Causal identification of single-cell experimental perturbation effects with CINEMA-OT

Mingze Dong[1,2], **Bao Wang**[3,4], **Jessica Wei**[4,5], **Curt Perry**[4,5], **Alexander Frey**[4,5], **Feriel Ouerghi**[4,5], **Ellen F. Foxman**[3,4†], **Jeffrey J. Ishizuka**[2,4,5†], **Rahul M. Dhodapkar**[6†], **David van Dijk**[1,7,8†]
[1] Interdepartmental Program in Computational Biology & Bioinformatics, Yale University
[2] Department of Pathology, Yale School of Medicine
[3] Department of Laboratory Medicine, Yale School of Medicine
[4] Department of Immunobiology, Yale School of Medicine
[5] Department of Medical Oncology, Yale School of Medicine
[6] Department of Ophthalmology and Visual Science, Yale School of Medicine
[7] Department of Internal Medicine (Cardiology), Yale School of Medicine
[8] Department of Computer Science, Yale University
† Co-corresponding authors
Correspondence should be addressed to: david.vandijk@yale.edu

Abstract

Recent advancements in single-cell technologies allow characterization of experimental perturbations at single-cell resolution. While methods have been developed to analyze data from such experiments, the application of a strict causal framework has not yet been explored for the inference of treatment effects at the single-cell level. In this work, we present a causal inference based approach to single-cell perturbation analysis, termed CINEMA-OT (Causal INdependent Effect Module Attribution + Optimal Transport). CINEMA-OT separates confounding sources of variation from perturbation effects to obtain an optimal transport matching that reflects counterfactual cell pairs. These cell pairs represent causal perturbation responses permitting a number of novel analyses, such as individual treatment effect analysis, response clustering, attribution analysis, and synergy analysis. We benchmark CINEMA-OT on an array of treatment effect estimation tasks for several simulated and real datasets and show that it outperforms other single-cell perturbation analysis methods. Finally, we perform CINEMA-OT analysis of two newly-generated datasets: (1) rhinovirus and cigarette smoke-exposed airway organoids, and (2) combinatorial cytokine stimulation of immune cells. In these experiments, CINEMA-OT reveals potential mechanisms by which cigarette smoke exposure dulls the airway antiviral response, as well as the logic that governs chemokine secretion and peripheral immune cell recruitment.

1 Introduction

Cellular responses to environmental signals are a fundamental component of biological functioning, playing an integral role in both homeostasis and disease [1]. For decades, controlled perturbation experiments have been used to reveal the underlying mechanisms of biological processes. Recent advances in single-cell technologies allow complex experiments measuring high dimensional phenotypes at high throughput under diverse stimulation conditions [2–8]. However, deriving biological insights from these experiments remains a challenge.

For the analysis of single-cell perturbation data, several approaches have been developed, which can be categorized as follows:

Preprint version.

1. *Differential expression* methods test for statistically significant differences in gene expression between cell populations. [9–13]

2. *Differential abundance* methods formulate the quantification problem as differential abundance testing in a continuous cellular state manifold. [14–17]

3. *Perturbation analysis* methods aim to reveal underlying biological processes by modeling perturbations in a latent space with either linear models or neural networks. [3–7, 18–21]

While techniques to characterize the effects of perturbations by averaging over populations have been extensively used in the analysis of single-cell data, methods allowing for causal single-cell perturbation analysis have not yet been extensively explored. In causal inference, the quantification of responses to perturbations is known as the *treatment effect estimation problem* [22]. In subsequent portions of this manuscript, we will borrow from the terminology of causal inference, referring to perturbations and treatments, as well as response and treatment effect, interchangeably. Ideal causal methods allow for the direct characterization of underlying confounding variation, which is not accounted for in any of the existing single-cell analysis methods.

Variability in cellular responses to treatment may be attributable to underlying confounding variation. [23]. In the case of scRNA-seq experiments, sources of variation such as cell cycle stage, microenvironment, and pre-treatment chromatin accessibility may all act as confounding factors when performing treatment effect estimation [18]. Collectively, confounding factors can be thought of as a cell's context that may both influence a cell's underlying gene expression profile, and condition treatment-induced gene signatures. Correct identification of confounders allows for appropriate causal matching of cell pairs between conditions, enabling treatment effect estimation at the single-cell level.

