denote source and learned descriptors. All configurations retain the complete model's feasible mean, controls, cell count, allocator scores, and realization procedure.

Method	DEOver↑	DEPrec↑	ES↑	DirAgr↑	LFCSpear↑	AUPRC↑	AUROC↑	PDCorr↑	MSE↓	MAE↓	PDS_1↑	PDS_2↑	PDS_{cos}↑
Source / source	0.2260	0.1736	0.7283	**0.8861**	**0.7113**	0.2506	0.7384	**0.4779**	**6.61e-03**	**5.60e-02**	**0.8402**	**0.8621**	**0.7460**
Learned / source	**0.2544**	0.2067	**0.7824**	**0.8861**	**0.7113**	0.3278	0.7668	**0.4779**	**6.61e-03**	**5.60e-02**	**0.8402**	**0.8621**	**0.7460**
Source / learned	0.2211	0.1734	0.7255	**0.8861**	**0.7113**	0.2528	0.7439	**0.4779**	**6.61e-03**	**5.60e-02**	**0.8402**	**0.8621**	**0.7460**
COHERE	0.2485	**0.2073**	0.7817	**0.8861**	**0.7113**	**0.3319**	**0.7681**	**0.4779**	**6.61e-03**	**5.60e-02**	**0.8402**	**0.8621**	**0.7460**

Paired AUPRC contrasts and their uncertainty are summarized in Appendix G.

F Computational Resources

Experiments were conducted using two NVIDIA RTX 5090 GPUs. Cell-Eval used three single-thread workers per evaluation. Table 52 reports summed process durations for the measured PBMC and Nadig stages used to fit COHERE and generate predictions. The measurements include recorded preparation and auxiliary fits but exclude raw preprocessing, prior extraction, some intermediate construction, and evaluation, providing a partial breakdown of computational cost.

Table 51 reports trainable parameters by component. The identity-residual head and dataset-level shrinkage scalar are fitted in separate stages.

Table 51: **Trainable parameter counts by component.** Shared and independent backbones are alternatives. Identity-head counts are separate from the single dataset-level shrinkage scalar. Counts exclude frozen representations and the common allocator.

Dataset	Shared backbone	Independent backbone	Identity head	Shrinkage scalar
PBMC	423088	1268400	179552	1
Nadig	8198	24582	348512	1
SciPlex3	24594	73746	145760	1
Frangieh21	24594	73746	348512	1
OP3	27416	82200	145760	1
McFaline	221346	663714	348512	1

Table 52: **Measured fitting and generation times (seconds).** Values are summed process durations for the measured stages, excluding evaluation. Validation rows sum four generation passes per stage.

Stage	PBMC	Nadig
Response statistics / basis	8907.20	7744.89
Allocator	286.91	267.03
Response-feature construction	35.29	17.79
Empirical full fit	139.30	84.58
Ten empirical auxiliary fits	357.06	407.18
Identity-residual full fit	54.56	48.31
Five identity-residual auxiliary fits	54.25	59.22
Residual-feature construction	7.51	15.93
Global scalar fit	14.65	15.75
Empirical validation generation	878.51	180.82
Identity-residual validation generation	782.61	170.79
Global validation generation	823.17	184.56
Selected test generation	2669.07	232.94

G Uncertainty of Core Comparisons

Paired bootstrap uses perturbation identity as the resampling unit: cytokines for PBMC and perturbed genes for Nadig. External comparisons preserve shared identity weights across contexts and then apply equal-context macro aggregation. We use 2,000 draws with fitted models, evaluation populations, and PDS candidate pools fixed. ES is recomputed from significant-DE counts in each draw. The resulting 95% intervals describe identity-sampling uncertainty conditional on the fitted models and are reported without adjustment for multiple comparisons.

