Single-cell disentangled representations for perturbation modeling and treatment effect estimation

Jianle Sun[1], Petar Stojanov[2], Kun Zhang[1,3*]

[1]Carnegie Mellon University, Pittsburgh, 15213, PA, United States of America.

[2]Broad institute of MIT and Harvard, Cambridge, 02142, MA, United States of America.

[3] Mohamed bin Zayed University of Artificial Intelligence, Masdar City Abu Dhabi, United Arab Emirates.

*Corresponding author(s). E-mail(s): kunz1@andrew.cmu.edu;
Contributing authors: jianles@andrew.cmu.edu;
Petar@broadinstitute.org;

Abstract

Dissecting cell-state-specific changes in gene regulation induced by perturbations is crucial for understanding biological mechanisms. However, single-cell sequencing provides only unmatched snapshots of cells under different conditions. This destructive measurement process hinders the estimation of individualized treatment effects (ITEs), which are essential for pinpointing these heterogeneous mechanistic responses. We develop scDRP, a generative framework that leverages disentangled representation learning with asymptotic correctness guarantees to separate perturbation-dependent and perturbation-independent latent variables via a sparsity regularized β-VAE. Assuming quantile-preserving effects of perturbations conditional on confounders, scDRP performs conditional optimal transport in the latent space to infer counterfactual states and estimate ITEs. Applied to simulated and real single-cell perturbation data, scDRP accurately estimates treatment effects and individual counterfactual responses, revealing cell type-specific functional gene module dynamics. Specifically, it captures distinct cellular patterns under rhinovirus and cigarette-smoke extract exposures, reveals heterogeneous responses to interferon stimulation across diverse immune cell types, and identifies distinct functional module activation in chronic myeloid leukemia cells following CRISPR knockouts targeting different genes. scDRP also

generalizes to unseen perturbation doses and combinations. Our framework provides a principled computational approach to extracting heterogeneous causal relationships from single-cell perturbation data, enabling a deeper understanding of cellular and molecular mechanisms.

Keywords: single-cell perturbation, treatment effect, disentangled representation, optimal transport, counterfactual prediction, quantile matching, biclustering

1 Introduction

Understanding cell-specific heterogeneous responses to chemical or genetic perturbations is crucial to deciphering the principles of life. Advances in single-cell sequencing technologies now enable large-scale profiling of multi-omics features across numerous perturbations, offering an unprecedented opportunity to dissect gene regulatory circuitry in development and disease [1]. However, the measurements consist of unmatched populations of cells under different perturbations; this destructive nature of sequencing experiments precludes direct comparison of a cell's state before and after perturbation, which is formally referred to as the individualized treatment effect (ITE), posing a fundamental challenge for causal inference.

To circumvent this, many prevailing perturbation modeling frameworks, such as scGEN [2] and State [3], focus on predicting population-average responses; although effective at the population level, they obscure treatment heterogeneity, which is crucial in complex systems like organoids, where subpopulations respond differently. Furthermore, typical methods aiming for individual-level prediction often resort to a branch of strong assumptions. CINEMA-OT [4], for instance, assumes that perturbation-induced latent signals are completely independent of confounders (the cell's intrinsic properties such as cell types), and that responses arise from linear combinations, neglecting the complex nonlinear relationships and interactions that govern cellular responses.

In this paper, to address this challenge, we propose a framework of single-cell disentangled representation learning with perturbations (scDRP). Leveraging the properties of genetic perturbations and recent advances in disentangled representation learning, scDRP employs a properly designed sparsity-regularized β-VAE to learn and separate perturbation-dependent and perturbation-independent latent variables. Under the assumption that perturbations exert quantile-preserving effects on the latent variables they influence when conditioning on confounders, our framework enables the identification of ITE. We also discuss the validity of this assumption. Biologically, this assumption implies that a cell's intrinsic relative rank within a homogeneous subpopulation (e.g., within the same cell type, with similar cell states) is preserved after treatments. For instance, a cell that is inherently the most sensitive to the perturbation (e.g., 95th percentile) among its control peers is assumed to remain the most sensitive (95th percentile) relative to the new perturbed population, even as the entire distribution shifts. Specifically, conditional optimal transport (OT) with squared Euclidean

cost, which is conceptually equivalent to quantile matching or monotone rearrangement in one-dimension [5], is performed in the disentangled latent space to generate counterfactual predictions and estimate individual causal responses (Fig.1a).

We demonstrate the strong performance of scDRP in estimating ITEs and predicting counterfactual perturbation responses across both simulated and real single-cell perturbation datasets. Biclustering analysis on the estimated ITE matrix reveals cell type-specific differences in functional gene module (FGM) responses under diverse perturbations (Fig.1b). Specifically, the re-analysis of rhinovirus and cigarette-smoke exposure data resolves distinct, fine-grained cellular patterns. Applying to additional datasets, we further uncover heterogeneous responses to interferon stimulation across diverse immune cell types, and delineate distinct functional module activation in chronic myeloid leukemia cells following targeted CRISPR knockouts. Furthermore, by performing interpolation within the latent and effect space, we enable inference of the effects of unobserved perturbation doses (Fig.1c) as well as combinations of multiple perturbations (Fig.1d), highlighting the model's ability to generalize to unseen scenarios.

Our theoretical analysis and empirical results suggest that scDRP will provide a powerful computational tool for dissecting large-scale single-cell perturbation datasets, advancing our understanding of cellular and molecular mechanisms underlying physiological and disease processes and paving the way for systematic *in silico* perturbation screening and rational therapeutic target discovery.

2 Results

2.1 Overview of the scDRP framework for inferring single-cell treatment effects

To overcome the fundamental limitation that single-cell sequencing cannot capture the same cells across control and perturbed states, we developed scDRP, a deep generative framework designed to infer individual treatment effects from unmatched population data. Our approach is grounded in the biological understanding that a cell's observed transcriptional state is a composite of its intrinsic, stable identity (e.g., cell type, maturation stage) and its dynamic response to external stimuli. As a consequence, if these two components could be computationally disentangled, one could predict how a specific control cell would react to a perturbation by maintaining its intrinsic identity while shifting its response state, effectively creating an *in silico* "counterfactual twin."

To implement this conceptual framework (Fig. 1a), scDRP posits a generative model where the high-dimensional gene expression profile $\mathbf{x}$ is a noisy function of low-dimensional latent representation $\mathbf{z}$ [6]. We structured this latent space into two distinct conditionally independent subspaces: perturbation-invariant components $\mathbf{z}_u$ and perturbation-induced components $\mathbf{z}_d$. Biologically, $\mathbf{z}_u$ captures intrinsic cellular properties (convoluted with confounders $\mathbf{w}$) that remain stable across conditions. In contrast, $\mathbf{z}_d$ captures the variation induced by cellular responses to perturbations, which is jointly affected by the perturbation $\mathbf{a}$ and confounders $\mathbf{w}$.

Technically, we operationalize this using a β-Variational Autoencoder (β-VAE) regularized with a Hilbert–Schmidt Conditional Independence Criterion (HSCIC) [7, 8]

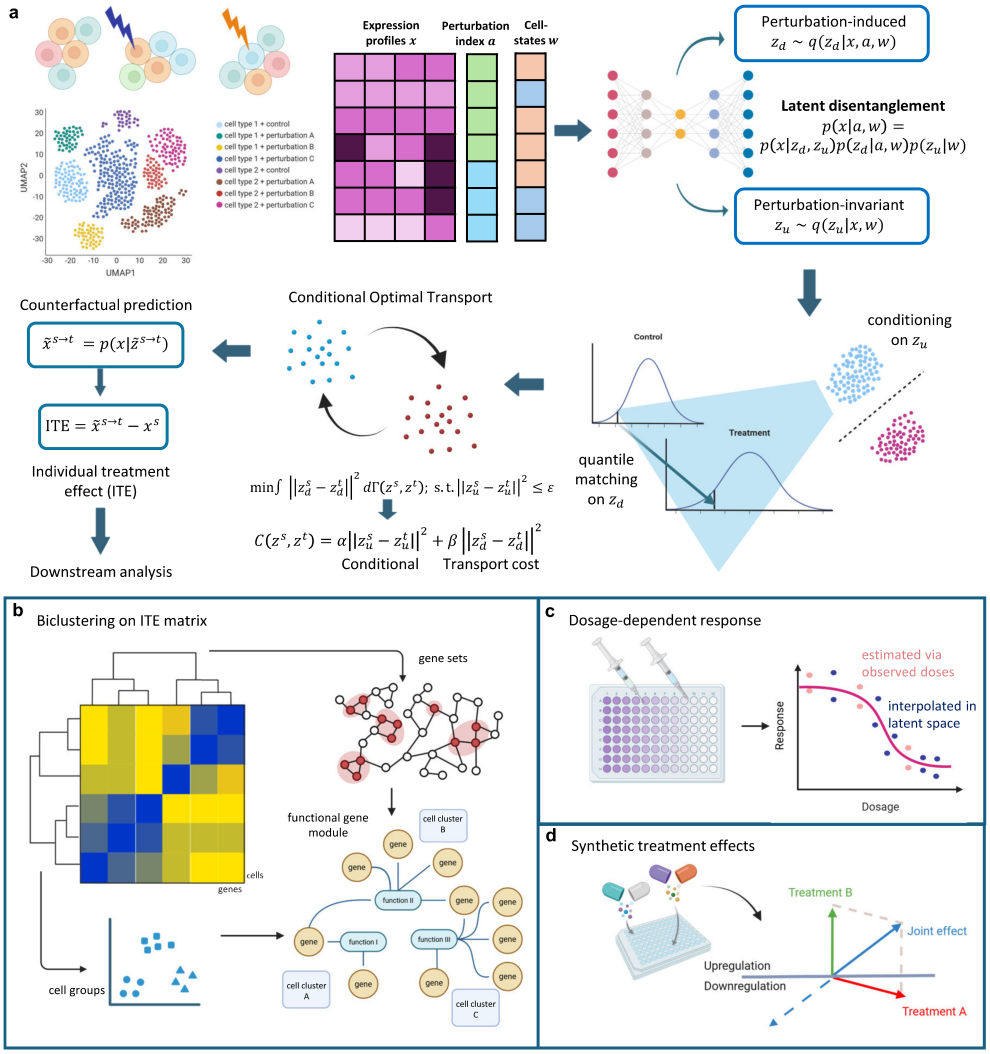


Fig. 1 Model overview. **a.** scDRP disentangles the latent space into perturbation-dependent cell responses ($\mathbf{z}_d$) and perturbation-independent cell identities ($\mathbf{z}_u$). By performing quantile matching on $\mathbf{z}_d$ conditional on $\mathbf{z}_u$, scDRP identifies individualized causal effects and formulates them as a conditional optimal transport problem in the latent space. **b.** Biclustering of the estimated ITE matrix reveals functional gene modules underlying perturbation responses across different conditions and cell populations. **c.** scDRP estimates effects of unseen dosages by interpolation in the latent space. **d.** scDRP infers unseen combinatorial perturbation effects by vector addition of the estimated ITEs.

and gated $\mathcal{L}_0$ norm [9]. This regularization enforces strict conditional independence between $\mathbf{z}_u$ and $\mathbf{z}_d$ given $\mathbf{w}$, ensuring that the decomposd "identity" features do not leak into the "response" space (see Supplementary Notes SN1 for identifiability discussion [10, 11]).

With the disentangled latent space, scDRP estimates ITEs by mapping cells from the control to the treated distribution. A key challenge in causal inference is determining which control cell corresponds to which treated cell. We address this by assuming that perturbations exert rank-preserving effects on the response latent variables $\mathbf{z}_d$ within homogeneous subpopulations (i.e., after controlling for confounders $\mathbf{w}$ or its surrogate $\mathbf{z}_u$). This corresponds to the conservative coupling that minimizes the overall ITEs (i.e., minimal change principle) [12], implying that a cell's relative sensitivity rank or quantile within its peer group is preserved after treatment.

To effectively conduct quantile matching within similar cells across different conditions, scDRP formulates it as a conditional OT problem as OT with a convex cost function is mathematically equivalent to quantile matching in 1-dimension [5]. We seek an optimal coupling matrix Γ^* that matches control and treated cells by minimizing the transport cost in the response space ($\mathbf{z}_d$) while strictly constraining them to share similar intrinsic identities ($\mathbf{z}_u$) by imposing a soft conditional constraint. Mathematically, this is achieved by solving for the coupling that minimizes $\int C(\mathbf{z}^s, \mathbf{z}^t) d\Gamma(\mathbf{z}^s, \mathbf{z}^t)$, where a conditional-penalized cost

$$C\left(\mathbf{z}^s, \mathbf{z}^t\right) = \underbrace{\alpha||\mathbf{z}_u^s - \mathbf{z}_u^t||_2^2}_{\text{soft conditional on identity}} + \underbrace{\beta||\mathbf{z}_d^s - \mathbf{z}_d^t||_2^2}_{\text{transport cost on response}} . \tag{1}$$

Then scDRP can generate counterfactual predictions and estimate ITE for each cell through barycentric projection with optimal coupling Γ^* (Method & Supplementary Notes SN2). We also illustrate the idea of conditional OT in disentangled latent space with a toy example (in Supplementary Notes SN3 & Supplementary Fig.S1-S3). This formulation allows scDRP to move beyond population averages and resolve heterogeneity at single-cell resolution.

2.2 scDRP accurately estimate cell-specific individual treatment effects on different genes

We first compared our approach with two representative existing methods, CINEMA-OT (COT) [4] and scGEN [2]. CINEMA-OT assumes that the mapping from the latent space to the observed expression profiles follows a linear mixing process, and it disentangles latent components based on the strong assumption that the perturbation-related variables z_d are completely independent of confounders. In contrast, scGEN models the mapping from latent factors to observed variables as a nonlinear process, but it does not impose any disentanglement in the latent space; instead, it performs downstream inference based on average changes. Different from both, our model assumes that observed variables are generated as a general nonlinear function of latent variables, while simultaneously enforcing disentanglement in the latent space. Furthermore, we relax the independence assumption to require only that z_d and z_u are

independent conditional on confounders (or their proxies, such as cell type), thereby making our framework more consistent with realistic biological settings.

We generated simulated data comprising five cell types subjected to five perturbation conditions under varying dimensionalities of z_d and z_u. Each setting was repeated 20 times to eliminate random variation. We estimated both the ITE and CATE under two settings: with known cell-type labels (observed confounding surrogates) and with unknown cell-type labels (unobserved confounding surrogates). We evaluated the average root mean square error (RMSE) of ITE and conditional average treatment effect (CATE) across all perturbations and all cell types. As shown in Fig.2a, CINEMA-OT performs poorly on complex nonlinear data, partly because it relies on linear disentanglement, and partly due to its overly strong assumption that confounders are completely independent of the treatment. Although scGEN employs nonlinear neural networks and performs relatively well in CATE estimation through mean-shift–based adaptation instead of latent disentanglement, it performs poorly in ITE estimation. In contrast, our method achieves consistently accurate and stable ITE estimation in both settings.

We discuss different configurations for applying optimal transport and barycentric projection in the disentangled latent space. As shown in Fig.2b, when cell types are known, the model is largely insensitive to hyperparameter choices and projection schemes, as the grouped optimal transport enables effective domain adaptation in the latent space. In contrast, when cell types are unknown, carefully balancing the weights between the conditional and transport cost components becomes more critical.

2.3 scDRP effectively predicts cellular responses to diverse perturbations *in silico* on real datasets

We evaluate the effectiveness of perturbation prediction by comparing the observed real treated samples with the pseudo-treated samples generated through *in silico* counterfactual prediction from the control group. The evaluation metrics include the RMSE and Pearson correlation coefficient (PCC) of the mean expression of each gene between observed treatment group and counterfactual predictions from control group (we compare the mean since the true counterfactual outcome of control cells are not accessible in real data), as well as the maximum mean discrepancy (MMD) between the observed and predicted distributions.

To rigorously evaluate the predictive capability of scDRP, we first assessed its performance using simulated data. Regardless of the accessibility of cell-type labels, scDRP consistently demonstrated robust prediction performance, characterized by significantly lower MMD and RMSE when comparing the predicted and true treated sample distributions, alongside a notably higher PCC (Supplementary Fig.S4).

Furthermore, we proceeded to validate the counterfactual prediction performance on 4 real datasets: Kang et al. [13], Dong et al. [4], Hagai, et al. [14], and Adamson et al. [15]. These datasets encompass both chemical and genetic perturbations, including experiments on pure cell lines subjected to different perturbations, as well as real tissue cells comprising multiple cell types and even multiple species, providing broad representativeness. As shown in Fig.2c (known cell-type setting) and Supplementary Fig.S5 (unknown cell-type setting), our method exhibits consistently superior performance

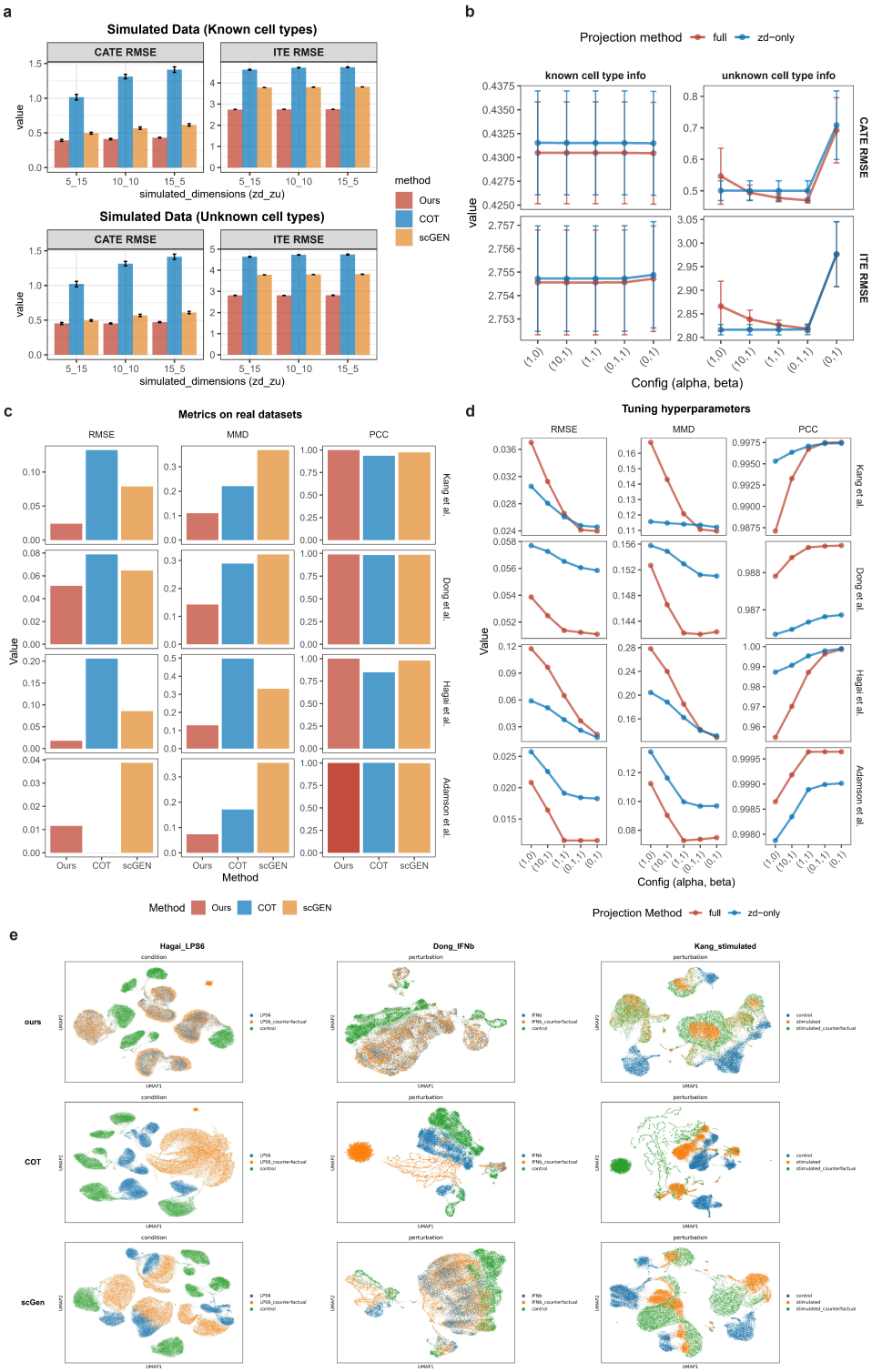


Fig. 2 Performance of effect estimation and counterfactual prediction on simulated and real datasets. **a.** RMSE of CATE and ITE estimation on all genes under known and unknown cell type scenarios. The x-axis shows the dimension settings for z_d and z_u in the simulation, and the different colors correspond to different methods. We repeat each simulated setting 20 times and report the mean±1.96SE. **b.** Sensitivity analysis on hyperparameters of conditional OT on simulated data. **c.** Performance on counterfactual predictions on real single-cell data, where evaluation metrics includes RMSE and PCC of the mean expression of each gene within top 3000 highly variable genes (HVGs) between observed treatment group and counterfactual predictions from control group (we compare the mean since the true counterfactual outcome of control cells are not accessible in real data), as well as MMD between the observed and predicted distributions. All metrics are averaged across all perturbations and cell types. **d.** Sensitivity analysis on hyperparameters of conditional OT on real data. **e.** UMAP visualization of observed and counterfactual predicted cells.

across nearly all datasets. UMAP visualizations of counterfactual pseudo-treated samples generated from control cells (Fig.2e) further confirm this: the counterfactual distributions produced by our method almost completely overlap with the observed treatment distributions, whereas other methods exhibit substantial gaps.