One well-established confounding factor that may affect treatment response is cell type. For example, widely used nucleoside analog chemotherapeutics such as 5-fluorouracil (5-FU) act selectively on cells in the DNA synthesis phase of the cell cycle, killing cancer cells while minimizing effects on healthy tissue [24]. Some mutations may also drive differential response to a stimulation, as is seen with some tumors in response to TGF-β [25]. Confounders may be latent or unobserved, such as different exposures of cells to a drug, which may have different effects at different concentrations within each individual cell.

We aim to solve this problem by introducing a causal framework permitting characterization of perturbation effects at the single cell level. We explicitly model the diversity of cellular responses to perturbation across confounding factors, which does not require prior knowledge of cell identifies. With this approach, we can identify cell state-conditioned transcriptional changes that drive pathology or pharmacologic response to perturbation.

In this paper, we present CINEMA-OT (Causal INdependent Effect Module Attribution + Optimal Transport), which applies independent component analysis (ICA) and filtering based on a functional dependence statistic to identify confounding factors and treatment-associated factors. CINEMA-OT then applies weighted optimal transport (OT) to achieve causal individual matching. The algorithm is based on a causal inference framework for modeling confounding signals and conditional perturbation effects at the single-cell level. We show that the model is uniquely indentifiable via the theory of ICA. The computed causal cell matching enables a multitude of novel downstream analyses, including but not limited to: individual treatment effect estimation, treatment synergy analysis, sub-cluster level biological process enrichment analysis, and attribution of perturbation effects.

We demonstrate the power of CINEMA-OT by benchmarking it on several simulated and real datasets and comparing it to widely used single-cell level perturbation analysis methods. We then perform CINEMA-OT analyses of two newly-generated datasets. In the first, we examine the effects of viral infection and cigarette smoke on innate immune responses in airway organoids. In the second, we perform combinatorial cytokine stimulation of *ex vivo* peripheral blood mononuclear cells in order to characterize how cytokines act in concert to shape immune responses.