Figure 6 summarizes uncertainty for joint shape adaptation, individual shape channels, and recipient adaptation. All differences are learned or adapted minus source, except the factorial interaction defined in Appendix C.8. Under the benchmark DE scoring protocol, joint shape adaptation has positive AUPRC intervals across all six datasets (Figure 6a), but recipient adaptation has opposite effects on PBMC and Nadig (panel b). Occurrence-channel effects are positive in both datasets; quantile-channel effects are negative on PBMC and positive on Nadig (panels c,d). For Nadig DE-ranking contrasts, Table 17 resamples identities with defined metrics, whereas Figure 6 resamples all 380 identities and averages the defined values in each draw. The procedures have identical point estimates and 95% interval endpoints that differ by at most 6×10^{-4}; they coincide on PBMC, where all identities have defined metrics.

H Additional Mechanism and Implementation Details

H.1 Retrained Mean–Occurrence Simplification

The simplified model jointly learns mean and occurrence transformations while retaining source positive quantiles, the fixed mean-derived basis, normalizers, allocator, and realization procedure. It omits quantile adaptation and identity correction, including the residual-head and shrinkage fits. The

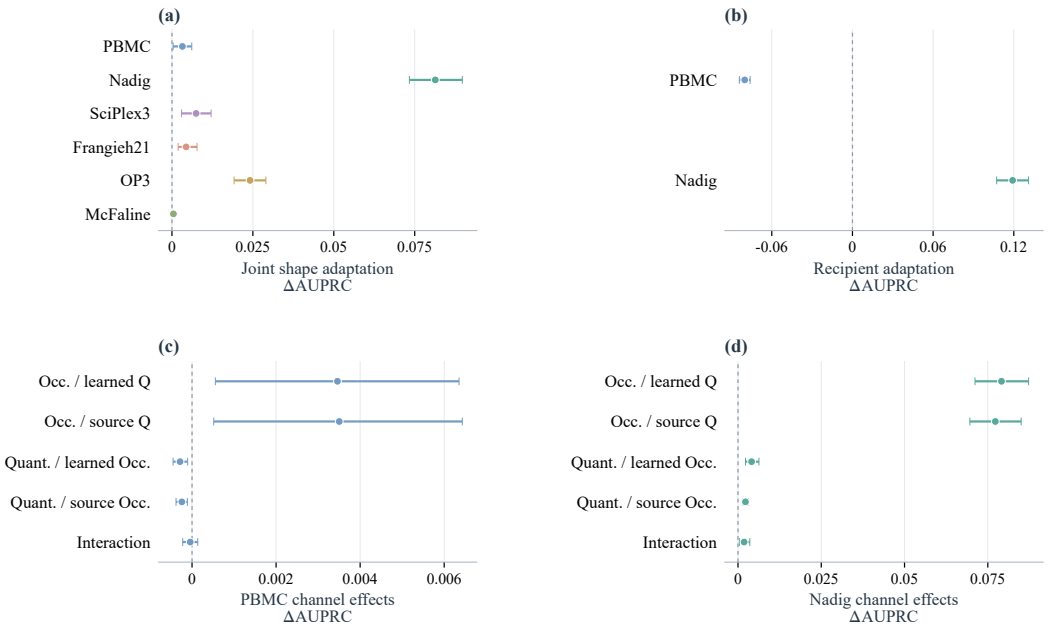


Figure 6: **Core mechanism contrasts under the benchmark DE scoring protocol.** Points and bars show paired AUPRC differences and 95% identity-bootstrap intervals. (a) Joint learned-minus-source shape at fixed complete-model means across six datasets. (b) Recipient adaptation minus direct transfer, with identity correction disabled. (c,d) Occurrence and positive-quantile channel effects and their interaction on PBMC and Nadig. Q denotes positive quantiles and Occ. occurrence; the channel after the slash is held fixed. Panels use different horizontal scales; dashed lines mark zero.