The selection of hyperparameters is also a subtle issue. In real-world data, when explicit cell-type labels are used, it is better to impose no additional condition or a weaker additional condition on $\mathbf{z}_u$ (Fig.2d, Supplementary Fig.S7-S8), while without known cell-type labels, applying a moderate conditional often achieves better results (Supplementary Fig.S6). It is also worth noted that, purely matching by confounding ($\mathbf{z}_u$, i.e., setting $\beta = 0$), similar to the approach adopted by CINEMA-OT, showed very poor results.

2.4 Improved ITE estimation and biclustering of ITEs enable finer-grained capture of heterogeneous response patterns

To demonstrate the behavior of our approach, which features more accurate ITE estimation via non-linear disentanglement and conditional OT, alongside biclustering-based (rather than uni-dimensional clustering) downstream analysis for elucidating the biological heterogeneity of perturbation responses, we analyzed the same scRNA-seq dataset of rhinovirus (RV) and cigarette-smoke extract (CSE) exposure in primary human bronchial organoids as CINEMA-OT [4]. The UMAP visualization demonstrates that our more accurate ITE estimates better preserve the heterogeneity in responses to exposure across different cells (Fig.3a).

Our analysis corroborates previous findings of CINEMA-OT [4] that RV treatment induces heterogeneous, cell-type-specific effects, including the activation of apoptotic pathways and general inflammatory responses. Our more accurate ITE estimation and biclustering approach, however, resolves these effects with greater granularity by identifying distinct functional gene modules (FGMs) (Fig.3b, Supplementary Table 1, Fig.S9). For instance, we specifically localize the intrinsic apoptotic signaling pathway to the Ionocyte-dominated FGM2 module and pinpoint distinct immune profiles, such as natural killer cell mediated cytotoxicity, to the Hillock-specific FGM3 module.

More importantly, our analysis reveals major biological effects of RV treatment that were not previously identified. While the prior study prominently highlighted a type I interferon signature, our ITE biclustering strategy uniquely uncovers two dominant biological programs: 1) a profound ciliary biogenesis signature (cilium organization, motile cilium assembly) localized to the Pre-ciliated/Doublet FGM1 module, and 2) a strong cell-cycle progression signature (chromosome segregation, spindle organization) enriching in both the Cycling basal (FGM8) and Brush+PNEC (FGM9) cells. These findings suggest that our approach successfully disentangles the RV ITEs into more discrete biological processes, revealing that cell-cycle and ciliogenesis perturbations [16, 17] are central, previously unrecognized components of the cell-type-specific response.

Similarly, our analysis confirms that CSE induces a heterogeneous, cell-type-specific response, with notable effects in Basal, Ciliated, and Secretory populations.

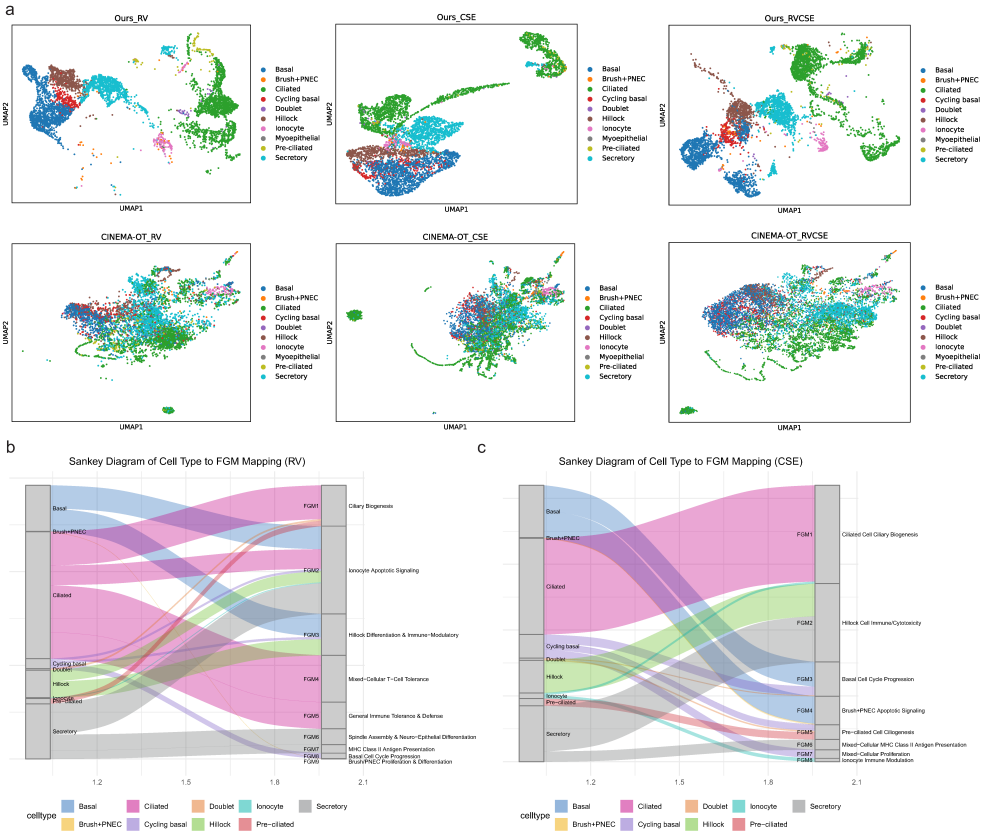


Fig. 3 Biclustering of improved ITEs reveals fine-grained heterogeneous response patterns under RV and CSE exposures. **a.** UMAP visualization of the ITE matrix, clustering cell subgroups by their response profiles across different exposures. **b.** Sankey diagram illustrating the enrichment of FGMs within specific cell types following RV exposure. **c.** Sankey diagram illustrating the enrichment of FGMs within specific cell types following CSE exposure.

Moreover, it uncovers a more complex set of biological programs than previously reported (Fig.3c, Supplementary Table 2, Fig.S10). While prior work highlighted metabolic perturbations, namely Reactive oxygen species pathway, Fatty acid metabolism, and Xenobiotic metabolism, our analysis identifies two dominant, previously unrecognized axes of the CSE ITE.

The first is a profound ciliary biogenesis failure, characterized by the overwhelming enrichment of cilium organization and cilium assembly pathways in both Ciliated (FGM1) and Pre-ciliated (FGM5) cell modules [18]. The second is a strong proliferative response, defined by chromosome segregation and cell cycle G2/M phase transition terms, which we localized to the Basal/Cycling basal compartment (FGM3) [19]. Furthermore, our approach resolves highly discrete immune and stress modules entirely

missed by the prior analysis, including a potent natural killer cell mediated cytotoxicity signature in Hillock (FGM2) and Ionocyte (FGM8) cells, and an intrinsic apoptotic signaling pathway specifically within the Brush+PNEC population (FGM4) [20].

2.5 Biclustering on estimated ITE matrix reveals cell-type specific immune response to interferon stimulation

We then analyzed the heterogeneous responses of different types of human peripheral blood mononuclear cells (PBMC) to interferon IFN-β stimulation [13], which is a potent antiviral and immunomodulatory cytokine that drives a highly coordinated immune program. Visualization of the ITE matrix obtained from scDRP reveals clear clustering of different cell types, demonstrating the heterogeneous effects of IFN-β on different immune cells, while the ITE obtained from CINEMA-OT fails to achieve this (Fig.4a). By further performing biclustering visualization of the ITE matrix obtained from scDRP reveals clear clustering of different cell types, demonstrating the heterogeneous effects of IFN on different immune cells. In contrast, the ITE obtained from CINEMA-OT fails to achieve this. On the ITE matrix stratified by cell type, followed by merging of the resulting biclusters [21], we identified 8 FGMs.

We assigned each cell to the bicluster-derived FGM it belongs to, and then computed the proportion of cells associated with each FGM among each cell type. The results (Fig.4b-c) reveal distinct patterns associated between cell types and FGMs in response to IFN-β stimulation, and we further link them with biological functions through GO enrichment analysis [22] (Fig.4d, Supplementary Table S3 & Supplementary Fig.S11). Effector cells, such as $CD4^+$ T, $CD8^+$ T, and NK cells, uniformly engage FGM3, which relates to antiviral defense and lymphocyte differentiation, confirming the primary role of IFN-β in activating the machinery for viral clearance [23]. The $CD4^+$ T cells are also associated with FGM4, which is related to functions such as protein folding and endoplasmic reticulum stress. This suggests that the activation of T ells places enormous demands on their protein synthesis and secretory pathways, accompanied by intracellular homeostatic stress. This heightened metabolic activity is intrinsically linked to cellular homeostatic mechanisms, including the unfolded protein response (UPR) which manages endoplasmic reticulum (ER) stress, suggesting that as T cells ramp up the synthesis and secretion of cytokines and enhance antigen presentation in response to IFN stimulation, they must also engage pathways that mitigate the intracellular stress caused by the high protein load [24, 25].

Meanwhile, antigen-presenting cells (APCs), including DCs (FGM6) and B cells (FGM7), converge on MHC-II Antigen Presentation programs, highlighting a rapid mechanism to potentiate adaptive immunity. Besides, the IFN-β stimulus links B cells with FGM8 (Immune Cell Migration) as well, indicating it also activates programs related to cell homing and localization [26]. This suggests that IFN-β prepares B cells to rapidly migrate to secondary lymphoid organs, facilitating essential T-B cell interactions necessary for antibody production [27].

Most critically, IFN-β drives divergent functional fates within the monocyte compartment: $FCGR3A^+$ monocytes are skewed toward an activating phenotype, engaging FGM1 (MHC-I/II Presentation and T Cell Regulation), whereas the $CD14^+$ monocytes preferentially upregulate FGM2 (Apoptosis and Metabolic Negative Regulation).

This functional dichotomy suggests that IFN-β employs FGM2 as a crucial negative regulatory mechanism, possibly by limiting the lifespan and inflammatory potential of the CD14$^+$ subset to tightly control the magnitude and duration of the immune response.

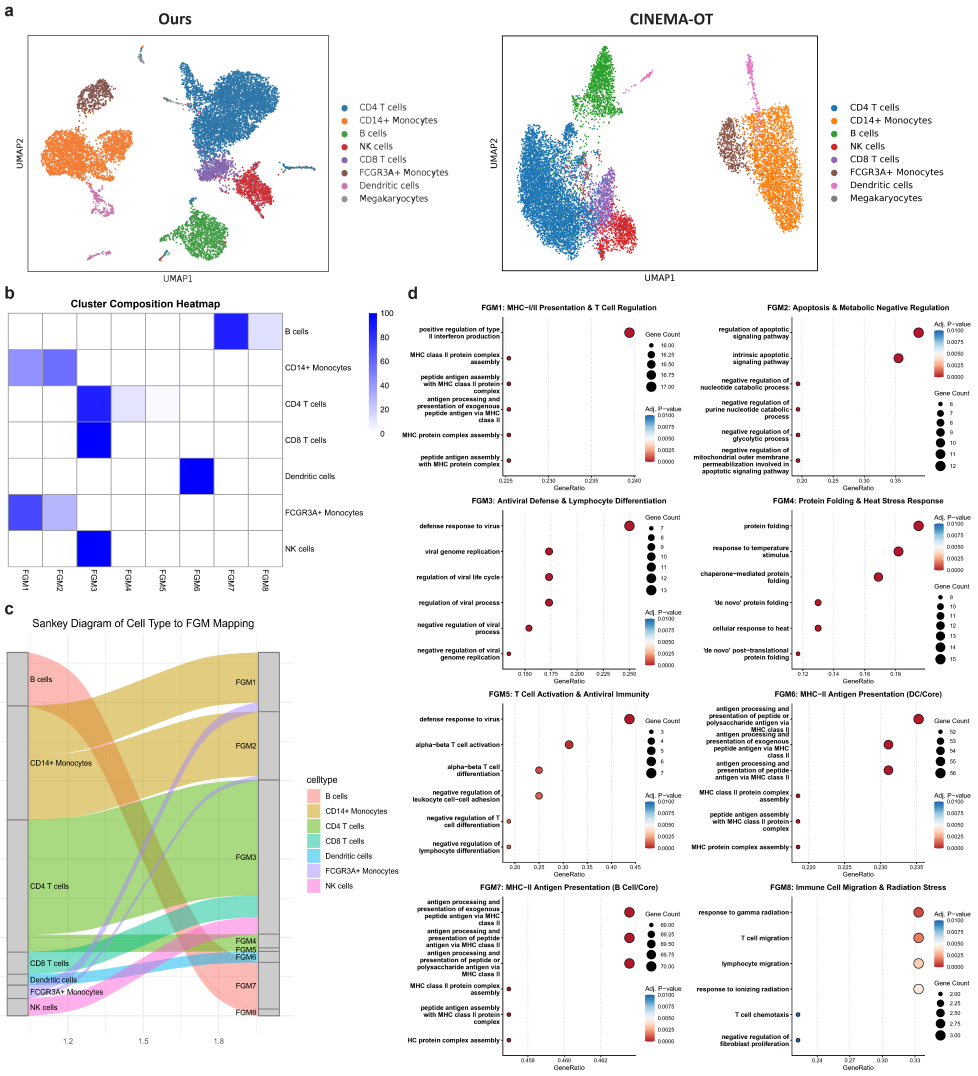


Fig. 4 Cell-type specific immune response to interferon stimulation. **a.** UMAP visualization of ITE matrix estimated via scDRP (ours) and CINEMA-OT. **b.** Heatmap on FGM composition among each cell type. **c.** Sankey diagram of cell type to FGM mapping. **d.** GO enrichment analysis reveals functional roles of each FGM. The summary on functions of each FGM is shown in the subtitle.

2.6 Chronic myeloid leukemia cells exhibit heterogeneous responses to CRISPR perturbations of different genes

Next, we analyzed the effects of eight different single-guide RNAs (sgRNAs) mediated CRISPR perturbations on chronic myeloid leukemia (K562) cells [15]. We estimated the ITE of control-group cells under each perturbation (Fig.5a), performed biclustering of the ITEs for each perturbation separately, and then aggregated the results to obtain seven FGMs.

Results (Fig.5b) show that FGM1 (MHC I antigen presentation and cytotoxicity) is the dominant functional module across almost all eight CRISPR perturbations, comprising 100% of the response for *UBASH3B* and over 75% for several other genes. This strongly confirms that the core response of K562 cells to genetic stress is channeled through the regulation of immune effector functions, apoptosis, and survival. For example, miR-125b targeting of BAK1 in K562 cells has been shown to suppress apoptosis and promote proliferative signals [28].

However, specific genes show a unique divergence into other critical pathways. FGM2, which relates to cell differentiation and oxidative stress, plays a major role, particularly in the perturbations of *BAK1* (53.1%), *OSR2* (48.7%), and *KLF1* (36.5%). KLF1 has been shown to orchestrate almost all aspects of terminal erythroid differentiation in both human and murine systems [29], and in CML leukemic stem cells, shifts in oxidative stress via mitochondrial metabolism are coupled to differentiation when certain regulatory pathways are inhibited (ULK1, etc.) [30]. This suggests these genes are critical nodes linking cell death signaling to the leukemia cell's differentiation status and oxidative tolerance, where their loss may promote a dedifferentiated, proliferative state.

Furthermore, specific transcription factor perturbations reveal distinct functional consequences relevant to the myeloid lineage and immune evasion (Fig.5c, Supplementary Table S4 & Supplementary Fig.S12). CEBPE's disruption, while primarily FGM1, uniquely activates FGM7 (Phagocytosis), reflecting its key role in regulating myeloid functional maturation, especially granulocytic differentiation [31]. The co-perturbation of *CEBPE* and *RUNX1T1* is the only one to engage FGM5 (MHC I Assembly) significantly, providing a direct target to manipulate the immunogenicity of the CML cells. Lastly, the unique link of *ETS2* to FGM4 (Gas Transport and Detoxification) suggests it is a regulatory nexus controlling the leukemia cell's metabolic stability and antioxidant defense. This is consistent with broader findings that ROS regulation and metabolic reprogramming are central in myeloid leukemias, with antioxidant pathways linked to survival, immune escape, and therapy resistance [32, 33], identifying a potential vulnerability for combined metabolic and genetic targeting.