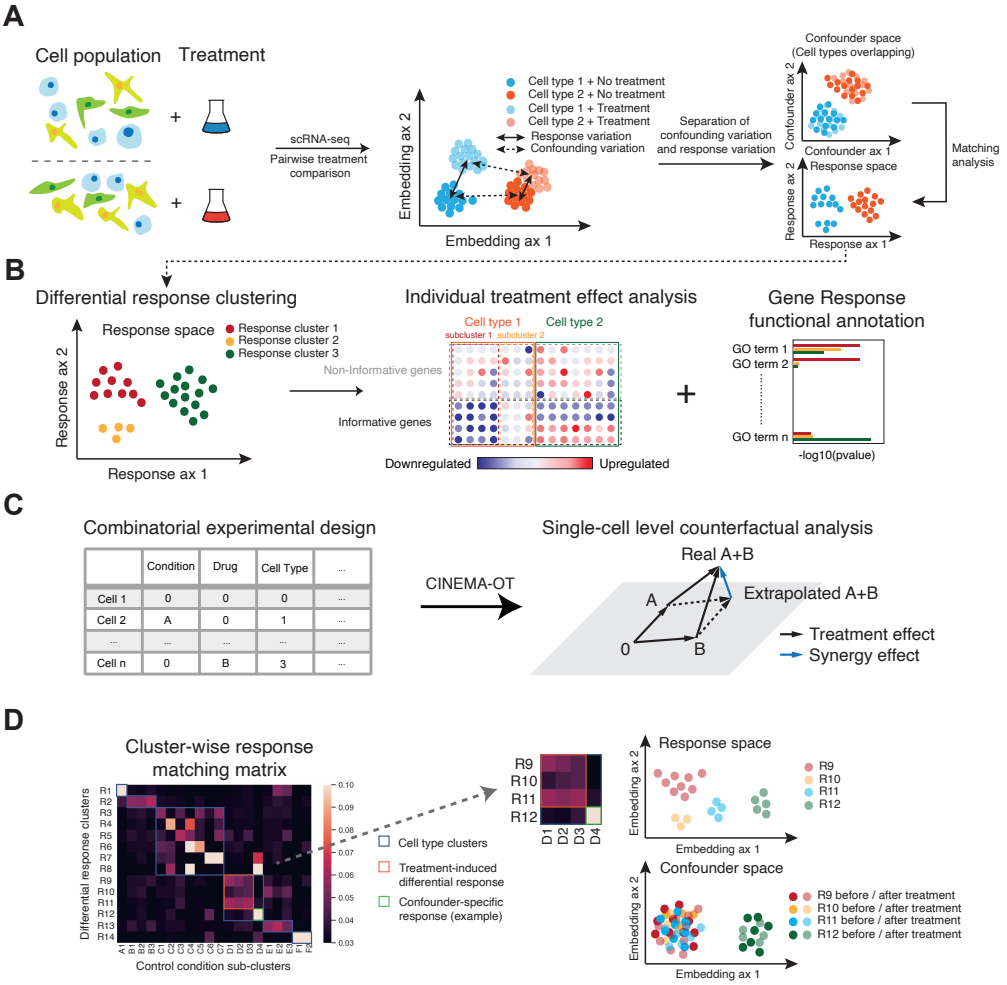

Figure 1: A causal framework for single-cell level perturbation effect analysis. **A.** In single-cell experiments, cells are separated by both treatment batches and latent cell states. Latent states of cells have confounding effects on the response to perturbation. Upon successfully separating confounding variation and response variation in the measurements, we can identify confounding signals for each cell and match counterfactual cell pairs across conditions to compute causal perturbation effects. **B.** After characterizing the differential response matrices at a single-cell level, we can subcluster cells by treatment responses. These responses may be further characterized by other tools, such as gene set enrichment analysis. **C.** We are able to quantify synergy in combinatorial perturbations by evaluating the dissimilarity of extrapolated phenotypes and true combinatorially perturbed phenotypes. **D.** CINEMA-OT can attribute divergent treatment effects to either explicit confounders or latent confounders using the cluster-wise response matching matrix.

2 Results

2.1 Confounder signal matching via CINEMA-OT

To perform causal single-cell perturbation effect inference, we have adopted the *potential outcome causal framework* [22, 26]. Ideally, to generate causal assertions about the effect of a perturbation on the transcriptional state of a given cell, we would like to measure the same cell both before and after the perturbation is applied. However, the process of obtaining transcript measurements from single cells is destructive, and an individual cell may only be measured once. In fact, to the best of our knowledge, all high throughput genomic measurement technologies are destructive. A solution would be to infer *counterfactual* cell pairs, which are inferred, causally-linked pairs—predictions of what a cell in one condition would look like in another condition. The potential outcome framework formalizes this concept by establishing a statistical model that describes outcome variables as a

function of confounding factors and treatment-associated factors. In our setting, to infer single-cell treatment effects, this would translate to distinguishing the effects of biological variation and treatment on treatment-associated gene expression.

In the potential outcome framework, a key difficulty for general unsupervised causal inference is the mixing of confounders with outcomes. In the field of causal inference, this is described as learning with both interventions and latent confounding [27]. In our case, a gene can contribute to confounding variation as well as treatment-associated variation. To apply the tools of classical causal inference, confounding factors must first be distinguished from treatment-associated factors. Notably, confounding factors and treatment-associated factors may be treated as a low dimensional function of the gene space.

To unmix confounding effects and treatment-associated effects, we propose two sufficient assumptions, which can be potentially relaxed:

> **Assumption 1: Independent sources and noise.** *Confounding factors and treatment events are pairwise independent random variables.*
>
> This assumption relies on treated and untreated cells being drawn from the same underlying set of cells. In practice, this is a central part of most single-cell experimental designs.
>
> **Assumption 2: Linearity of source signal combinations.** *Confounding gene signatures can be modeled as a linear combination of confounding sources plus an independent noise term. The outcome gene signatures can be modeled as arbitrary functions of confounding factors and treatment events plus an independent noise term.*
>
> While modeling total expression as a linear combination of relevant factors may be unable to capture complex nonlinear relationships between confounding variables, this assumption is necessary in our case, as the nonlinear source separation problem is generally not identifiable. Moreover, we use the confounders as a intermediate metric to construct a matching across similar cells, therefore full characterization of nonlinear dependency is not necessary in our method. Meanwhile, our assumption allows nonlinear interactions between confounders and treatments, thereby permitting modeling of general non-linear treatment effects, including confounder-specific treatment effects.