retained losses are $(\mathcal{L}_\delta/\nu_\delta + \mathcal{L}_\pi/\nu_\pi)/3$, preserving the full model's loss scaling and denominator floors. For each dataset, the simplified model is trained for 8,000 updates with seed 42; checkpoints at 2,000, 4,000, 6,000, and 8,000 updates are evaluated on the validation set. Candidates are ranked within four metric families: mean accuracy (PDCorr, MSE, MAE), perturbation discrimination (the three PDS metrics), DE recovery (DEOver, DEPrec, AUPRC, AUROC), and direction agreement (DirAgr). Family ranks receive equal weight; tied selections favor the earlier update. The selected updates are 2,000 for PBMC, Frangieh21, and OP3, and 4,000 for Nadig, SciPlex3, and McFaline. All six selected predictors are evaluated on test data. Each simplified predictor is compared with the corresponding COHERE model on the same test populations and controls. External results average query-level metrics within each context and then weight contexts equally; ES is computed from significant-gene counts within each context. Paired bootstrap samples retain all context records for each sampled perturbation identity.

Figure 7 shows that simplification has different effects across datasets. On PBMC, the simplified model changes PDCorr by less than 3×10^{-5} and increases AUPRC by 0.00033; on Nadig, AUPRC and PDCorr decrease by 0.00721 and 0.00556, respectively. On SciPlex3, Frangieh21, and McFaline, the paired AUPRC-difference intervals include zero. Mean accuracy nevertheless differs: Frangieh21 MSE increases by 0.7%, and McFaline PDCorr decreases by 0.00609. On OP3, the simplified model has lower AUPRC and PDCorr by 0.00168 and 0.00465, respectively, but achieves a 5.9% reduction in MSE and a 0.07641 increase in ES (Table 53). These results show that the benefits of the complete predictor vary across datasets and metrics. Because both branches are omitted together, this comparison does not separate the contributions of quantile adaptation and identity correction. The simplified adapters have 422,944, 8,196, 24,588, 24,588, 27,408, and 221,292 learned parameters in Figure 7 dataset order, excluding frozen representations and the common allocator.

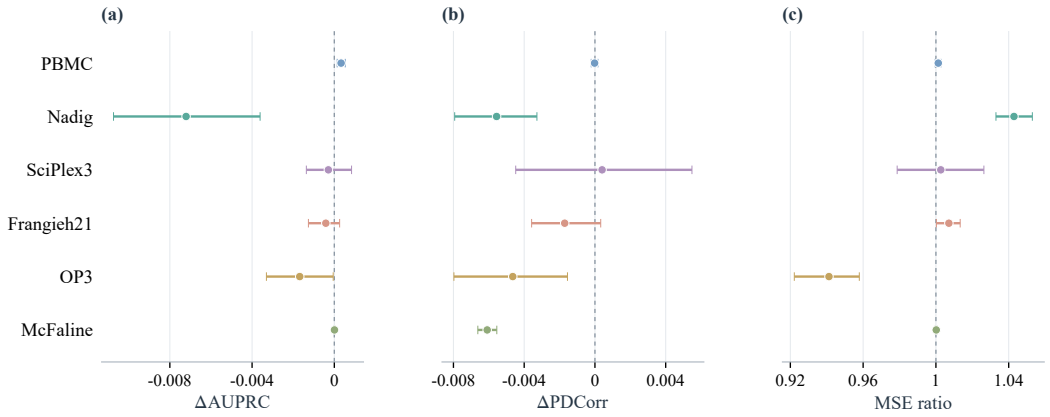


Figure 7: **Retrained mean–occurrence predictors.** Differences from COHERE in (a) AUPRC and (b) PDCorr; (c) aggregate MSE ratios (mean–occurrence/COHERE). Paired 95% identity-bootstrap intervals are shown. Positive differences and ratios below one favor the simplified predictor.

Table 53: **Mean–occurrence predictors versus COHERE: all 13 metrics.** Both models use matched test populations and controls. The better value within each pair is shown in bold; ties at the displayed precision are highlighted for both models.