2.7 scDRP predicts dose-sensitive cellular response to chemical perturbations by interpolating on latent space

To assess the model's capacity of generalizing to unseen conditions, we examined cellular responses across a range of chemical perturbation doses. We analysed the sciplex4 dataset [34], which profiles mammary epithelial and alveolar basal epithelial cells treated with seven compounds at multiple doses. To capture dose-specific

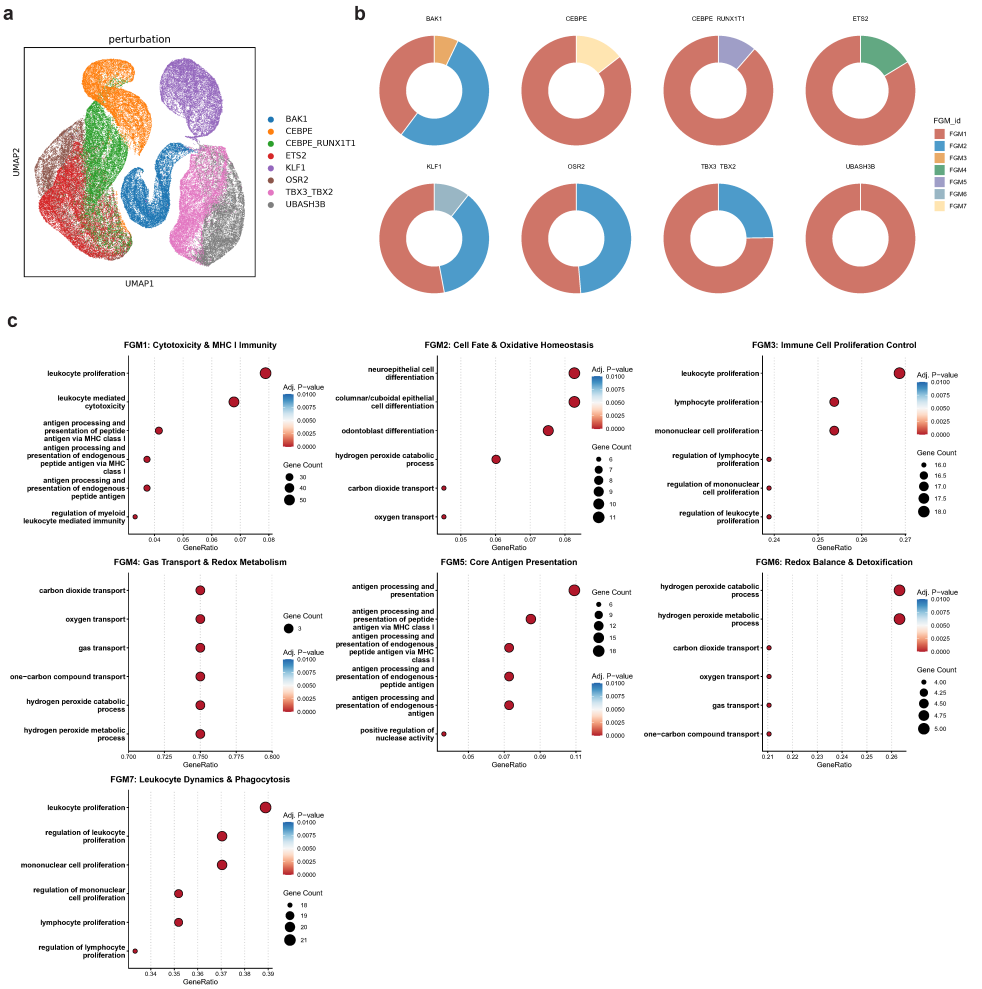


Fig. 5 Heterogeneous responses to CRISPR perturbations. **a.** UMAP visualization of the ITE matrix for control cells under different perturbations. **b.** Proportions of FGMs under different perturbations. **c.** GO enrichment analysis reveals functional roles of each FGM. The summary on functions of each FGM is shown in the subtitle.

effects, we extended the perturbation index $\mathbf{a}$ with an additional dimension encoding dose values v. For doses with observed cells, we inferred the effects by applying optimal transport in the disentangled latent space, a procedure we term Estimation. For unobserved doses, we employ conditional Gaussian process regression to impute the mean and variance of $\mathbf{z}_d$, for unobserved dose values, and apply residual correction to ensure quantile alignment of samples before and after perturbation while keeping $\mathbf{z}_u$ unchanged. The VAE decoder is then used to generate counterfactual samples for

causal effect estimation, which is a strategy we term Extrapolation (see Methods for details).

We estimated the effects of seven chemical perturbations on two cell types in the SciPlex4 dataset using two different methods. For each cell type, we selected the five genes most strongly affected and plotted the CATE–dosage curves (Fig.6a). The interpolated curves closely overlapped with those estimated at observed dose values, demonstrating the strong generalization ability of our method. In contrast, direct estimates at observed doses exhibited greater variability, as they contained more noise, compared to the smoothed interpolation results obtained from the Gaussian process.

We can obtain more detailed information about perturbation effects from the dose–response curves. For example, in alveolar basal epithelial cells, Acetate stimulation leads to a dose-dependent upregulation of *AKR1B1* expression, which may be due to enhanced flux through the glucose-derived polyol pathway under high acetate conditions, increased production of NADPH to buffer oxidative stress, or activation of metabolic reprogramming that favors enzymes like *AKR1B1* [35, 36]. Indeed, *AKR1B1* is known to mediate metabolic reprogramming and enhance stress tolerance in contexts of hyperglycemia and increased energy demand [37].

In contrast, in mammary epithelial cells, stimulation with an ACLY inhibitor exerts a sustained suppressive effect on GFRA1 expression, likely because inhibition of ACLY reduces cytosolic acetyl-CoA supply, thereby limiting lipid and protein acetylation needed for growth factor signaling and transcriptional activation; additionally, decreased ACLY activity can impair metabolic pipelines (e.g. lipogenesis, histone acetylation) [38], crucial for maintaining GFRA1 expression.

2.8 scDRP helps infer synthetic effects of multiple treatments via vector composition on ITE space

To further demonstrate the model's extrapolative ability, we considered the task of inferring the synthetic effects of multiple perturbations given observations of single perturbations. We conceptualize this as a form of vector composition (see Methods). We validated this approach using a dataset [4] of human PBMCs subjected to single and combined stimulations, including IFNβ + TNFα, IFNβ + IFNγ, and IFNβ + IL-6. We compared the absolute CATE of each gene within each cell type using ITEs estimated directly from samples subjected to combined stimulations with those calculated by vector addition of the ITEs from each single stimulation.

As shown in the Fig.6b, one can observe a very strong linear correlation, which demonstrates the effectiveness of our method in inferring combinatorial effects. At the same time, note that our method can only estimate the absolute magnitude of the combined effect, but cannot determine whether the final effect is an upregulation or downregulation of gene expression (as clearly illustrated by the absolute value function effect in the Fig.6c).

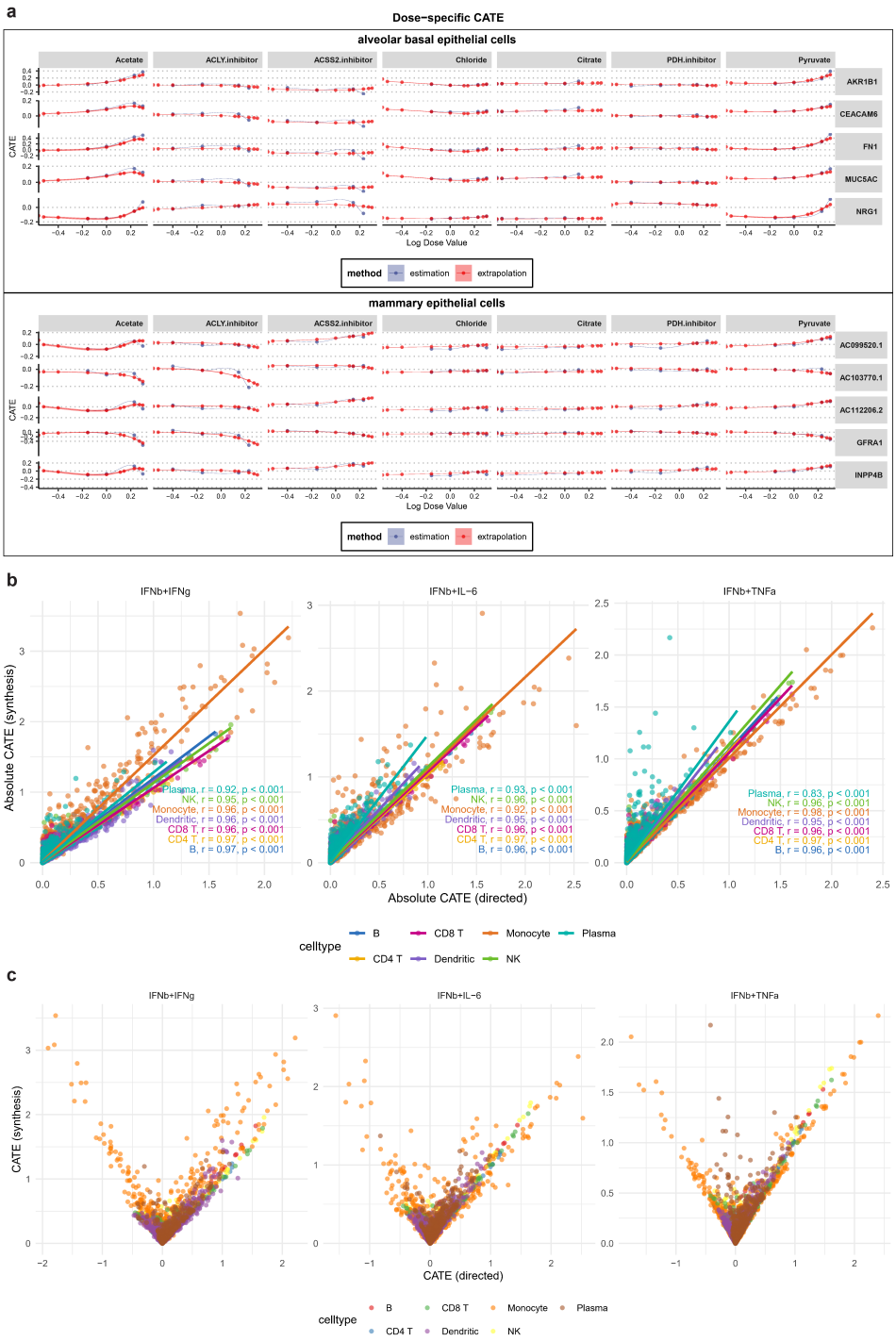


Fig. 6 scDRP generalizes to unseen dosage values and treatment combinations. **a.** Dose-response (CATE) curves estimated via estimation (observed dosages) and extrapolation (latent interpolation) for two cell types and seven perturbations in sciplex4 dataset. **b.** Absolute CATE of combined treatments obtained via direct estimation vs synthetic inference from individual treatments for top 3000 genes. **c.** CATE of combined treatments obtained via direct estimation vs synthetic inference from individual treatments for top 3000 genes.

3 Discussions

Disentangled representation learning has recently received widespread attention to address the non-paired data challenge in single-cell omics [39]. In this study, we introduce scDRP, a framework that enables the estimation of ITEs and counterfactual responses from single-cell perturbation data. By combining disentangled representation learning with conditional optimal transport, scDRP provides a principled way to isolate the true causal impact of perturbations from confounding variation. Our results demonstrate that this formulation not only improves the accuracy of counterfactual prediction but also facilitates interpretable biological insights by linking latent dimensions to functional gene modules. Latent-space interpolation and ITE vector composition illustrate the model's capacity to generalize beyond observed experimental conditions, positioning scDRP as a versatile computational framework for identifying perturbation effects in high-dimensional single-cell systems.

We address the fundamental problem of causal inference that only (at most) one potential outcome is observable for any given cell-through latent space decoupling and counterfactual matching. Given the complexity of biological systems and substantial technical noise, traditional linear models carry a high risk of model misspecification [40]. We therefore opted for a flexible neural network-based approach, which allows us to effectively perform the decoupling while accommodating technical noise and batch effects. We achieve the disentanglement based on the assumption that, given proxies of confounders (such as cell type), the perturbation-dependent and perturbation-invariant latent variables are independent. When cell-type information is unavailable, only partial disentanglement can be achieved. However, by incorporating the distance in $\mathbf{z}_u$ (which encodes confounding information) into the cost function of conditional OT as a soft conditional term, the model retains considerable robustness even in the absence of explicit cell-type labels.

The core idea of counterfactual matching in the latent space stems from the principle that, after controlling for confounders, a treatment may shift the overall population distribution but should not alter the relative position of each individual within the population, that is, the exogenous noise is independent of the treatment, which is commonly adopted in counterfactual inference. This actually corresponds to the Fréchet-Hoeffding upper bound on the joint distribution induced by the conditional marginal distributions of potential outcomes after controlling for confounders, specifying the conservative coupling (i.e., the minimal change assumption) that minimizes the individual causal effects [12]. We implement a soft-constrained conditional OT to jointly perform confounder control ($\mathbf{z}_u$) and quantile matching ($\mathbf{z}_d$), inspired by the equivalence between one-dimensional OT under convex loss and quantile mapping. Exploring alternative estimation strategies, such as independent residual correction or coupled per-dimension quantile matching, represents an important avenue for future work. Future extensions may also incorporate temporal or spatial modalities to further disentangle dynamic regulatory programs and cellular interactions.

While we primarily validated and analyzed scDRP on single-cell RNA-seq perturbation data in this study, the framework can be readily generalized to perturbation datasets from other omics, such as scATAC-seq [41]. We believe that as more advanced perturbation sequencing technologies are developed and more comprehensive datasets

are accumulated across diverse cell and tissue types with richer perturbations, scDRP will serve as a powerful computational tool. It will help elucidate the cellular and molecular mechanisms in biological systems under perturbations at an unprecedented resolution and facilitate the discovery of novel drug targets.

4 Methods

4.1 Disentangled representations with minimal-change principle

Suppose we have the observed gene expression $\mathbf{x}$, perturbation index $\mathbf{a}$, and cell type index $\mathbf{w}$ (optional), as well as some technical covariates like sequencing batch index $\mathbf{c}$ (optional). We assume observed $\mathbf{x} \in \mathbb{R}^p$ is generated from latent gene programs $\mathbf{z} \in \mathbb{R}^d$ ($d \ll p$) and that only a subset of $\mathbf{z}$ is affected by the perturbation. Under the assumption that $\mathbf{z}_d$ are $\mathbf{z}_u$ are conditionally independent given confounding $\mathbf{w}$, we employee the following VAE architecture to learn the disentangled representations of gene expressions.

- Generative network: We assume the generating process: $p(\mathbf{x}) = p(\mathbf{z})p(\mathbf{c})p(\mathbf{x}|\mathbf{z},\mathbf{c})$, where $p(\mathbf{z}) = \prod p(\mathbf{z}_d|\mathbf{a},\mathbf{w})p(\mathbf{z}_u|\mathbf{w})$, $p(\mathbf{z}_d|\mathbf{a},\mathbf{w}) \overset{d}{=} \mathcal{N}(\boldsymbol{\mu}_{aw}, \boldsymbol{\sigma}^2_{aw}\mathbf{I}_{n_d})$ and $p(\mathbf{z}_u|\mathbf{w}) \overset{d}{=} \mathcal{N}(\boldsymbol{\mu}_w, \boldsymbol{\sigma}^2_w\mathbf{I}_{n_u})$.
  There are flexible choices to specify $p(\mathbf{x}|\mathbf{z},\mathbf{c})$ and reconstruction loss (log likelihood) accordingly. For example, (1) if the log-normalized data are provided as $\mathbf{x}$, we can use a half Gaussian distribution $|\mathcal{N}(f(\mathbf{z},\mathbf{c}),\Sigma)|$ and use mean squared error (MSE) loss as the reconstruction loss; (2) if the raw count data are provided, we may assume that $p(\mathbf{x}|\mathbf{z},\mathbf{c})$ follows a zero-inflation negative binomial (ZINB) model [42] $x_{ng} \sim \pi_g\delta_0 + (1-\pi_{ng})\mathrm{NB}(l_n\rho_{ng}, r_g) \overset{d}{=} \mathrm{ZINB}(l_n\rho_{ng}, r_g, \pi_g)$, where l_n is the library size of cell n, ρ_{ng} is the mean expression proportion of gene g in cell n generated by some injective functions $\rho_{ng} = f(\mathbf{z},\mathbf{c})$, r_g is the dispersion factor of gene g, π_{ng} is the dropout rate of gene g in cell n, and we have $\mathcal{L}_{\mathrm{recon}} = -\sum_n \log\{P_{\mathrm{ZINB}}(\mathbf{z}_n; l_n\boldsymbol{\rho}_n, \mathbf{r}, \boldsymbol{\pi})\}$.
- Inference network: We use the following inference network to approximate the generative process: with encoders $q(\mathbf{z}_d|\mathbf{x},\mathbf{a},\mathbf{w})$ and $q(\mathbf{z}_u|\mathbf{x},\mathbf{w})$, and we have

$$\mathcal{L}_{\mathrm{KL}} = \mathrm{KL}(q(\mathbf{z}_d|\mathbf{x},\mathbf{a},\mathbf{w})\|p(\mathbf{z_z}d|\mathbf{a},\mathbf{w})) + \mathrm{KL}(q(\mathbf{z}_u|\mathbf{x},\mathbf{w})\|p(\mathbf{z}_u|\mathbf{w})),$$

where KL denotes the Kullback-Leibler divergence.

To encourage disentanglement, we use β-VAE (which introduces a tunable hyperparameter β on the KL term) [6] and impose an additional independence loss [43] on VAE ELBO by Hilbert-Schmidt conditional Independence Criterion (HSCIC) [7, 8, 44] to impose $\mathbf{z}_d \perp \mathbf{z}_u|\mathbf{w}$, i.e., $\mathcal{L}_{\mathrm{ind}} = \mathrm{HSCIC}(\mathbf{z}_d, \mathbf{z}_u \mid \mathbf{w})$, where

$$\mathrm{HSCIC}(X,Y|Z) = \mathrm{Tr}(\mathbf{Q}G_Y\mathbf{Q}G_{\ddot{X}}),$$

which measures conditional dependence through the projection operator $\mathbf{Q} = \mathbf{I} - \mathbf{V}$ with $\mathbf{V} = (\mathbf{G}_Z + n\epsilon\mathbf{I})^{-1}\mathbf{G}_Z$, multiplied with the centered Gram matrix with the extended variable kernel $\mathbf{G}_{\ddot{X}} = \mathbf{G}_X \odot \mathbf{G}_Z$.

To encourage identifiability via sparsity constraints [11, 45, 46] and automatically select the optimal latent dimensions of $\mathbf{z}_d$ and $\mathbf{z}_u$, we introduce a gate variable $\mathbf{m}$ on $\mathbf{z}$ (i.e., use $\mathbf{m} \odot \mathbf{z}$ as the real latent factors for generating process) and then imposing hard concrete distribution on m and shrinkage the latent dimension by $\mathbb{E}[||\mathbf{z}||_0] = \sum_{i=1}^{d} p_{m_i}$, where p_{m_i} is the activation probability of each gate calculate via the hard concrete distribution [9].