Given these assumptions, we can strictly prove that the identification of confounding factors is equivalent to solving the blind source separation problem (BSS), which can be done by the ICA algorithm (see Methods for the proof).

After ICA transformation, the gene count matrix is decomposed into a linear combination of confounding signals, treatment-associated signals, and independent noise signals. In practice, we filter the noise signal by PCA dimensionality reduction prior to ICA. To identify treatment signals, we use a non-parametric distribution-free test based on Chatterjee's coefficient to quantify functional dependence between each identified signal and the ground truth treatment signal [28] (Figure 2A).

Finally, with the identified confounding factors, we apply optimal transport with entropic regularization to generate causally matched counterfactual cell pairs. Optimal transport is a natural choice for this matching procedure as it preserves mass, is robust to outliers, and avoids collapsing matches at the boundaries of separated clusters within the data [29, 30]. While solving the optimal transport problem is often prohibitively resource-intensive for large-scale biological data, CINEMA-OT considers the tractable case of entropic regularization [31, 29]. Optimal transport with entropic regularization can be formulated as a convex optimization problem which can be solved efficiently using the alternating direction method (Sinkhorn-Knopp algorithm [31, 29]).

2.2 Causal matching in the setting of differential abundance

A treatment may change the distribution of cell densities, e.g. cells may die or proliferate in response to some perturbation. Thus, we may have an additional factor of differential confounder abundance across experimentally perturbed datasets. This factor can come from either real biological effects or batch effects. The differential abundance confounder can affect the performance of CINEMA-OT since in this case the underlying confounders are no longer independent of the treatment event

and assumption (1) is violated. Indeed, our experiments have shown that while CINEMA-OT can tolerate moderate levels of differential abundance (Extended Figure 2), it can fail when high levels of differential abundance are present (Extended Figure 3).

In order to correct for the effects caused by differential abundance, we have developed an iterative reweighting procedure for CINEMA-OT. In each iteration step, we estimate likelihood weights of the treatment events (for details, see Methods) (Figure 2C). By balancing the local distribution of cells in the confounder space using these weights, we can correctly match cells.

We note that this functionality should be used only when required. As iterative reweighting within CINEMA-OT relaxes assumption 1 in our framework, the identifiability of our model can no longer be guaranteed, which may reduce its ability to identify certain classes of cellular responses. For example, aligning cell types with differential abundance across treatment conditions eliminates the possibility that one cell type can convert to another cell type upon treatment. In such a case, cell type can be a treatment-induced factor instead of a confounder. Accounting for this, in order to integrate prior biological knowledge, CINEMA-OT also provides an option to assign weights according to user-provided labels (e.g. cell-types). In this case, CINEMA-OT can assign weights using confounder labels instead of automatically balancing over all possible covariates.

2.3 Causal matching enables individual treatment effect and synergy analyses

With the matched counterfactual cell pairs computed by CINEMA-OT, we are able to obtain two key outputs: (1) the matching correspondence matrix across treatment conditions, and (2) the individual treatment effect (ITE) for each cell with its counterfactual pair across treatments (Figure 1A).

Individual treatment effect matrices are gene by cell matrices which can be clustered and visualized by existing scRNA-seq computational pipelines. By clustering over an ITE matrix, we can identify sub-clusters within cell types with heterogeneous response to treatments. We may perform statistical analysis to identify the significantly affected genes and identify their coordinated biological function by gene set enrichment analysis (Figure 1B).