Method	DEOver↑	DEPrec↑	ES↑	DirAgr↑	LFCSpear↑	AUPRC↑	AUROC↑	PDCorr↑	MSE↓	MAE↓	PDS_1↑	PDS_2↑	PDS_{cos}↑
PBMC													
Mean–occurrence	0.59712	**0.59093**	**0.85870**	**0.85701**	0.71579	**0.64378**	**0.64604**	0.89570	9.233e-05	**0.00382**	**0.98543**	0.97685	**0.99063**
COHERE	**0.59725**	0.59087	0.85364	0.85700	**0.71584**	0.64345	0.64563	**0.89572**	**9.221e-05**	**0.00382**	**0.98543**	**0.97711**	**0.99063**
Nadig													
Mean–occurrence	0.23760	0.20207	0.77172	0.88302	**0.71130**	0.32472	0.76270	0.47236	0.00690	0.05685	**0.84120**	**0.86321**	**0.75821**
COHERE	**0.24852**	**0.20733**	**0.78167**	**0.88607**	0.71128	**0.33193**	**0.76812**	**0.47792**	**0.00661**	**0.05598**	0.84019	0.86208	0.74600
SciPlex3													
Mean–occurrence	**0.06461**	0.08861	0.45177	0.77937	0.52887	0.11950	**0.72389**	**0.48113**	9.577e-04	**0.01094**	0.80835	**0.82913**	**0.82572**
COHERE	0.06455	**0.08933**	**0.46264**	**0.78541**	**0.53059**	**0.11979**	0.72148	0.48073	**9.551e-04**	0.01098	**0.80917**	0.82791	0.81538
Frangieh21													
Mean–occurrence	**0.03390**	0.02463	**0.29959**	**0.54770**	**0.32372**	0.03638	0.54473	0.29060	0.00271	0.02348	0.63086	**0.67354**	**0.76060**
COHERE	0.03354	**0.02471**	0.29011	0.54666	0.31909	**0.03679**	**0.54522**	**0.29231**	**0.00270**	**0.02338**	**0.63377**	0.67274	0.75492
OP3													
Mean–occurrence	0.08041	**0.16230**	**0.33563**	**0.81231**	0.59937	0.25511	0.82362	0.53596	**0.00980**	**0.04042**	**0.81830**	**0.80883**	**0.76808**
COHERE	**0.08714**	0.16204	0.25922	0.81221	**0.60918**	**0.25679**	**0.82881**	**0.54061**	0.01042	0.04200	0.80611	0.79903	0.76518
McFaline													
Mean–occurrence	**0.00953**	**0.01038**	**-0.06472**	**0.38887**	0.43687	**0.00987**	**0.59363**	0.27658	5.770e-04	0.01046	0.51877	0.52476	**0.55009**
COHERE	0.00897	**0.01038**	-0.06885	0.38725	**0.43803**	0.00986	0.59291	**0.28268**	**5.769e-04**	**0.01045**	**0.51878**	**0.52575**	0.54668

H.2 PBMC Negative-Transfer Localization

Validation-set pathway interventions. The following fixed-weight interventions use the empirical adapter without an identity residual, retain source quantiles, and share the same allocator and realization procedure. They compare four combinations of source or adapted mean and occurrence coordinates. The occurrence coordinate is residual logit ℓ: replacing the mean can also change realized support, so these are pathway interventions rather than independent interventions on the generated mean and zero fraction.

Figure 8 shows contrasting adaptation effects: mean adaptation reduces validation AUPRC on PBMC under either occurrence setting, whereas both adapted pathways improve it on Nadig. On PBMC, changing only the mean decreases AUPRC by 0.0904 relative to direct source transfer. The paired interval for this difference is $[-0.0987, -0.0803]$; the interaction interval includes zero. Nadig's interaction is $+0.0301$ $[0.0192, 0.0410]$. The PBMC analysis covers four validation cytokines; on