The final loss is hence

$$\mathcal{L} = \mathcal{L}_{\text{recon}} + \beta\mathcal{L}_{\text{KL}} + \lambda_h\mathcal{L}_{\text{ind}} + \lambda_z\mathbb{E}||\mathbf{z}||_0.$$

4.2 Individual effect estimation and counterfactual prediction

In our generative model framework, the counterfactual outcomes for control group cells after treatment can be generated using the decoupled latent representations. Suppose we obtain the disentangled $\mathbf{z}_u^s \in \mathbb{R}^{n\times d_u}$, $\mathbf{z}_u^t \in \mathbb{R}^{m\times d_t}$, $\mathbf{z}_d^s \in \mathbb{R}^{n\times d_d}$, $\mathbf{z}_d^t \in \mathbb{R}^{m\times d_d}$, the perturbation-invariant latent components $\mathbf{z}_u$ will remain unchanged, while the key issue is to transfer $\mathbf{z}_d$ for samples observed in the control domain to the treatment domain. To achieve this, we transfer $\mathbf{z}_{d,i}$ from source distribution (in control group) $\mathbf{z}_d^s$ to the target distribution (in the corresponding treatment group) $\mathbf{z}_d^t$ while conditioning on similar $z_{u,i}$. We resort to conditional rank-preserving assumption, which gives a conservative estimate that minimizes the overall ITE (minimal change principle), corresponding to the Fréchet-Hoeffding upper bound on conditional coupling $(\mathbf{z}_d^s, \mathbf{z}_d^t|\mathbf{z}_u)$ [12].

Ideally, it can be achieved by defining the cost as

$$C\left(\mathbf{z}^s, \mathbf{z}^t\right) = \mathbb{I}_{D(\mathbf{z}_u^s,\mathbf{z}_u^t)=0}C(\mathbf{z}_d^s, \mathbf{z}_d^t) + \infty \cdot \mathbb{I}_{D(\mathbf{z}_u^s,\mathbf{z}_u^t)\neq 0}C(\mathbf{z}_d^s, \mathbf{z}_d^t), \tag{2}$$

to constraint the shift of $\mathbf{z}_d$ distribution under the condition of the same $\mathbf{z}_u$. However, in reality, it would be extremely hard or impossible to find samples in two domains that are exactly matched. Even if we have some class labels as surrogates (e.g. cell types), we can do conditional transfer within each class, but the heterogeneity within each class (e.g. subclass) is still neglected. And therefore, we relax to

$$C\left(\mathbf{z}^s, \mathbf{z}^t\right) = \mathbb{I}_{D(\mathbf{z}_u^s,\mathbf{z}_u^t)\leq\varepsilon}C(\mathbf{z}_d^s, \mathbf{z}_d^t) + B \cdot \mathbb{I}_{D(\mathbf{z}_u^s,\mathbf{z}_u^t)>\varepsilon}C(\mathbf{z}_d^s, \mathbf{z}_d^t), \tag{3}$$

where B is a large positive number, which is equivalent to the following cost function

$$\begin{aligned} C\left(\mathbf{z}^s, \mathbf{z}^t\right) &= \alpha D(\mathbf{z}_u^s, \mathbf{z}_u^t) + \beta C(\mathbf{z}_d^s, \mathbf{z}_d^t) \\ &= \alpha||\mathbf{z}_u^s - \mathbf{z}_u^t||_2^2 + \beta||\mathbf{z}_d^s - \mathbf{z}_d^t||_2^2, \end{aligned}$$

where α controls the strength of conditionals and β controls the weights of transport cost, and we can get the optimal coupling matrix via optimal transport (OT) solvers

$$\Gamma^* = \arg\min_{\Gamma} \int C(\mathbf{z}^s, \mathbf{z}^t) d\Gamma(\mathbf{z}^s, \mathbf{z}^t) \in \mathbb{R}^{n \times m}. \tag{4}$$

We take both distance measure $D(\mathbf{z}_u^s, \mathbf{z}_u^t)$ and transport cost $C(\mathbf{z}_d^s, \mathbf{z}_d^t)$ as squared Euclidean distances. The rationale is twofold. First, the strict convexity of this cost ensures that in the one-dimensional case, optimal transport is equivalent to quantile matching [5], while in the multi-dimensional case it can be viewed as the sum of costs across each dimension. This makes it conceptually consistent with our assumption that perturbations correspond to rank-preserving shifts along each latent dimension. Second, since we perform disentanglement using a VAE, where the KL prior encourages the latent space to approximate a low-dimensional Gaussian distribution, adopting squared Euclidean distance as both the distance metric and the transport cost in the latent space is natural and effective.

When $\alpha = 1$ and $\beta = 0$, the procedure resembles a smoothed version of kNN, where matching between pre- and post-perturbation cells relies solely on the invariant component z_u. In contrast, when $\alpha = 0$ and $\beta = 1$, the procedure amounts to applying no conditional control, directly transferring the distribution of z_d from the control group to the treatment group in a approximately quantile-preserving manner [5]. In practice, when cell types are known, we first perform domain adaptation in the latent space within each cell-type group, in which case a relatively small value of α can be used. By contrast, when cell types are unknown, domain adaptation is applied jointly across all samples, and a larger value of α is required.

Upon obtaining the optimal coupling matrix, we induce distributional shift through barycentric projection. We consider two alternative strategies: (i) performing barycentric projection over the entire latent space,

$$\tilde{\mathbf{z}}_i^{s \to t} = \frac{\sum_{j=1}^m \Gamma_{ij} \mathbf{z}_j^t}{\sum_{j=1}^m \Gamma_{ij}}, \tag{5}$$

and (ii) applying barycentric projection to $\mathbf{z}_d$ only while keeping the original $\mathbf{z}_u$ unchanged.

$$\tilde{\mathbf{z}}_i^{s \to t} = \left(\mathbf{z}_{u,i}^s, \frac{\sum_{j=1}^m \Gamma_{ij} \mathbf{z}_{d,j}^t}{\sum_{j=1}^m \Gamma_{ij}} \right). \tag{6}$$

We can then generate counterfactual samples for control group cells after treatment as $\tilde{\mathbf{x}}_i^{s \to t} \sim p(\mathbf{x} | \tilde{\mathbf{z}}_i^{s \to t}, \mathbf{c}_i)$, and the individual treatment effect (ITE) for the control cells when being treated can be estimated as

$$\hat{\boldsymbol{\tau}}_i = \mathbf{x}_i^{a=1} - \mathbf{x}_i^{a=0} = \tilde{\mathbf{x}}_i^{s \to t} - \mathbf{x}_i^s, \tag{7}$$

and the conditional average treatment effect (CATE), or more precisely conditional average treatment effect of the untreated (CATU), within a specific cell type q is $\text{CATE}_q = \frac{\sum_i \mathbb{I}(\mathbf{w}_i = q) \boldsymbol{\tau}_i}{\sum_i \mathbb{I}(\mathbf{w}_i = q)}$.

4.3 Biclustering on the ITE matrix

The ITE matrix contains information about the heterogeneity of responses to the same perturbation across different cells and genes. To uncover this heterogeneity, we employed the BCPlaid algorithm [47] (available via R package biclust, with the default parameters) to perform biclustering analysis on the estimated ITE matrix, aiming to identify subpopulations of cells and gene sets that exhibit specific response patterns.

We first perform biclustering on cell groups under different conditions (e.g., distinct cell types or perturbations), and then merge and summarize the resulting biclusters. During summarization, we first remove genes that are broadly present across all biclusters within a given condition, as these typically reflect housekeeping or background signals rather than condition-specific differences. Next, we merge biclusters within each condition, followed by merging biclusters across different conditions [21].

We merge biclusters within the same condition using the overlap score (OS) to pay more attention on exclusion, defined as

$$OS_{i,j} = \frac{e_{ij}}{\max(|g_i|, |g_j|)} \tag{8}$$

where e_{ij} denotes the number of genes only appears in biclusters i and j, and g_i represents the set of genes in bicluster i. Bicluster pairs with $OS > 0.03$ are combined. We subsequently merge them by emphasizing the degree of overlap since biclusterings across different conditions are independent, quantified using the "intersection-over-union" (IoU) criterion

$$IoU_{i,j} = \frac{|g_i \cap g_j|}{|g_i \cup g_j|} \tag{9}$$

Pairs of biclusters with $IoU > 0.3$ are combined. Through this two-step merging procedure, we ultimately obtain functional gene modules (FGMs) that capture condition-specific biological signals.

Cells are subsequently assigned to FGMs based on their bicluster membership, allowing us to characterize changes in FGM proportions across different conditions. Furthermore, we apply the R package clusterProfiler [22] to perform enrichment analysis on these differential FGMs, thereby linking the identified modules to specific functional pathways and biological processes.

4.4 Unseen dose-specific responses predicted via latent space interpolation

To quantitatively describe how cells response to a particular treatment with covariates such as chemical dose values (v, typically log-transformed), we concatenate the perturbation dose values as one component with the one-hot encoded perturbation index to form a new perturbation index (a), which is then input into the model. For the dose values of cells in the observational dataset that have been subjected to the corresponding perturbation (v_{obs}), we can directly apply the previously mentioned method to estimate the effect size of that dose treatment on the cells (The process we refer to as Estimation).

In contrast, for dose values that have never been observed in the dataset (v_{unseen}), assuming $\mathbf{z}_u$ remaining unchanged, we interpolate the unknown $\mathbf{z}_d(v_{\text{unseen}})$ by using Gaussian Process (GP) regression (conditioning on $\mathbf{w}$) to predict the dose-dependent mean $\boldsymbol{\mu}_{z_d}(v)$ and standard deviation $\boldsymbol{\sigma}_{z_d}(v)$, with hyperparameters optimized by maximizing the log marginal likelihood. Personalized counterfactuals were generated based on a residual preservation principle. For each cell in the control condition, we calculated its standardized latent residual, $\boldsymbol{\epsilon}_i = (\mathbf{z}_{d,i}^{(c)} - \boldsymbol{\mu}^{(c)}) \oslash \boldsymbol{\sigma}^{(c)}$, which captures its unique deviation from the control population. Assuming this residual is conserved, the counterfactual latent state for a new dose v_{new} was synthesized by transplanting this residual onto the predicted distribution: $\mathbf{z}_{d,i}^{(\text{new})} = \hat{\boldsymbol{\mu}}(v_{\text{new}}) + \boldsymbol{\epsilon}_i \otimes \hat{\boldsymbol{\sigma}}(v_{\text{new}})$. The final high-dimensional cell profile was then generated by passing the complete counterfactual vector, $[\mathbf{z}_{d,i}^{(\text{new})}, \mathbf{z}_{u,i}^{(c)}]$, through the pre-trained VAE decoder (The process we refer to as Extrapolation).

4.5 Synthetic effect estimation of multiple treatments through ITE vector composition

To infer the combined effects of multiple perturbations from a single-cell dataset that includes responses to different individual perturbations, we analogize the superposition of different perturbations to the addition of vectors (Fig.1d). Through decoupled learning, we can estimate the ITE for each cell i in the control group after receiving perturbation A (τ_i^A) and perturbation B (τ_i^B) separately. We can then calculate the correlation coefficient r_{AB} between τ^A and τ^B among all cells, which can be interpreted as the cosine of the angle between the vectors representing the effects of these two treatments. Subsequently, using the parallelogram law of vector addition and the cosine theorem, we can infer the absolute synthetic effect τ_i^{AB} when the cells simultaneously receive treatments A and B as

$$|\tau_i^{AB}| = \sqrt{(\tau_i^A)^2 + (\tau_i^B)^2 - 2\tau_i^A \tau_i^B r_{AB}} \tag{10}$$

Given that angles summing to 360 degrees share an identical cosine, we can only estimate the magnitude of the combined effect, while its direction (up- or down-regulation) cannot be discerned (Fig.1d). This is evident from the distinct, absolute value-like shape of the plot (Fig.6c).

4.6 Simulations

We simulate single-cell perturbation dataset with the following procedures. We first simulate N cells in the control group, where the latent components are generated from Gaussian mixture model $z_d \sim \mathcal{N}(\mu_{at}^D, \sigma_{at}^{D^2})$ and $z_u \sim \mathcal{N}(\mu_t^I, \sigma_w^{I^2})$ with cell type labels t sampled from a Dirichlet distribution $Dir(K)$. Then, for treatment group, we use the same z_u and transform z_d to $\mathcal{N}(\mu_{at}^D + \delta_{at}, \sigma_{at}^{D^2})$, where $\delta_{at} \sim \mathcal{N}(1, 0.5^2)$, and ensure that the ranks of the corresponding samples remain consistent before and after the shift. The mean expression of gene g for each sample μ_{ng} is generated as $\mu_{ng} = \log\left\{\exp\left(f_g(z_d^{(n)}, z_u^{(n)})\right) + 1\right\}$, where $f_g(\cdot)$ is a non-linear neural network,

where we impose selective masks on latent components when generating different genes to increase the sparsity. The dispersion factors r_g and dropout rates π_g are sampled to make highly expressed genes have lower Biological Coefficient of Variation (BCV, here we specify dispersion factor $r = 1/\alpha = 1/\text{BCV}^2$) and lower dropout rate, i.e., $r_g = \sqrt{\frac{5}{\mu_g/\sum_g \mu_g + 10} + 0.1 + \mathcal{N}(0, 0.01^2)}$, $\pi_g = \frac{0.4}{1+\exp\left\{0.1(\mu_g - \text{median}[\mu_g])/\sum_g \mu_g\right\}} + \mathcal{N}(0, 0.01^2)$. The gene expressions are then sampled from zero-inflated negative binomial distributions $x_{ng} \sim \pi_g \delta_0 + (1 - \pi_g)\text{NegBin}(\mu_{ng}, r_g)$. We randomly sample $N/2$ cells in control and $N/2$ cells in treatment groups to get the final observed single-cell dataset.

4.7 Evaluations

For the simulated data, we computed the RMSE of the estimated ITE and CATE on all genes. For the real datasets, we compared the observed treatment cells with the counterfactually generated "pseudo-treatment" samples from the control cells. We evaluated the RMSE and PCC of their mean gene expression levels within top 3000 HVGs as we cannot access the individual-level counterfactual truth. We also evaluated the MMD between the observed treatment and counterfactually predicted distributions. The MMD is calculated using top 200 HVGs, and estimated based on the empirical distributions $\hat{P}_X$ and $\hat{P}_Y$ is estimated by the V-statistic (for convenience in vectorized implementation) [48],

$$\text{MMD}^2[\hat{P}_X, \hat{P}_Y] = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{i'=1}^{n} k(x_i, x_{i'}) + \frac{1}{m^2}\sum_{j=1}^{m}\sum_{j'=1}^{m} k(y_j, y_{j'}) - \frac{2}{nm}\sum_{i=1}^{n}\sum_{j=1}^{m} k(x_i, y_j), \tag{11}$$

where $k(\cdot,\cdot)$ is the Gaussian (Radial Basis Function, RBF) kernel $k(x, y) = \exp\left(-\frac{\|x-y\|^2}{2\sigma^2}\right)$. The bandwidth σ is chosen using the median heuristic, where it is set equal to the median of the Euclidean distances between all pairs of samples in the dataset. We also visualized the results using UMAP.

Supplementary information. Supplementary materials are available online.

Acknowledgements. We sincerely appreciate Dr Richard Border and Dr. Martin Jinye Zhang for their valuable comments on this work. Fig.1 is created with BioRender (https://BioRender.com).

Declarations

- Funding: We would like to acknowledge the support from NSF Award No. 2229881, AI Institute for Societal Decision Making (AI-SDM), the National Institutes of Health (NIH) under Contract R01HL159805, and grants from Quris AI, Florin Court Capital, and MBZUAI-WIS Joint Program, and the Al Deira Causal Education project.
- Conflict of interest: No conflict of interest.
- Ethics approval and consent to participate: Not applicable.

- Consent for publication: Not applicable.
- Data availability: The Kang and Hagai datasets are available via pertpy package (http://pertpy.readthedocs.io/en/latest/index.html). The Adamson and sciplex4 datasets are available via scPerturb (https://zenodo.org/records/13350497). The Dong dataste are available from DRYAD (https://datadryad.org/dataset/doi:10.5061/dryad.4xgxd25g1).
- Code availability: Codes of scDRP are available at https://github.com/sjl-sjtu/scDRP. CINEMA-OT and scGen are implemented with Python library pertpy (http://pertpy.readthedocs.io/en/latest/index.html).