In addition, in a combinatorial perturbation experiment, we are able to define a synergy effect metric by comparing the predicted effect of combining multiple treatments with the observed effect of combined treatment (Figure 1C). We define this synergy metric by estimating the difference between the true sample under combinatorial treatment (A+B) and the predicted sample by adding the effects of treatment A and treatment B, thus assuming purely linear, non-interactive effects. If no difference is measured, we may conclude that there are no nonlinear or interaction effects between the treatments. If non-zero synergy is present, this points to some interaction between treatments A and B. Synergy is computed for every cell-gene pair, resulting in a matrix of equivalent form to expression or ITE matrices - a unique feature of CINEMA-OT.

Another important task in perturbation effect analysis is the attribution of treatment effects. Differential response can be driven either by differences in explicit confounding factors or by latent factors, such as treatment heterogeneity. Because CINEMA-OT provides a single-cell level matching as one output, the task can be solved by coarse-graining the matching matrix. Responses that cluster both in response as well as confounder space may be attributed to explicit confounding factors. Conversely, responses that cluster well in the response space but do not demonstrate clustering in the confounder space may be attributed to latent factors (Figure 1D). Such an analysis can be performed either at the cell type level or at the sub-cluster level to reveal underlying heterogeneity. In order to further identify genes with significant confounder-specific treatment effects, we quantify the confounder effect size via a causal regression model and estimate significance using the ratio of confounder-explained effect size to the residual norm (see Methods for additional details).

2.4 Validation of CINEMA-OT using simulated data

There are a number of existing methods that perform single-cell level perturbation effect analysis [3–7, 18–20, 32, 33]. Supplementary Figure 1 comprises a summary table of currently available methods and their capabilities. To investigate how CINEMA-OT differs from these methods, we perform extensive benchmarking on a number of tasks.