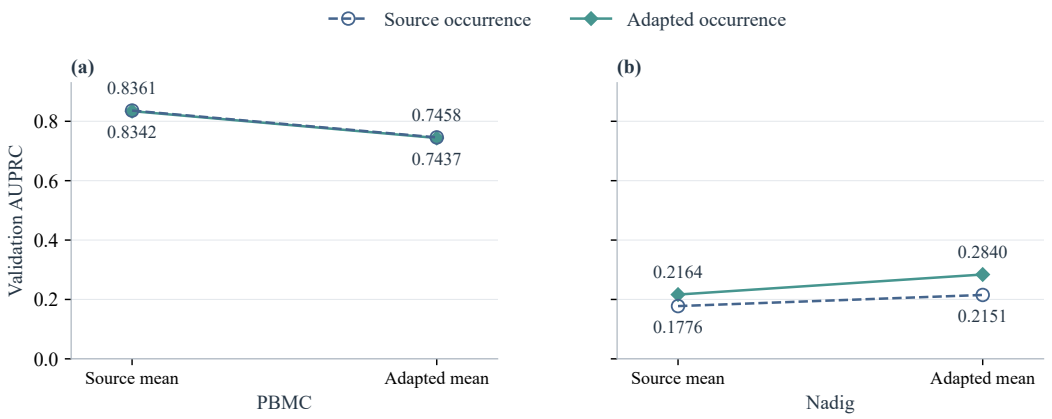


Figure 8: **Mean and occurrence pathway interventions on validation data.** Mean response comes from direct source transfer or recipient adaptation; lines distinguish source and adapted occurrence-logit channels. Source positive quantiles, allocator, and the realization procedure are fixed, and identity correction is disabled. Panels share an AUPRC scale; labels give point estimates.

Nadig, AUPRC is defined for 53 of the 60 validation identities across all four interventions. These validation interventions characterize fitted prediction pathways and complement the test comparisons.

Test-set error localization. A separate analysis of COHERE and direct-source-transfer test outputs covers all 62 PBMC identities. COHERE reduces identity-balanced local mean MSE by 7.78%, but the mean increase in the native significance score is 1.26 on reference non-DE genes versus 0.23 on reference DE genes. The mean within-identity fraction of non-DE genes becoming false positives is 15.15%, compared with 6.05% losing false-positive status. Among reference non-DE genes with finite real and predicted LFCs under both methods, LFC MAE increases from 0.294 for direct transfer to 0.456 for COHERE. These threshold transitions do not decompose AUPRC. They identify increased predicted significance and more DE false positives among reference non-DE genes, rather than only a uniform attenuation of reference DE effects. The empirical adapter and the complete COHERE predictor are nearly identical on PBMC, limiting attribution to the identity residual. Local donor–cell-type error analyses and identity-pooled native DE use different control and aggregation definitions.

DE recovery within donor–cell-type contexts. We also test each PBMC cytokine response against PBS controls from the same donor and cell type, recalculating real BH labels within that context. Contexts require at least 20 real response cells and 20 controls: 3,583 of 4,452 satisfy this condition, and 220 of these have single-class labels and undefined AP. Stable-score AP is averaged across eligible contexts within each cytokine and then across cytokines. COHERE attains 0.26676 versus 0.23599 for direct transfer, with a paired difference of 0.03077 $[0.01756, 0.04369]$. This local diagnostic changes controls, labels, and aggregation relative to the pooled benchmark. Its reversed ordering shows that recipient adaptation can improve within-context DE recovery while reducing pooled DE recovery; it does not isolate which aspect of pooling causes the difference.

H.3 Baseline Configurations and Selection Budgets

The following settings specify the baseline implementations evaluated in this study. Local fitting uses the training split defined in Appendix A, and validation selects within each model's predefined candidate set. Source responses, target training panels, and target controls follow the benchmark-specific rules in Appendix A, including the retained OP3 split exception. Learned baselines use processed log1p inputs, with model-specific latent/PCA representations below. Training budgets differ and are reported in their actual units. Seed 42 is used for local fitting and inference; the released PerturbDiff checkpoints are used directly rather than retrained locally.

Table 54: **Baseline architectures and training budgets for the primary benchmarks.** Budgets and checkpoint candidates describe the runs evaluated in this study. Upstream training costs for reused checkpoints are not included in local training budgets.