References

[1] Peidli, S., Green, T.D., Shen, C., Gross, T., Min, J., Garda, S., Yuan, B., Schumacher, L.J., Taylor-King, J.P., Marks, D.S., *et al.*: scPerturb: harmonized single-cell perturbation data. Nature Methods **21**(3), 531–540 (2024)

[2] Lotfollahi, M., Wolf, F.A., Theis, F.J.: scGen predicts single-cell perturbation responses. Nature Methods **16**(8), 715–721 (2019)

[3] Adduri, A.K., Gautam, D., Bevilacqua, B., Imran, A., Shah, R., Naghipourfar, M., Teyssier, N., Ilango, R., Nagaraj, S., Dong, M., et al.: Predicting cellular responses to perturbation across diverse contexts with state. bioRxiv, 2025–06 (2025)

[4] Dong, M., Wang, B., Wei, J., O. Fonseca, A.H., Perry, C.J., Frey, A., Ouerghi, F., Foxman, E.F., Ishizuka, J.J., Dhodapkar, R.M., *et al.*: Causal identification of single-cell experimental perturbation effects with cinema-ot. Nature Methods **20**(11), 1769–1779 (2023)

[5] Chewi, S., Niles-Weed, J., Rigollet, P.: Statistical optimal transport (2024)

[6] Higgins, I., Matthey, L., Pal, A., Burgess, C., Glorot, X., Botvinick, M., Mohamed, S., Lerchner, A.: beta-VAE: Learning basic visual concepts with a constrained variational framework. In: International Conference on Learning Representations (2017)

[7] Fukumizu, K., Bach, F.R., Jordan, M.I.: Dimensionality reduction for supervised learning with reproducing kernel hilbert spaces. Journal of Machine Learning Research **5**(Jan), 73–99 (2004)

[8] Sheng, T., Sriperumbudur, B.K.: On distance and kernel measures of conditional dependence. Journal of Machine Learning Research **24**(7), 1–16 (2023)

[9] Louizos, C., Welling, M., Kingma, D.P.: Learning sparse neural networks through l_0 regularization. In: International Conference on Learning Representations (2018)

[10] Khemakhem, I., Kingma, D., Monti, R., Hyvarinen, A.: Variational autoencoders and nonlinear ica: A unifying framework. In: International Conference on Artificial Intelligence and Statistics, pp. 2207–2217 (2020). PMLR

[11] Zheng, Y., Zhang, K.: Generalizing nonlinear ica beyond structural sparsity. Advances in Neural Information Processing Systems **36**, 13326–13355 (2023)

[12] Balakrishnan, S., Kennedy, E., Wasserman, L.: Conservative inference for counterfactuals. Journal of Causal Inference **13**(1), 20230071 (2025)

[13] Kang, H.M., Subramaniam, M., Targ, S., Nguyen, M., Maliskova, L., McCarthy, E., Wan, E., Wong, S., Byrnes, L., Lanata, C.M., *et al.*: Multiplexed droplet single-cell rna-sequencing using natural genetic variation. Nature Biotechnology **36**(1), 89–94 (2018)

[14] Hagai, T., Chen, X., Miragaia, R.J., Rostom, R., Gomes, T., Kunowska, N., Henriksson, J., Park, J.-E., Proserpio, V., Donati, G., *et al.*: Gene expression variability across cells and species shapes innate immunity. Nature **563**(7730), 197–202 (2018)

[15] Adamson, B., Norman, T.M., Jost, M., Cho, M.Y., Nuñez, J.K., Chen, Y., Villalta, J.E., Gilbert, L.A., Horlbeck, M.A., Hein, M.Y., *et al.*: A multiplexed single-cell crispr screening platform enables systematic dissection of the unfolded protein response. Cell **167**(7), 1867–1882 (2016)

[16] Faris, A.N., Ganesan, S., Chattoraj, A., Chattoraj, S.S., Comstock, A.T., Unger, B.L., Hershenson, M.B., Sajjan, U.S.: Rhinovirus delays cell repolarization in a model of injured/regenerating human airway epithelium. American Journal of Respiratory Cell and Molecular Biology **55**(4), 487–499 (2016)

[17] Gagliardi, T.B., Goldstein, M.E., Song, D., Gray, K.M., Jung, J.W., Ignacio, M.A., Stroka, K.M., Duncan, G.A., Scull, M.A.: Rhinovirus c replication is associated with the endoplasmic reticulum and triggers cytopathic effects in an in vitro model of human airway epithelium. PLoS pathogens **18**(1), 1010159 (2022)

[18] Brekman, A., Walters, M.S., Tilley, A.E., Crystal, R.G.: Foxj1 prevents cilia growth inhibition by cigarette smoke in human airway epithelium in vitro. American journal of respiratory cell and molecular biology **51**(5), 688–700 (2014)

[19] Crystal, R.G.: Airway basal cells. the "smoking gun" of chronic obstructive pulmonary disease. American journal of respiratory and critical care medicine **190**(12), 1355–1362 (2014)

[20] Lan, M.-Y., Ho, C.-Y., Lee, T.-C., Yang, A.-H.: Cigarette smoke extract induces cytotoxicity on human nasal epithelial cells. American journal of rhinology **21**(2), 218–223 (2007)

[21] Gong, Y., Xu, J., Wu, M., Gao, R., Sun, J., Yu, Z., Zhang, Y.: Single-cell biclustering for cell-specific transcriptomic perturbation detection in ad progression. Cell Reports Methods **4**(4) (2024)

[22] Xu, S., Hu, E., Cai, Y., Xie, Z., Luo, X., Zhan, L., Tang, W., Wang, Q., Liu, B., Wang, R., *et al.*: Using clusterprofiler to characterize multiomics data. Nature Protocols **19**(11), 3292–3320 (2024)

[23] Crouse, J., Bedenikovic, G., Wiesel, M., Ibberson, M., Xenarios, I., Von Laer, D., Kalinke, U., Vivier, E., Jonjic, S., Oxenius, A.: Type i interferons protect t cells against nk cell attack mediated by the activating receptor ncr1. Immunity **40**(6), 961–973 (2014)

[24] Pramanik, J., Chen, X., Kar, G., Henriksson, J., Gomes, T., Park, J.-E., Natarajan, K., Meyer, K.B., Miao, Z., McKenzie, A.N., *et al.*: Genome-wide analyses reveal the ire1a-xbp1 pathway promotes t helper cell differentiation by resolving secretory stress and accelerating proliferation. Genome Medicine **10**(1), 76 (2018)

[25] So, J.-S.: Roles of endoplasmic reticulum stress in immune responses. Molecules and Cells **41**(8), 705–716 (2018)

[26] Dahlgren, M.W., Plumb, A.W., Niss, K., Lahl, K., Brunak, S., Johansson-Lindbom, B.: Type i interferons promote germinal centers through b cell intrinsic signaling and dendritic cell dependent th1 and tfh cell lineages. Frontiers in Immunology **13**, 932388 (2022)

[27] Swanson, C.L., Wilson, T.J., Strauch, P., Colonna, M., Pelanda, R., Torres, R.M.: Type i ifn enhances follicular b cell contribution to the t cell–independent antibody response. Journal of Experimental Medicine **207**(7), 1485–1500 (2010)

[28] Li, Q., Wu, Y., Zhang, Y., Sun, H., Lu, Z., Du, K., Fang, S., Li, W.: mir-125b regulates cell progression in chronic myeloid leukemia via targeting bak1. American Journal of Translational Research **8**(2), 447 (2016)

[29] Tallack, M.R., Perkins, A.C.: Klf1 directly coordinates almost all aspects of terminal erythroid differentiation. IUBMB Life **62**(12), 886–890 (2010)

[30] Ianniciello, A., Zarou, M.M., Rattigan, K.M., Scott, M., Dawson, A., Dunn, K., Brabcova, Z., Kalkman, E.R., Nixon, C., Michie, A.M., *et al.*: Ulk1 inhibition promotes oxidative stress–induced differentiation and sensitizes leukemic stem cells to targeted therapy. Science Translational Medicine **13**(613), 5016 (2021)

[31] Shyamsunder, P., Shanmugasundaram, M., Mayakonda, A., Dakle, P., Teoh, W.W., Han, L., Kanojia, D., Lim, M.C., Fullwood, M., An, O., *et al.*: Identification of a novel enhancer of cebpe essential for granulocytic differentiation. Blood, The Journal of the American Society of Hematology **133**(23), 2507–2517 (2019)

[32] Romo-González, M., Ijurko, C., Hernández-Hernández, Á.: Reactive oxygen species and metabolism in leukemia: a dangerous liaison. Frontiers in Immunology **13**, 889875 (2022)

[33] Kaweme, N.M., Zhou, S., Changwe, G.J., Zhou, F.: The significant role of redox system in myeloid leukemia: from pathogenesis to therapeutic applications. Biomarker Research **8**(1), 63 (2020)

[34] Srivatsan, S.R., McFaline-Figueroa, J.L., Ramani, V., Saunders, L., Cao, J., Packer, J., Pliner, H.A., Jackson, D.L., Daza, R.M., Christiansen, L., *et al.*: Massively multiplex chemical transcriptomics at single-cell resolution. Science **367**(6473), 45–51 (2020)

[35] Lee, J.V., Berry, C.T., Kim, K., Sen, P., Kim, T., Carrer, A., Trefely, S., Zhao, S., Fernandez, S., Barney, L.E., *et al.*: Acetyl-coa promotes glioblastoma cell adhesion and migration through ca2+–nfat signaling. Genes & Development **32**(7-8), 497–511 (2018)

[36] Schwab, A., Siddiqui, A., Vazakidou, M.E., Napoli, F., Böttcher, M., Menchicchi, B., Raza, U., Saatci, Ö., Krebs, A.M., Ferrazzi, F., *et al.*: Polyol pathway links glucose metabolism to the aggressiveness of cancer cells. Cancer Research **78**(7), 1604–1618 (2018)

[37] Chen, X., Chen, C., Hao, J., Qin, R., Qian, B., Yang, K., Zhang, J., Zhang, F.: Akr1b1 upregulation contributes to neuroinflammation and astrocytes proliferation by regulating the energy metabolism in rat spinal cord injury. Neurochemical Research **43**(8), 1491–1499 (2018)

[38] Batchuluun, B., Pinkosky, S.L., Steinberg, G.R.: Lipogenesis inhibitors: therapeutic opportunities and challenges. Nature Reviews Drug Discovery **21**(4), 283–305 (2022)

[39] Sun, J., Liang, C., Wei, R., Zheng, P., Bai, L., Ouyang, W., Yan, H., Ye, P.: scmrdr: A scalable and flexible framework for unpaired single-cell multi-omics data integration. arXiv preprint arXiv:2510.24987 (2025)

[40] Tejada-Lapuerta, A., Bertin, P., Bauer, S., Aliee, H., Bengio, Y., Theis, F.J.: Causal machine learning for single-cell genomics. Nature Genetics, 1–12 (2025)

[41] Rubin, A.J., Parker, K.R., Satpathy, A.T., Qi, Y., Wu, B., Ong, A.J., Mumbach, M.R., Ji, A.L., Kim, D.S., Cho, S.W., *et al.*: Coupled single-cell crispr screening and epigenomic profiling reveals causal gene regulatory networks. Cell **176**(1), 361–376 (2019)

[42] Lopez, R., Regier, J., Cole, M.B., Jordan, M.I., Yosef, N.: Deep generative modeling for single-cell transcriptomics. Nature Methods **15**(12), 1053–1058 (2018)

[43] Lopez, R., Regier, J., Jordan, M.I., Yosef, N.: Information constraints on auto-encoding variational bayes. Advances in Neural Information Processing Systems **31** (2018)

[44] Fukumizu, K., Gretton, A., Sun, X., Schölkopf, B.: Kernel measures of conditional dependence. Advances in neural information processing systems **20** (2007)

[45] Zheng, Y., Ng, I., Zhang, K.: On the identifiability of nonlinear ica: Sparsity and beyond. Advances in Neural Information Processing Systems **35**, 16411–16422 (2022)

[46] Li, Z., Fan, S., Zheng, Y., Ng, I., Xie, S., Chen, G., Dong, X., Cai, R., Zhang, K.: Synergy between sufficient changes and sparse mixing procedure for disentangled representation learning. The Thirteenth International Conference on Learning Representations (2025)

[47] Turner, H., Bailey, T., Krzanowski, W.: Improved biclustering of microarray data demonstrated through systematic performance tests. Computational Statistics & Data Analysis **48**(2), 235–254 (2005)

[48] Gretton, A., Borgwardt, K.M., Rasch, M.J., Schölkopf, B., Smola, A.: A kernel two-sample test. The Journal of Machine Learning Research **13**(1), 723–773 (2012)

Supplementary Materials

Supplementary Notes

SN1 Identifiability theory of non-linear disentanglement

1.1 General Theory

Recent advances in the theory of disentangled and nonlinear ICA models have progressively relaxed the conditions required for identifiability, i.e., the ability to recover unique latent variables up to subgroup or component-wise invertible transformations and permutations.

Zheng et al. [1] first established Structural Sparsity as a sufficient condition for identifiability in fully unsupervised nonlinear ICA, showing that when each latent factor uniquely determines at least one subset of observed variables (i.e., the intersection of their parents identifies that source), the true latent components can be recovered without any auxiliary variables.

Building upon this, Zheng and Zhang [2] generalized the theory to more realistic settings by introducing partial sparsity, undercompleteness (more observed than latent variables), and partial source dependence. They proved that if structural sparsity and independence hold for only a subset of sources, those sources remain component-wise identifiable, while the remaining (possibly dependent) ones are identifiable up to a subspace. The framework also allows auxiliary variables with only a small number of distinct values to assist identifiability, extending nonlinear ICA to mixed or grouped latent structures.

Most recently, Li et al. [3] unified two previously separate lines of identifiability research, i.e., sufficient distributional changes across auxiliary variables and sparse mixing structures, into a single complementary framework. They demonstrated that sparsity in the mixing process can compensate for insufficient domain variation, while diverse conditional distributions can mitigate violations of sparsity. Under smoothness, conditional independence, and either moderate distributional changes or partial sparsity, both subspace- and component-wise identifiability can be achieved with milder assumptions than required by classical nonlinear ICA results.

These works broaden the identifiable regimes for disentangled representation learning, where the theoretical insights are particularly relevant to single-cell generative modeling, where identifying biologically meaningful latent factors, such as cell identity or state, is paramount. In practice, the structural sparsity corresponds to the modular nature of gene regulatory networks, where each latent factor (e.g., a specific transcription factor) regulates only a sparse

subset of marker genes. Simultaneously, the auxiliary variables manifest as experimental metadata, such as perturbation conditions (e.g., drug treatment or gene knockdown). By leveraging the sparse connectivity of gene modules alongside the distributional shifts induced by these perturbations, generative models can achieve robust identifiability, effectively disentangling intrinsic cellular programs from extrinsic experimental effects.

1.2 Identifiability under noisy observations

While the aforementioned theories primarily focus on deterministic or noiseless mixing processes, real-world generative mechanisms are inevitably corrupted by observation noise, which fundamentally alters the identifiability landscape. Khemakhem et al. [4] provided the rigorous identifiability results for deep latent-variable models under additive noise by embedding nonlinear ICA theory into the variational autoencoder framework. Specifically, for a noisy generative process of the form $x = f(z) + \varepsilon$, where f is an injective nonlinear function and ε denotes independent noise, we introduced a conditionally factorized latent prior $p(z \mid u)$ whose parameters vary with an auxiliary observed variable u. By modeling each latent component with an exponential family whose sufficient statistics are modulated by u, the joint distribution $p(x, z \mid u)$ becomes identifiable up to component-wise invertible transformations and permutations, even in the presence of noise and under undercomplete settings. This result demonstrates that distributional variation across auxiliary variables can overcome the fundamental non-identifiability induced by nonlinear mixing and observation noise, extending classical nonlinear ICA identifiability to realistic stochastic generative models. Importantly, their theory establishes consistency of VAE-based maximum likelihood estimation, showing that both latent variables and generative mechanisms can be recovered (up to trivial indeterminacies) rather than merely matching the marginal data distribution.

Classical structural sparsity-based identifiability results enforce sparsity in the Jacobian of the mixing function to restrict the dependency graph between latent and observed variables. However, explicitly computing and regularizing the Jacobian becomes prohibitively expensive in deep generative models. Instead, we adopt a complementary surrogate by constraining the effective latent dimensionality through learnable gating variables with ℓ_0 regularization. This mechanism encourages only a small subset of latent components to actively participate in the generative process, thereby implicitly inducing sparse dependency structures between latent and observed variables. From a functional perspective, restricting latent participation reduces the expressiveness of admissible mixing functions and limits the class of nonlinear reparameterizations responsible for non-identifiability. While not equivalent to explicit Jacobian sparsity, this latent selection strategy serves as a practical approximation that promotes structural sparsity in deep models with significantly lower computational overhead.

1.3 Regularization on sparsity

Classical structural sparsity-based identifiability results enforce sparsity in the Jacobian of the mixing function to restrict the dependency graph between latent and observed variables. However, explicitly computing and regularizing the Jacobian becomes prohibitively expensive in deep generative models. Instead, we adopt a complementary surrogate by constraining the effective latent dimensionality through learnable gating mechanism $\tilde{z} = g \odot z,\ g \in \{0,1\}^n$ with ℓ_0 regularization [5]. This mechanism encourages only a small subset of latent components to actively participate in the generative process, thereby implicitly inducing sparse dependency structures between latent and observed variables. Moreover, by encouraging only a small number of active gates, the model automatically infers the effective latent dimensionality, mitigating the overparameterization that otherwise exacerbates non-identifiability in nonlinear generative models. Though slightly stronger than classical entry-wise Jacobian sparsity, this latent selection strategy serves as a commonly used and practical approximation that promotes structural sparsity in deep models with significantly lower computational overhead [6].

SN2 Conditional optimal transport on disentangled latent space.

Recall that we assume that the observed variables $\boldsymbol{x}$ are generated by both domain (perturbation)-dependent latent factors $\boldsymbol{z}_d \sim p_d(\boldsymbol{z}_d|\boldsymbol{a},\boldsymbol{w})$ and domain-invariant latent factors $\boldsymbol{z}_u \sim p_u(\boldsymbol{z}_u|\boldsymbol{w})$, where $\boldsymbol{a}$ stands for the domain index, $\boldsymbol{w}$ stands for the confounding, and we assume that $\boldsymbol{z}_d \perp \boldsymbol{z}_u|\boldsymbol{w}$.

The idea of counterfactual matching in the disentangled latent space is that, we assume that the effect of changing domains (perturbations) $\boldsymbol{a}$ on $\boldsymbol{z}_d$ is rank-preserving, conditional on confoundings. That's to say, perturbations might have heterogeneous effects on the latent factors of cells belonging to different cell types or cell states. However, after controlling for these confounding factors (cell type or cell state), the effect of the perturbation on the cells within that subgroup should be similar, meaning that the relative positions (distribution rank or quantile) of these cells before and after the perturbation should be preserved.

We resort to optimal transport (OT) for the implementation of quantile matching, i.e.,

$$\Gamma^* = \mathrm{argmin}_\Gamma \int \left|\left|\boldsymbol{z}_d^s - \boldsymbol{z}_d^t\right|\right|^2 d\Gamma(\boldsymbol{z}^s, \boldsymbol{z}^t),$$

since there is an equivalence between OT with a strictly convex cost (like square Euclidean distance) and quantile matching in one-dimension [7, 8]. In multi-dimensional latent space, although inexact, we can view this as a trade-off in quantile matching across different dimensions, because the squared Euclidean distance can be seen as the summation of the squared Euclidean distances in each dimension.

When confounding $\boldsymbol{w}$ is not observed or not fully observed (e.g. only known a surrogate like

cell type labels), $\boldsymbol{z}_u$ can serve as a representation of confounding. It is hard to impose hard conditionals on $\boldsymbol{z}_u$, i.e.,

$$\Gamma^* = \text{argmin}_\Gamma \int \left|\left|\boldsymbol{z}_d^s - \boldsymbol{z}_d^t\right|\right|^2 d\Gamma(\boldsymbol{z}^s, \boldsymbol{z}^t),$$

$$\text{s.t.}\, \boldsymbol{z}_u^s = \boldsymbol{z}_u^t,$$

and we relax to a soft version,

$$\Gamma^* = \text{argmin}_\Gamma \int \left|\left|\boldsymbol{z}_d^s - \boldsymbol{z}_d^t\right|\right|^2 d\Gamma(\boldsymbol{z}^s, \boldsymbol{z}^t),$$

$$\text{s.t.}\, \left|\left|\boldsymbol{z}_u^s - \boldsymbol{z}_u^t\right|\right|^2 \leq \varepsilon.$$

and this is equivalent to solving

$$\Gamma^* = \text{argmin}_\Gamma \int \left(\alpha\left|\left|\boldsymbol{z}_u^s - \boldsymbol{z}_u^t\right|\right|^2 + \beta\left|\left|\boldsymbol{z}_d^s - \boldsymbol{z}_d^t\right|\right|^2\right) d\Gamma(\boldsymbol{z}^s, \boldsymbol{z}^t).$$

We can then generate counterfactual latent factors by barycentric projections on the full latent space

$$\tilde{\boldsymbol{z}}_i^{s\to t} = \frac{\sum_{j=1}^m \Gamma_{ij}\boldsymbol{z}_j^t}{\sum_{j=1}^m \Gamma_{ij}},$$

or on the $\boldsymbol{z}_d$ space only

$$\tilde{\boldsymbol{z}}_i^{s\to t} = \left(\boldsymbol{z}_{u,i}^s, \frac{\sum_{j=1}^m \Gamma_{ij}\boldsymbol{z}_{d,j}^t}{\sum_{j=1}^m \Gamma_{ij}}\right).$$

SN3 Toy examples to illustrate the idea of conditional optimal transport on disentangled latent space.