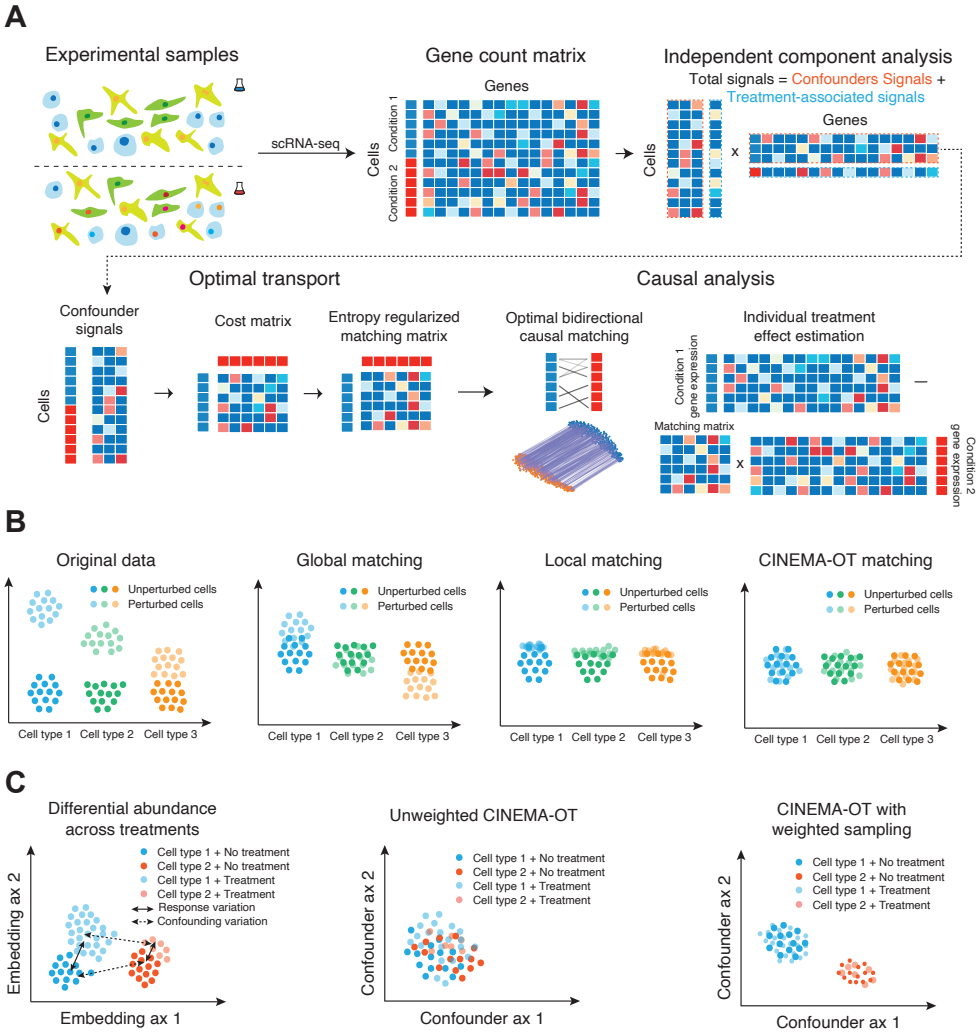

Figure 2: Overview of the CINEMA-OT framework. **A.** scRNA-seq count data is first decomposed into confounder variation and treatment-associated variation using ICA. Cells are then matched across treatment conditions by entropy-regularized optimal transport in the confounder space to generate a causal matching plan. The smooth matching map can then be used to estimate individual treatment effects. **B.** Illustration of the properties of CINEMA-OT compared to other common matching schemes. Global matching may have poor performance when there are confounder-specific heterogeneous responses to treatment. Local matching may be susceptible to boundary effects. By contrast, CINEMA-OT balances these concerns by enforcing preservation of distributional mass. **C.** Differential abundance effects may cause spurious matching by CINEMA-OT. When such effects are present, iterative reweighting may be used to balance cell populations and learn true underlying confounding signals.

We systematically compare methods including Mixscape, Harmony-Mixscape, optimal transport without CINEMA-OT, and a recently proposed variational inference approach (Contrastive-VI), to CINEMA-OT. Our comparison is based on three categories of metrics:

1. *Cell distribution equalization after treatment effect removal.* In data sets with or without ground truth, we can measure the validity of treatment effects by examining cell distributions in the gene expression space after treatment effect removal. If different treatments are applied to the same confounder distribution (Assumption 1, see Section 2.1), then these distributions should overlap well after treatment effects are removed. Metrics for evaluating

treatment effect removal include average silhouette width (ASW), principal components regression score (PCR), and graph connectivity.

2. *Differential response cluster preservation.* If a cell population has divergent responses to a perturbation, the cell population would form clustering structures in the response space. Therefore, preservation of such clustering structures in the estimated treatment effects is essential for perturbation effect identification. In this study, we evaluate the cluster preservation level via an adjusted Rand index (ARI) in individual treatment effect (ITE) matrices.

3. *Attribution accuracy.* Differential response patterns can be attributed to either confounder-specific effects (e.g. cell-type specific effects) or latent-factor driven effects (e.g. treatment drug dose distribution). In simulated data, the attribution accuracy can be measured via independence between confounding factors and responses conditioned on ground truth response labels. In our study this is evaluated by the PC regression score (PCR) in ITE matrices.

To obtain data with ground truth, we simulate data using the R package Splatter and Python package Scsim. Splatter is a popular package [34] for single-cell *in silico* data generation, and Scsim is a python implementation of the Splatter framework with additional support for simulating gene regulation programs as trajectories. [35] The genes in our simulated data are separated into three subsets, corresponding to the underlying trajectory, cell types, and treatment-associated genes respectively. We model the dependence between confounders and ground-truth treatment effects in several settings: *1. Overall treatment effects* modeling coherent global gene expression shifts across confounders; *2. Confounder-specific treatment effects* modeling diverging responses driven by underlying confounders, such as cell-type specific response; *3. Latent-factor driven treatment effects* modeling the differential treatment effect caused by unobservable latent confounders.

We used combinations of the aforementioned categories of metrics to evaluate the ability of CINEMA-OT and other methods to recover the dependence between confounders and treatment effects in each setting (see Methods for details). Finally, we combine these settings by modeling the differential response probabilities as conditional distributions on confounder labels, covering both confounder-specific effects and latent-factor driven effects (Figure 3B).