Model	Implementation / architecture	Budget and validation candidates
Means	Training-group averages; no optimizer	No checkpoint selection; source-free conventions in Appendix A
CPA	cpa-tools 0.8.8, scvi-tools 0.20.3; Gaussian nonvariational model, latent 128, encoder/decoder 3×256, generalized sigmoid dose response with log-transformed doses	20 epochs; candidates 5/10/15/20; selected PBMC 10, Nadig 5
CellFlow	Official OTFM implementation; train-fitted PCA 100; condition 256, time MLP 3×1024, flow 3×2048, decoder 3×4096	500,000 microiterations, accumulation 20 (25,000 optimizer updates); candidates 125002, 250002, 375002, 500000
scGen	Two hidden layers of 800, latent 100, dropout 0.2, KL weight 5×10^{-5}	20,000 batch steps; candidates 1k/5k/10k/20k
scPILOT	Two layers of 800, latent 100, dropout 0.2; exact latent OT	20,000 batch steps (40,000 optimizer calls); same four candidates
TxPert	STRING GATv2 top-20 neighbors, four layers, hidden 128, two heads, latent 64, decoder 512	20,000 steps; same four candidates; full-vocabulary Nadig configuration
STATE	v0.11.1; bidirectional Llama, hidden 768, eight layers, 12 heads, FFN 3072; rotary positional embeddings and key–value caching disabled; one-hot perturbations and batch encoder	100,000 updates; candidates 5k, 10k, 25k, 50k, 75k, 100k; selected PBMC 50k, Nadig 75k
PerturbDiff	Released scratch and finetuned primary-data checkpoints; finetuned generation in 12,626 genes before HVG extraction	No local fitting or checkpoint search; released checkpoints reused

CPA uses batch 2048, autoencoder learning rate 5×10^{-4} and weight decay 10^{-6}. CellFlow uses batch 1024 and Adam learning rate 5×10^{-5}; PCA uses at most 200,000 training cells, and primary-data perturbation features are ESM2. Its selected microiterations are PBMC 375002 and Nadig 500000. scGen uses Adam 10^{-3} and batch 256. scPILOT uses Adam 10^{-3}, weight decay 10^{-6}, optimizer epsilon 0.01, batch 256, and the discriminator-enabled loss with KL weight 0.2, reconstruction weight 1.0, prior feature-matching weight 10^{-3}, and adversarial classification weight 0.05. TxPert uses AdamW 10^{-3}, zero weight decay, batch 64 and dropout 0.2; the evaluated configuration retains basal-conditioned residual prediction. STATE uses 64-cell sets and 16 sets/update, energy loss, Adam 10^{-4}, zero weight decay, no scheduler and clipping at 10.

For scGen/scPILOT/TxPert, the primary-data selector averages validation ranks within numerical, PDS, DE and biological families; the biological family contains DirAgr, LFCSpear and ES. Their external selector instead uses DirAgr alone as the fourth family (11 core metrics), with earlier checkpoints breaking ties. STATE uses the 11-core, four-family selector. The COHERE selector is specified in Appendix B; the baseline selectors retain the model-specific reporting families described here. Primary CPA/CellFlow also use the 13-metric four-family selector (biology includes DirAgr, LFCSpear and ES); external CPA/CellFlow/PerturbDiff use the 11-core selector. Each uses within-candidate metric ranks followed by equal family weighting and an earlier-candidate tie break.

External models are separately trained. STATE retains 100,000 updates and the six-checkpoint schedule on each external dataset. The external scGen and scPILOT models, and TxPert where applicable, use checkpoints at 1,000, 5,000, 10,000, and 20,000 updates. TxPert is evaluated only on genetic-perturbation datasets. CPA uses 88 epochs on SciPlex3 (candidates: 22, 44, 66, 88), 64 on Frangieh21 (16, 32, 48, 64), 49 on OP3 (13, 25, 37, 49), and 15 on McFaline (4, 8, 12, 15). CellFlow is trained for 500,000 microiterations on each external dataset. External PerturbDiff is trained from scratch; the four external candidate schedules are 5k/10k/15k/20k. The training and selection costs of the reused primary PerturbDiff checkpoints are not quantified here; the comparisons are not matched for total training compute.