We constructed synthetic datasets, $\boldsymbol{X}_s$ (Source) and $\boldsymbol{X}_t$ (Target), in a latent space. Each data point $\boldsymbol{x} = [\boldsymbol{z}_u, \boldsymbol{z}_d]^T$ is composed of an **invariant component** ($\boldsymbol{z}_u$) and a **domain-specific component** ($\boldsymbol{z}_d$).

The $\boldsymbol{z}_u$ component is designed to represent conserved cell identity (e.g., cell types), featuring two distinct, well-separated clusters, symmetrically centered around ± 2, ensuring that this fundamental structure is shared between $\boldsymbol{X}_s$ and $\boldsymbol{X}_t$. In contrast, the $\boldsymbol{z}_d$ component models the domain shift or perturbation effect. While the source $\boldsymbol{z}_d$ is centered at 0, the target $\boldsymbol{z}_d$ is generated with cluster-specific shifts ($\pm$ 2), resulting in a clear domain separation that is dependent on the underlying $\boldsymbol{z}_u$ identity.

Specifically, the distributions are formulated as:

1. Invariant Factors ($\boldsymbol{z}_u$):

$$z_{u,s}, z_{u,t} \sim Mixture\big(Dist_u(-2,0.25), Dist_u(2,0.5)\big)$$

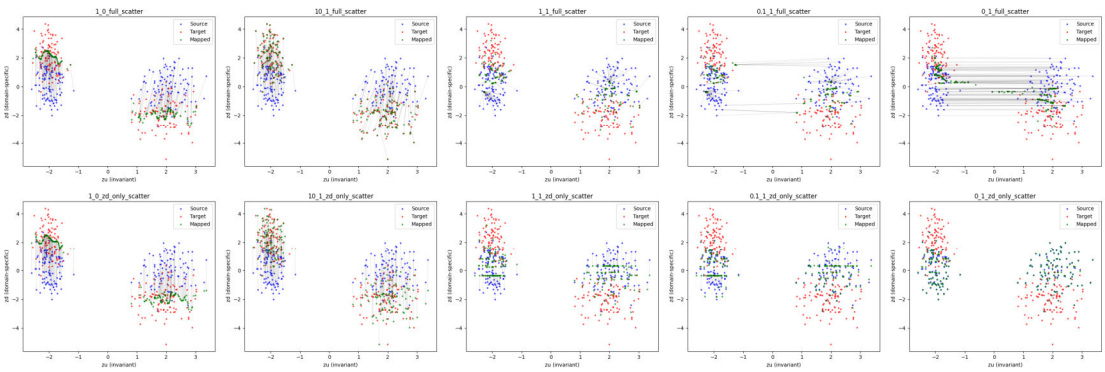
2. Domain-Specific Factors ($\boldsymbol{z}_d$):

$$z_{d,s} \sim Dist_d(0,1)$$

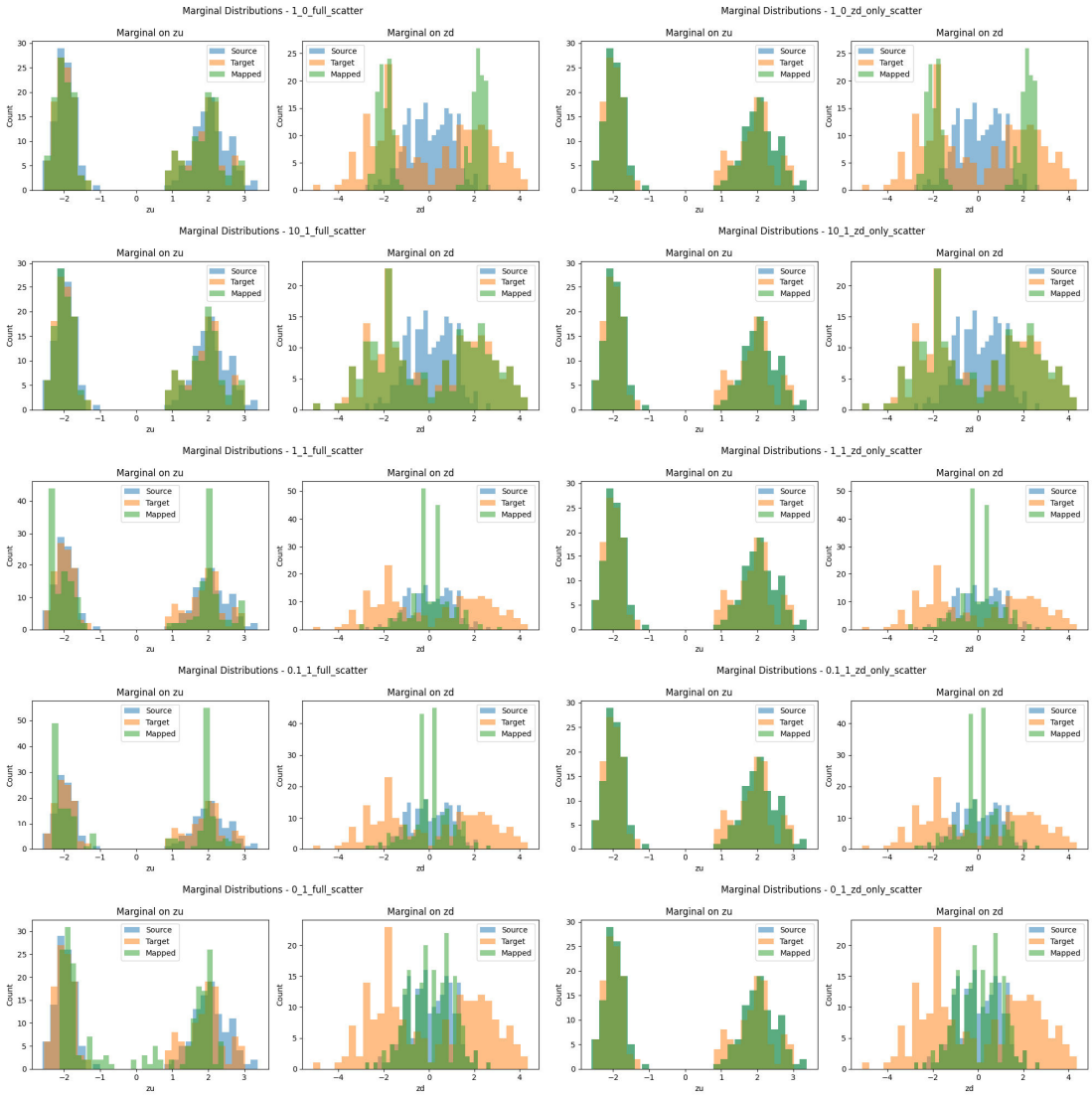
$$z_{d,t} \sim Mixture\big(Dist_d(2,1), Dist_d(-2,1)\big)$$

By allowing $Dist_u$ and $Dist_d$ to be selected from distributions such as Normal, Laplace, or Uniform, we can explore the robustness of the conditional optimal transport under various distributional differences, ensuring the generated data faithfully captures the need to align $\boldsymbol{z}_d$ while strictly preserving the $\boldsymbol{z}_u$ clustering structure.

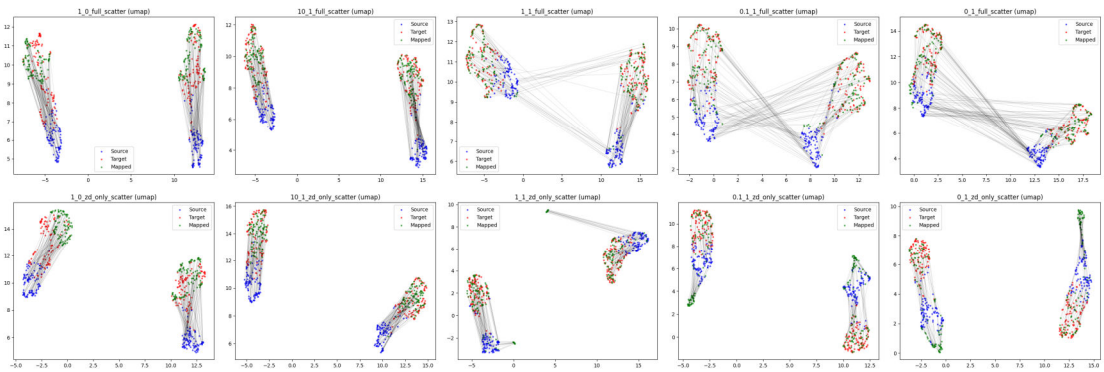
We first conducted simulations in a setting where both $\boldsymbol{z}_d$ and $\boldsymbol{z}_u$ are 1-dimensional, and presented the scatter plots of the counterfactual generations $\boldsymbol{x}$ (Supplementary Fig.S1) and the marginal distributions of $\boldsymbol{z}_d$ and $\boldsymbol{z}_u$ (Supplementary Fig.S2). It can be seen that when no conditioning is performed ($\alpha = 0$) or a very small α is set, significant cross-matching occurs between different groups. Although the marginals can be aligned, there is a clear issue of confounding interference. The issue becomes even more severe when performing barycentric mapping on $\boldsymbol{z}_d$ only, as the lack of conditioning leads to a more severe distribution shift. Conversely, when $\boldsymbol{z}_u$ is fully utilized for confounding matching without $\boldsymbol{z}_d$ transfer ($\beta = 0$), we observe that while the marginals of $\boldsymbol{z}_u$ can be aligned, the marginals of $\boldsymbol{z}_d$ exhibit significant differences. The simulations conducted on a multi-dimensional latent space further corroborate this point (Supplementary Fig.S3). Therefore, using an appropriate level of conditioning is crucial.



Supplementary Fig.S1 The scatter plots of the counterfactual generations in 1-dimensional latent-space toy example. The title of each subplot shows the configuration (alpha_beta_projection method) used.



Supplementary Fig.S2 The marginal distributions of $\boldsymbol{z}_d$ and $\boldsymbol{z}_u$ in 1-dimensional latent-space toy example. The title of each subplot shows the configuration (alpha_beta_projection method) used.



Supplementary Fig.S3 The scatter plots of the counterfactual generations in multi-dimensional latent-space toy example ($dim(\boldsymbol{z}_d) = dim(\boldsymbol{z}_u) = 10$ here). The title of each subplot shows the configuration (alpha_beta_projection method) used.

Reference:

[1] Zheng, Y., Ng, I., & Zhang, K. (2022). On the identifiability of nonlinear ICA: Sparsity and beyond. *Advances in neural information processing systems*, *35*, 16411-16422.

[2] Zheng, Y., & Zhang, K. (2023). Generalizing nonlinear ICA beyond structural sparsity. *Advances in Neural Information Processing Systems*, *36*, 13326-13355.

[3] Li, Z., Fan, S., Zheng, Y., Ng, I., Xie, S., Chen, G., ... & Zhang, K. (2025). Synergy Between Sufficient Changes and Sparse Mixing Procedure for Disentangled Representation Learning. In *The Thirteenth International Conference on Learning Representations*.

[4] Khemakhem, I., Kingma, D., Monti, R., & Hyvarinen, A. (2020). Variational autoencoders and nonlinear ica: A unifying framework. In *International conference on artificial intelligence and statistics* (pp. 2207-2217). PMLR.

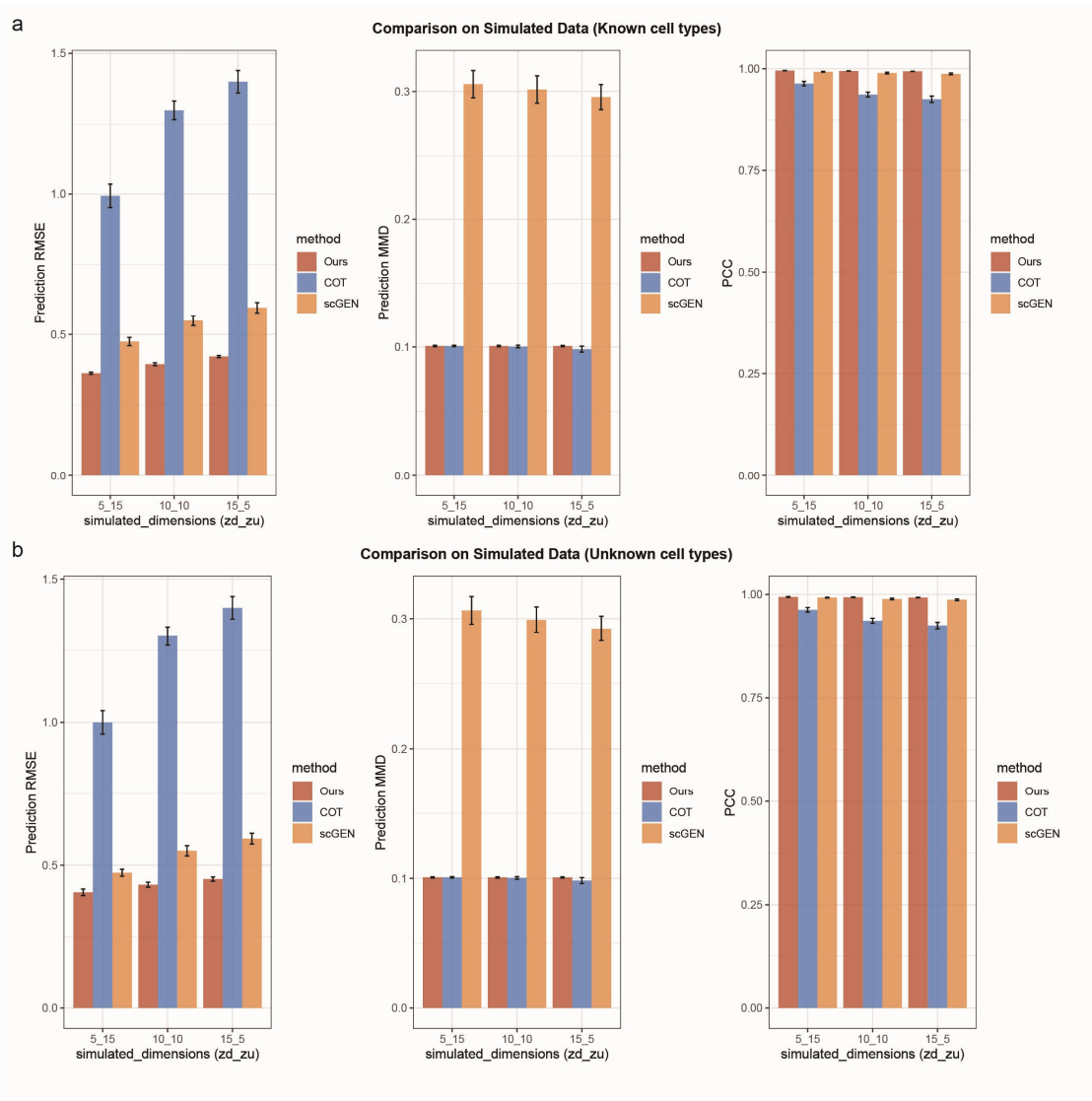
[5] Louizos, C., Welling, M., & Kingma, D. P. (2018). Learning Sparse Neural Networks through L0 Regularization. In *International Conference on Learning Representations*.

[6] Xie, S., Kong, L., Gong, M., & Zhang, K. (2023). Multi-domain image generation and translation with identifiability guarantees. In *The Eleventh international conference on learning representations*.

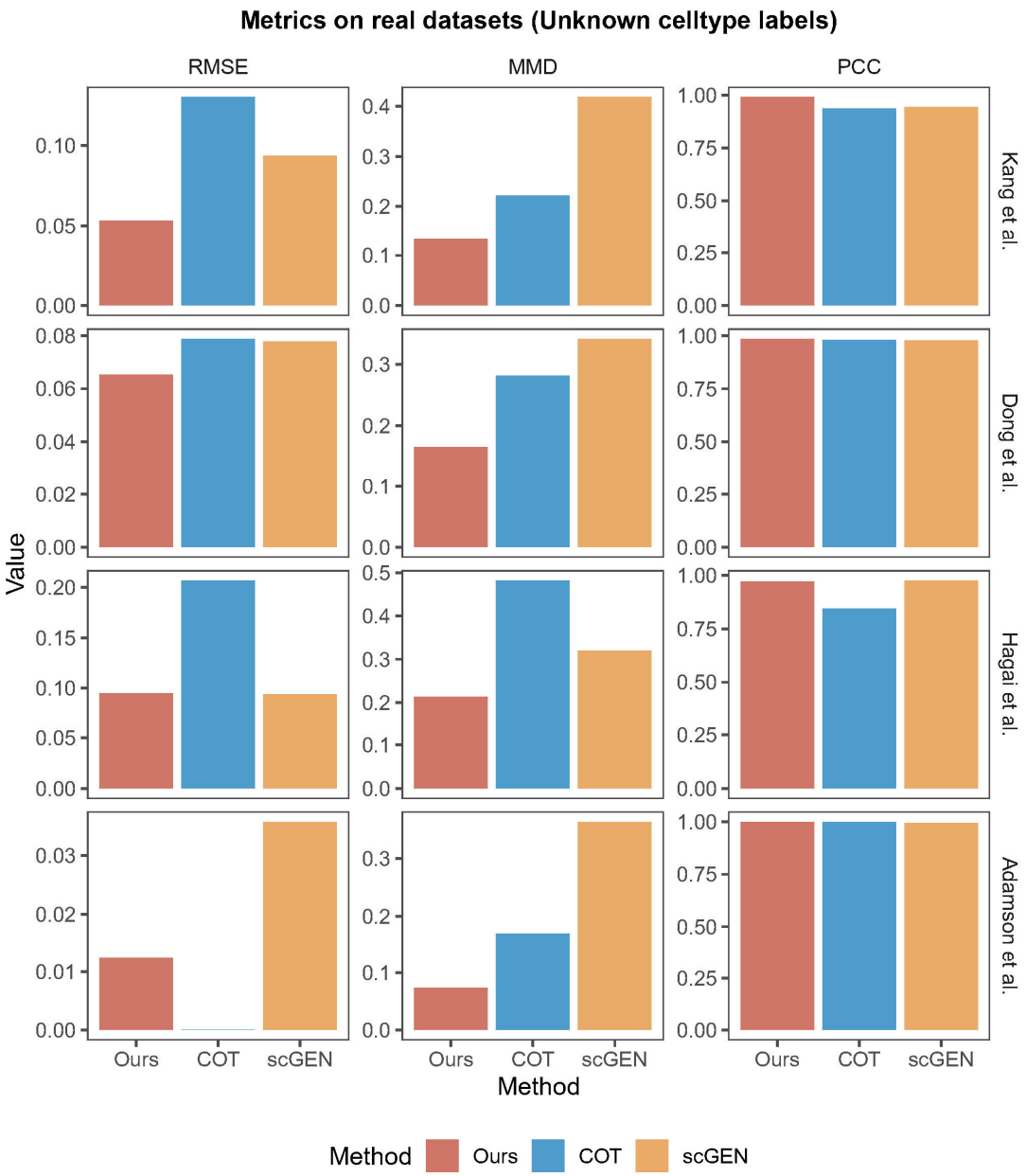
[7] Chewi, S., Niles-Weed, J., & Rigollet, P. (2024). Statistical optimal transport. *Springer*.