Our quantitative assessment of these synthetic datasets shows that CINEMA-OT achieves comparable performance to other tested methods in differential response cluster preservation (ARI > 0.95) (Figure 3C). Furthermore, CINEMA-OT attains superior performance in treatment effect removal, attribution of diverging responses to confounder states, and additional benchmarks when compared to other tested methods (Figure 3C, Supplementary Figure 2B).

2.5 Validation of CINEMA-OT using real data

In order to evaluate the performance of CINEMA-OT in a real setting, we use two publicly available single-cell transcriptomics datasets: (1) sequencing of entorhinal cortex in patients with Alzheimer's disease and unaffected controls [36]; (2) the sci-Plex4 drug perturbation dataset [8], which measures the response of the A549 and MCF7 cell lines to perturbation with 17 drugs. In both datasets, we do not know the ground truth perturbation effects or ground truth confounders at the single-cell level. Therefore, we only give the quantitative comparison for perturbation effect elimination while performing qualitative comparison for other categories of metrics.

In the Alzheimer's disease dataset, we focus on qualitative comparison of perturbation effect removal and differential response cluster preservation. While the first comparison can be conducted in an unsupervised manner, for the second comparison we integrate prior knowledge to evaluate the preservation of clusters of interest [37]. One notable example gene selected is SPP1, which has been previously described as up-regulated in some cell types of patients with Alzheimer's disease (e.g. microglia, some neuronal subtypes), but not in others (e.g. endothelial cells) [38, 36]. Our comparison here highlights that while Mixscape and Contrastive-VI can achieve either differential response cluster preservation or treatment effect removal, respectively, CINEMA-OT succeeds in both tasks (Supplementary Figure 6-7).

In the Sciplex dataset, we investigate the response to perturbation with Pracinostat, a histone deacytelase (HDAC) inhibitor, with the combinatorial induction of exogenous acetate, citrate, and pyruvate. HDAC inhibitors act as antitumoral agents through antagonizing the pro-transcriptional effects of