[8] Balakrishnan, S., Kennedy, E., & Wasserman, L. (2025). Conservative inference for counterfactuals. *Journal of Causal Inference*, *13*(1), 20230071.

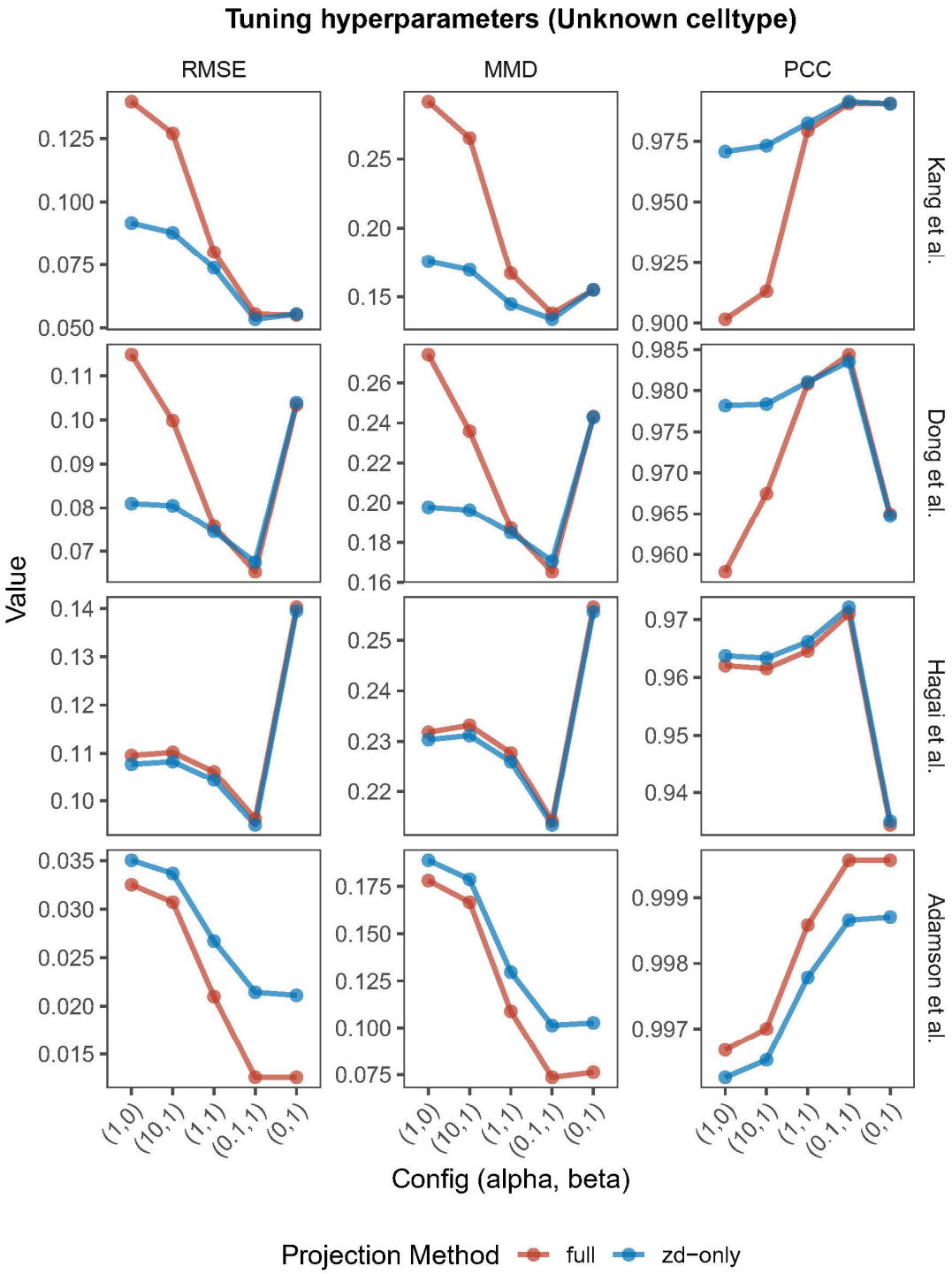
Supplementary Figures S4-S12



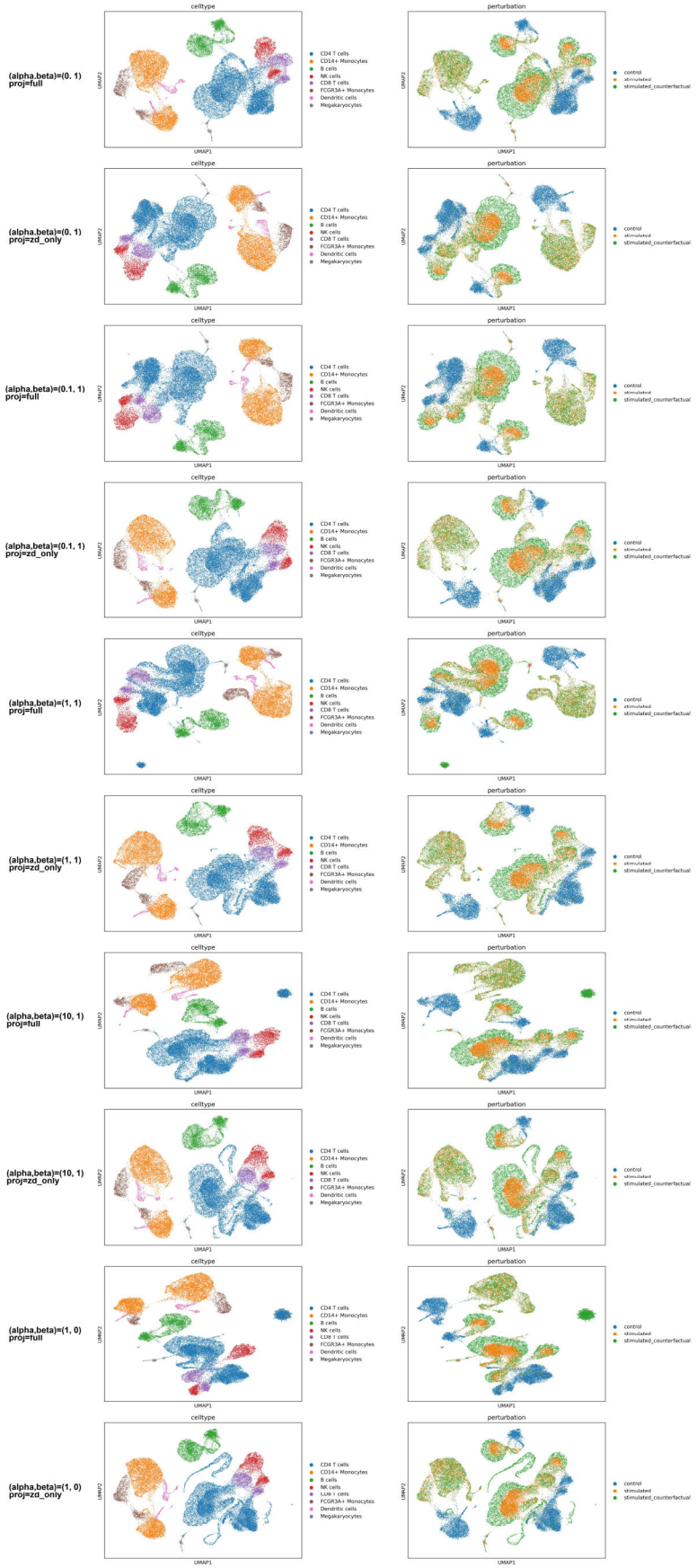
Supplementary Fig.S4 Performance on counterfactual predictions with simulated data under known cell-type labels (a) and unknown cell-type labels (b) settings. We evaluated the RMSE and Pearson correlation coefficient (PCC) of the mean expression of each gene between observed treatment group and counterfactual predictions from control group, as well as the maximum mean discrepancy (MMD) between the observed and predicted distributions. All metrics are averaged across all perturbations and cell types in each simulation. The x-axis shows the dimension settings for z_d and z_u in the simulation, and the different colors correspond to different methods. We repeat each simulated setting 20 times and report the mean$\pm$1.96SE.



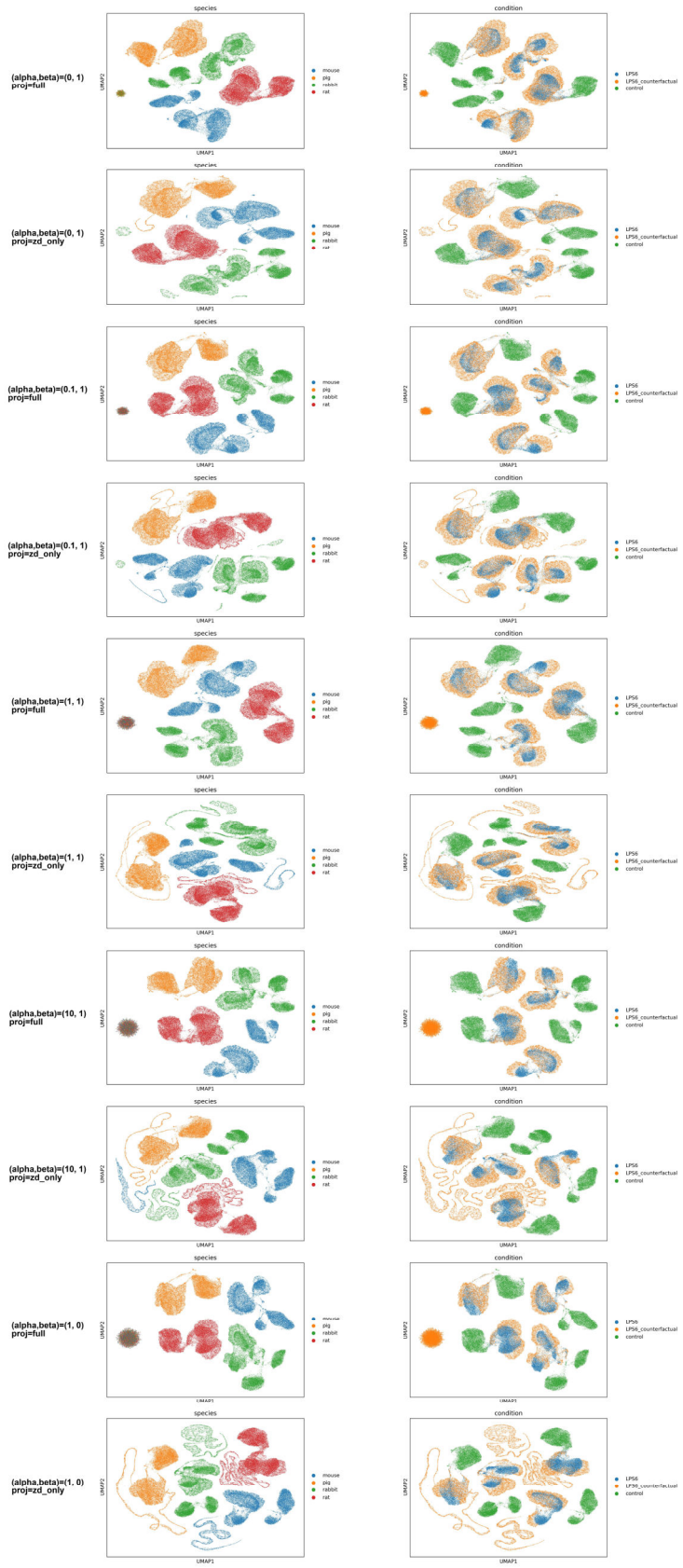
Supplementary Fig.S5 Performance on counterfactual predictions on real single-cell data without cell-type label information. The evaluation metrics includes RMSE and PCC of the mean expression of each gene between observed treatment group and counterfactual predictions from control group (we compare the mean since the true counterfactual outcome of control cells are not accessible in real data), as well as MMD between the observed and predicted distributions. All metrics are averaged across all perturbations and cell types.



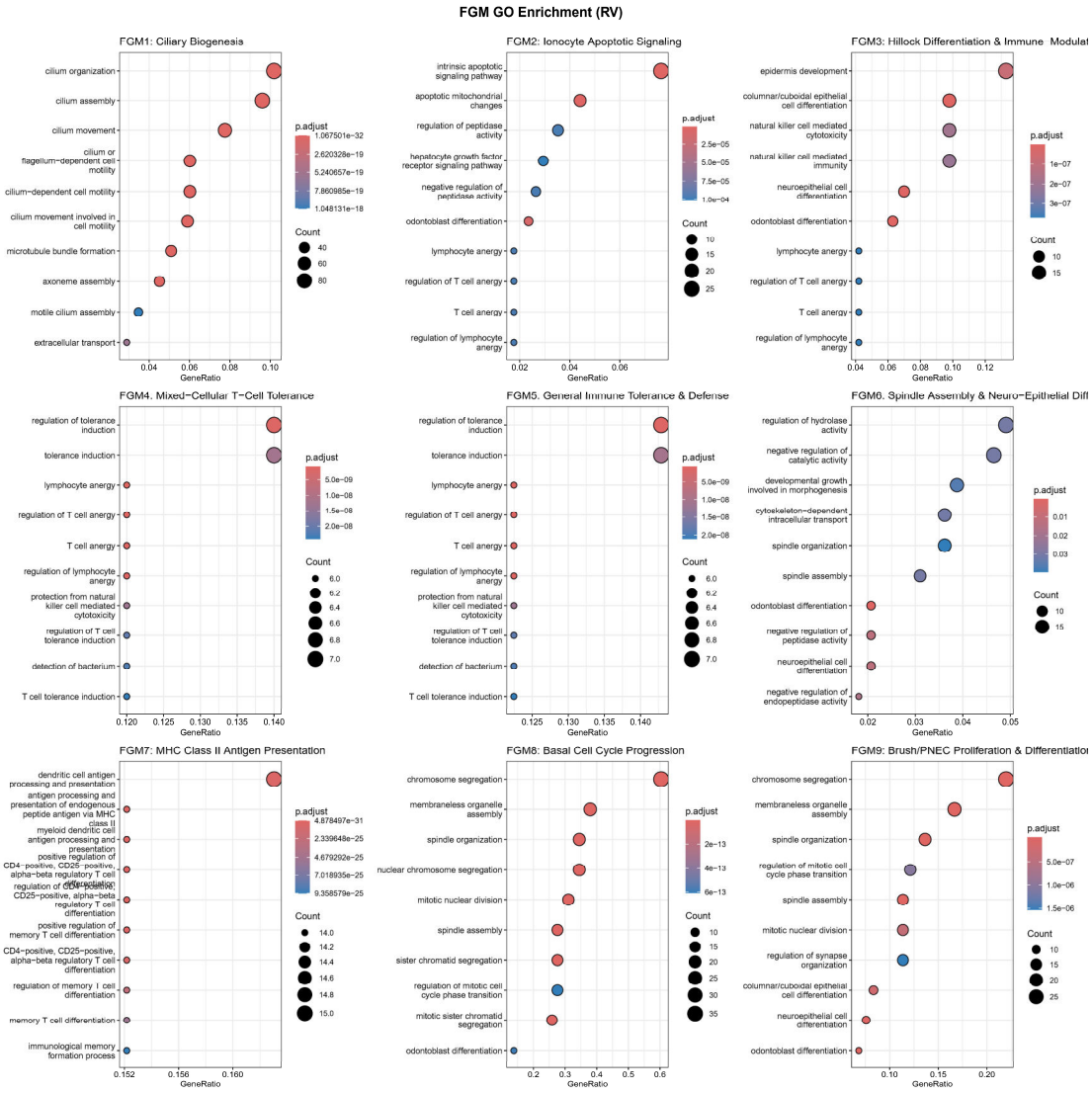
Supplementary Fig.S6 Sensitivity analysis on hyperparameters of conditional OT on real data under the unknown cell-type labels setting.



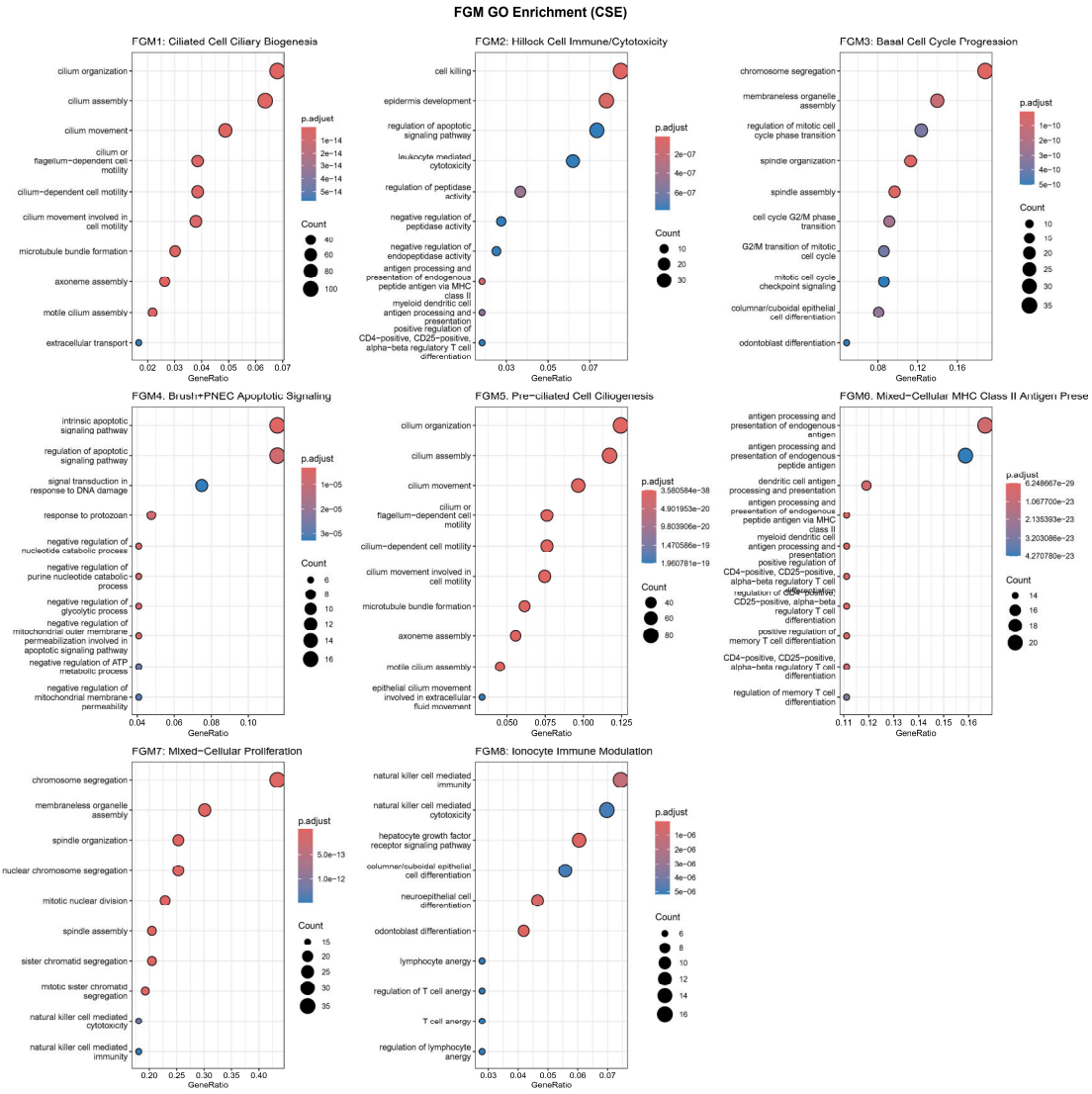
Supplementary Fig.S7 UMAP visualization of counterfactual predictions on Kang dataset under different settings of alpha, beta, and barycentric projection methods (known cell types).



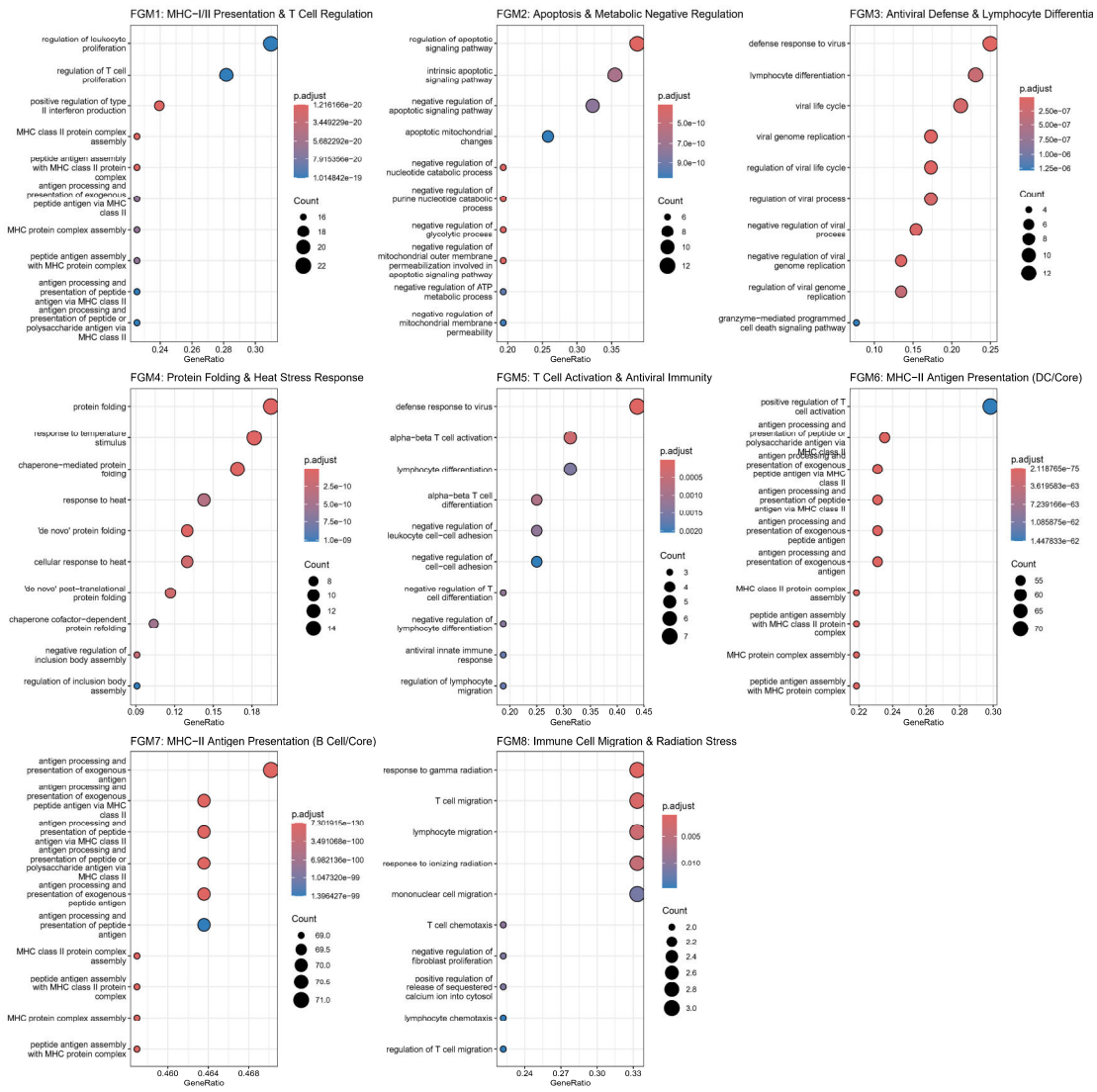
Supplementary Fig.S8 UMAP visualization of counterfactual predictions on Hagai dataset under different settings of alpha, beta, and barycentric projection methods (known cell types).



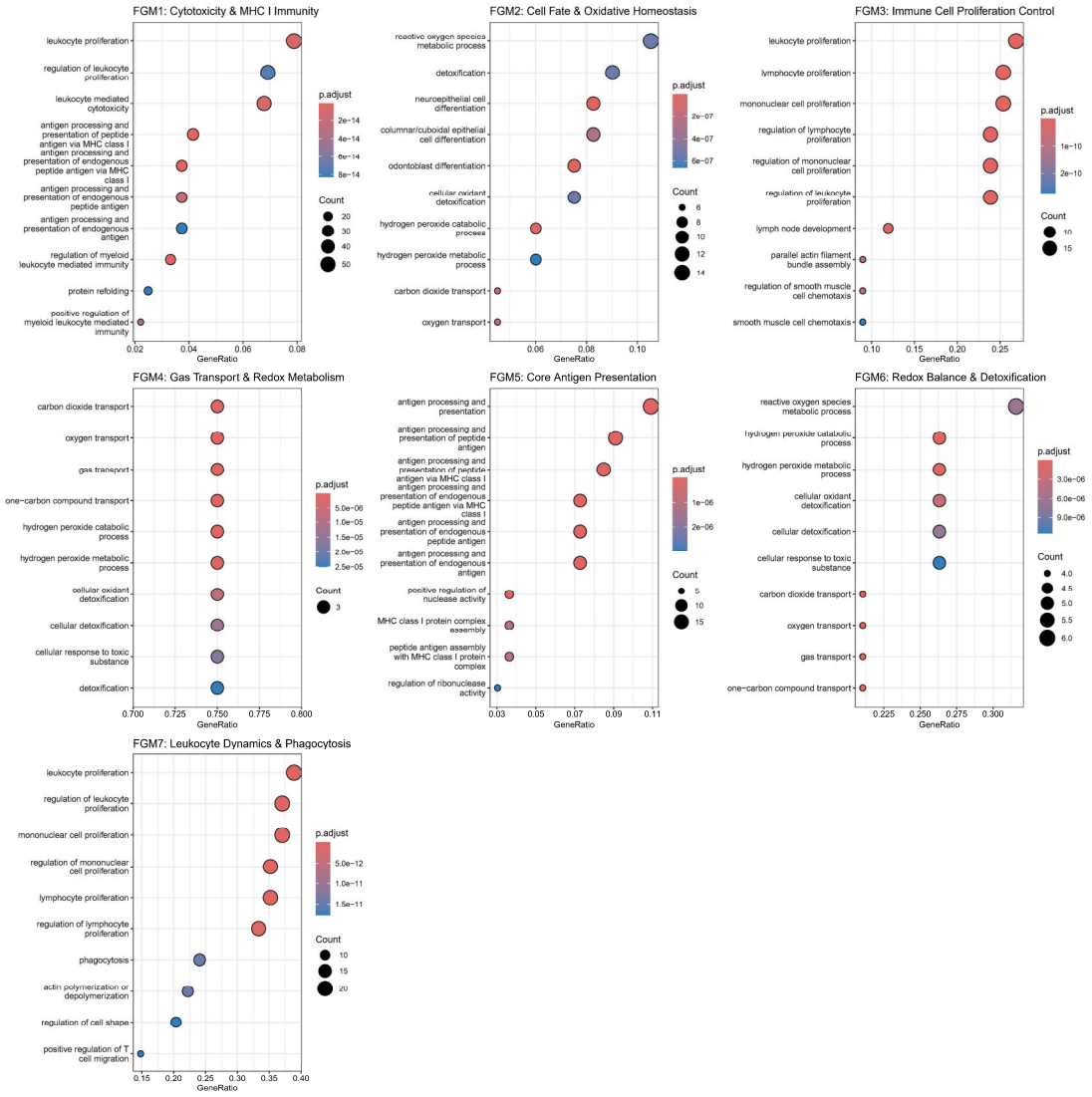
Supplementary Fig.S9 Details on GO enrichment of FGM identified in primary human bronchial organoids under rhinovirus (RV) exposure.



Supplementary Fig.S10 Details on GO enrichment of FGM identified in primary human bronchial organoids under cigarette-smoke extract (CSE) exposure.



Supplementary Fig.S11 Details on GO enrichment of FGM identified from Kang dataset.



Supplementary Fig.S12 Details on GO enrichment of FGM identified from Adamson dataset.

Supplementary Tables

Supplementary Table S1 Details on the FGM identified in primary human bronchial organoids under rhinovirus (RV) exposure.

FGM ID	Module Name	Functional Description
FGM1	Ciliary Biogenesis Module	This module is overwhelmingly dominated by Pre-ciliated cells and Doublets. Its functional signature is characterized by an exceptionally strong enrichment for pathways related to ciliary construction, including 'cilium organization', 'cilium assembly', 'axoneme assembly', and 'motile cilium assembly'. This reflects a highly active program of ciliogenesis localized to this specific cell population.
FGM2	Ionocyte Apoptotic Signaling Module	A highly specific module composed almost exclusively of Ionocytes. It is functionally defined by the 'intrinsic apoptotic signaling pathway' and 'apoptotic mitochondrial changes', indicating a distinct, cell-type-specific apoptotic response to the RV treatment.
FGM3	Hillock Differentiation & Immune-Modulatory Module	This module is driven by Hillock cells and exhibits a dual function. It is enriched for epithelial development terms like 'columnar/cuboidal epithelial cell differentiation' while concurrently showing a strong signature for innate and adaptive immune modulation, including 'natural killer cell mediated cytotoxicity' and 'regulation of T cell anergy'.
FGM4	Mixed-Cellular T-Cell Tolerance Module	A compositionally heterogeneous module with contributions from Ciliated, Basal, and Secretory cells. Functionally, it shows a highly coherent signature for immune suppression, strongly enriched for 'regulation of tolerance induction', 'T cell anergy', and 'T cell tolerance induction'.
FGM5	General Immune Tolerance & Defense Module	Similar to FGM4, this is another mixed-cellularity module defined by immune tolerance pathways such as 'tolerance induction' and 'regulation of T cell anergy'. It uniquely includes terms for 'protection from natural killer cell mediated cytotoxicity', suggesting a role in both modulating and evading immune responses.

FGM6	Spindle Assembly & Neuro-Epithelial Differentiation Module	This mixed module, with notable contributions from Secretory and Basal cells, is linked to cell division and differentiation. It is significantly enriched for 'spindle organization' and 'spindle assembly', alongside specialized differentiation terms 'odontoblast differentiation' and 'neuroepithelial cell differentiation'.
FGM7	MHC Class II Antigen Presentation Module	A key immune response module with broad cellular contribution, primarily from Secretory, Basal, and Ciliated cells. Its function is unambiguously centered on adaptive immunity, defined by enrichment for 'dendritic cell antigen processing and presentation' and 'antigen processing and presentation of endogenous peptide antigen via MHC class II'.
FGM8	Basal Cell Cycle Progression Module	This module is almost entirely composed of 'Cycling basal' cells and represents a canonical cell cycle program. It is strongly enriched for all phases of mitosis, including 'chromosome segregation', 'spindle organization', 'mitotic nuclear division', and 'mitotic sister chromatid segregation', reflecting active proliferation of the basal stem cell compartment.
FGM9	Brush/PNEC Proliferation & Differentiation Module	Driven almost exclusively by the Brush+PNEC cell population, this module captures a proliferative response within this rare cell type. It is highly enriched for cell cycle terms like 'chromosome segregation' and 'spindle organization', but also includes epithelial differentiation terms, suggesting a coupled proliferation and differentiation program.

Supplementary Table S2 Details on the FGM identified in primary human bronchial organoids under cigarette-smoke extract (CSE) exposure.

FGM ID	Module Name	Functional Description
FGM1	Ciliated Cell Ciliary Biogenesis Module	This module is predominantly composed of Ciliated cells. Its functional signature is a strong enrichment for pathways related to ciliary structure and function, including 'cilium organization', 'cilium assembly', 'cilium movement', and 'motile cilium assembly'.

FGM2	Hillock Cell Immune/Cytotoxicity Module	This module is highly specific to Hillock cells. It is functionally defined by immune response pathways, notably 'leukocyte mediated cytotoxicity' and 'regulation of peptidase activity', suggesting a role in cell-mediated defense and protein regulation under CSE stress.
FGM3	Basal Cell Cycle Progression Module	Driven by Basal and Cycling basal cells, this module represents a classic proliferative response. It is strongly enriched for all phases of mitosis, including 'chromosome segregation', 'spindle organization', and 'cell cycle G2/M phase transition'.
FGM4	Brush+PNEC Apoptotic Signaling Module	This module is uniquely dominated by the Brush+PNEC cell population. Its GO terms are tightly focused on 'intrinsic apoptotic signaling pathway' and 'regulation of apoptotic signaling pathway', indicating a specific pro-apoptotic ITE in this cell type.
FGM5	Pre-ciliated Cell Ciliogenesis Module	This module is almost exclusively composed of Pre-ciliated cells. Similar to FGM1 but in a precursor population, it is overwhelmingly enriched for the machinery of ciliary construction, including 'cilium organization', 'cilium assembly', and 'axoneme assembly'.
FGM6	Mixed-Cellular MHC Class II Antigen Presentation Module	This is a heterogeneous module with contributions from Basal, Ciliated, and Secretory cells. Its function is clearly defined by adaptive immunity, showing strong enrichment for 'antigen processing and presentation of endogenous peptide antigen via MHC class II' and 'dendritic cell antigen processing and presentation'.
FGM7	Mixed-Cellular Proliferation Module	A compositionally mixed module, with inputs from Cycling basal and Secretory cells. Like FGM3, it is defined by cell cycle progression, with enrichment for 'chromosome segregation', 'spindle organization', and 'spindle assembly', suggesting a proliferative response outside the main basal compartment.
FGM8	Ionocyte Immune Modulation Module	This module is driven by Ionocytes and is characterized by immune modulation pathways. It is enriched for

'natural killer cell mediated cytotoxicity', 'regulation of T cell anergy', and 'lymphocyte anergy'.

Supplementary Table S3 Details on the FGM identified in Kang dataset.

FGM NO.	SHORT NAME	DETAILED FUNCTIONAL DESCRIPTION (GO ENRICHMENT SUMMARY)
FGM 1	MHC I/II Presentation & T Cell Regulation	Highly enriched in MHC I/II antigen processing and presentation (both classes) and the Regulation of leukocyte and T cell proliferation. Also includes positive regulation of Type II IFN production.
FGM 2	Apoptosis & Metabolic Negative Regulation	Centered on the Regulation of apoptotic signaling pathway and Negative regulation of metabolic processes, including purine nucleotide catabolism and glycolysis, often via mitochondrial pathways.
FGM 3	Antiviral Defense & Lymphocyte Differentiation	Core function is the Defense response to virus and Lymphocyte differentiation. Key features include Viral genome replication regulation and Granzyme-mediated programmed cell death signaling.
FGM 4	Protein Folding & Heat Stress Response	Highly specialized for Protein folding (‘de novo’, chaperone-mediated) and the Response to heat or temperature stimulus. Reflects the cellular machinery maintaining proteostasis.
FGM 5	T Cell Activation & Antiviral Immunity	Characterized by a strong Defense response to virus and the Alpha−beta T cell activation and differentiation, including negative regulatory aspects of these processes.
FGM 6	MHC II Antigen Presentation (DC/Core)	Highly specific for the Antigen processing and presentation of exogenous peptide via MHC class II. Also includes MHC class II protein complex assembly and Positive regulation of T cell activation.
FGM 7	MHC II Antigen Presentation	Functional profile is nearly identical to FGM6, focusing on Antigen processing and presentation of exogenous antigen via

	(B Cell/Core)	MHC class II, confirming this as a central APC response.
FGM 8	Immune Cell Migration & Radiation Stress	Defined by Mononuclear/Lymphocyte/T cell migration and chemotaxis. Also includes the Response to gamma and ionizing radiation and Positive regulation of calcium ion release.

Supplementary Table S4 Details on the FGM identified in Adamson dataset.

FGM ID	SHORT NAME	DETAILED FUNCTIONAL DESCRIPTION (GO TERMS)
FGM1	Cytotoxicity & MHC I Immunity	Highly concentrated on leukocyte mediated cytotoxicity, MHC class I antigen processing and presentation (of peptide and endogenous antigen), and leukocyte proliferation. This module is a core driver of CD8+ T cell and NK cell immune responses.
FGM2	Cell Fate & Oxidative Homeostasis	Involved in specific cell differentiation processes (e.g., odontoblast, neuroepithelial cell differentiation) and the regulation of reactive oxygen species metabolic process, hydrogen peroxide catabolism, and general detoxification.
FGM3	Immune Cell Proliferation Control	Focused entirely on the positive and negative regulation of leukocyte, lymphocyte, and mononuclear cell proliferation. It also includes lymph node development. This module controls the numerical expansion of immune cells during a response.
FGM4	Gas Transport & Redox Metabolism	Primary functions include carbon dioxide transport, oxygen transport, and one-carbon compound transport. Also heavily involved in cellular oxidant detoxification and hydrogen peroxide catabolic process, maintaining the cell's metabolic and redox stability.
FGM5	Core Antigen Presentation	Highly specific for the machinery of the MHC I pathway: antigen processing and presentation (via MHC I), MHC class I protein complex assembly, and peptide antigen assembly. This is a fundamental component of cell-mediated immunity.
FGM6	Redox Balance &	A metabolic and protective module focused on ROS and H2 O2 catabolism, cellular oxidant detoxification, and the cellular response to toxic substances. Also includes gas transport

	Detoxification	functions.
FGM7	Leukocyte Dynamics & Phagocytosis	Functions related to leukocyte/lymphocyte proliferation, phagocytosis, T cell migration, and actin polymerization/cell shape regulation. This module governs the rapid growth, movement, and particle-clearing ability of immune cells.