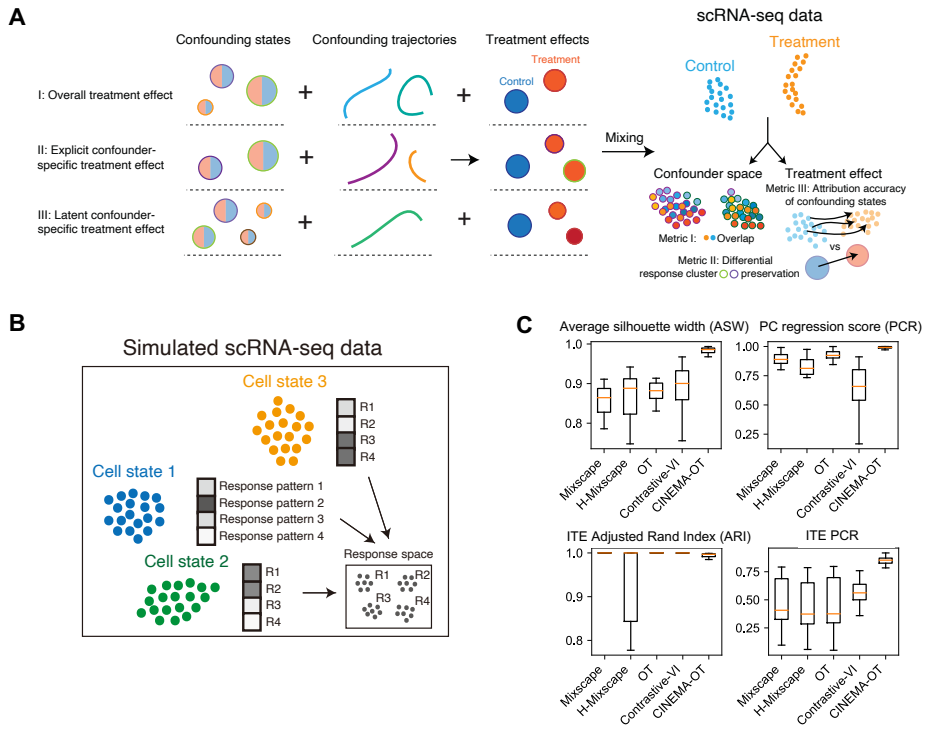

Figure 3: Benchmarking of CINEMA-OT against other methods for single-cell perturbation analysis. **A.** Illustrations of various factors to be considered in the task of single-cell level perturbation effect quantification. We evaluate CINEMA-OT against other methods using three classes of metrics: (I) mixing of samples in confounder space, (II) differential response cluster preservation in individual treatment response matrices, and (III) attribution accuracy of differential responses to confounding effects. **B.** Illustration of our conducted simulation study. **C.** Quantification of different validation metrics on synthetic data for CINEMA-OT and other methods.

histone deacetylation and silencing the expression of oncogenic factors through chromatin remodeling [39]. As HDAC inhibitors act partly through the deprivation of acetyl-CoA, we expect that the relative abundance of acetyl-CoA precursors within a cell would modulate the effect of HDAC inhibitor exposure, and acetyl-CoA precursors can be considered confounders [8] (Figure 3D). We find that CINEMA-OT can both match the distributions across treatment conditions and preserve the confounding labels, giving the best overall performance among the tested methods (Supplementary Figure 2C, see Methods for additional details).

Cell cycle stage is a well studied confounding covariate for experimental perturbation effects. However, in the Sci-plex dataset, we cannot evaluate the accuracy of cell cycle stage preservation since HDAC inhibition induces a G2/M cell cycle arrest [8]. As a result, to evaluate cell cycle preservation for the different methods, we quantify cell cycle information preservation after matching cells across cell lines. In order to get the cell cycle signal, we use the Tricycle package [40], which provides a prediction of a cell's phase in the cell cycle. We visualize cell cycle information preservation after matching cells across cell lines. CINEMA-OT is the only method that successfully matches cells between conditions while largely preserving the cell cycle information (Supplementary Figure 5).

2.6 CINEMA-OT identifies heterogeneous response patterns and synergistic effects of cigarette smoke exposure in rhinovirus infection

In order to demonstrate CINEMA-OTs ability to perform single-cell level experimental perturbation analysis, we have applied CINEMA-OT to new scRNA-seq data of rhinovirus infection in primary human bronchial organoids (Figure 4A). The experiment comprises 4 conditions: cigarette smoke extract (CSE) exposure, rhinovirus (RV) infection, the combination of rhinovirus and cigarette smoke (RVCSE), as well as a control condition (mock). While rhinovirus infection has been previously

studied [41], the goal of our study was to probe cellular defense responses to viral infection from each airway epithelial cell type in the presence or absence of a common environmental insult known to impact the outcome of rhinovirus infection: cigarette smoke. Previous studies of viral infection using this model considered gene expression in each cell type but not heterogenous response patterns, which may be of biological and clinical relevance in understanding the tissue response to respiratory virus infections [41, 42].

We first performed preprocessing of the dataset and annotated 8 cell clusters in total, including major cell types in the airway (basal, secretory, ciliated) and other rare (ionocyte, PNEC, and brush) or transitional (hillock and pre-ciliated) cell types (Figure 4B). We then performed CINEMA-OT analysis on mock-RV and mock-CSE condition pairs to identify single-cell level treatment effects. As expected, in response to rhinovirus infection, most epithelial cell types exhibited robust induction of the interferon response with upregulation of several interferon stimulated genes, including ISG15, IFI44, STAT1, MX1, and others (Figure 4C, 4D). [41]. In response to cigarette smoke extract exposure, a subset of epithelial cells increased expression of genes associated with metabolism of reactive oxygen species (PRDX1, TXN) and fatty acid and xenobiotic metabolism (ADH1C, ALDH3A1, CYP1B1). Interestingly, responses to CSE were primarily enriched in particular cell subpopulations including hillock, ciliated, and secretory cells, in contrast to the global interferon response seen following rhinovirus infection (Figure 4E, 4F). This demonstrates a functional division of defense mechanisms in airway epithelium, with cell type specific responses to different insults.

After analysis of the effect of cigarette smoke and viral infection individually, positive and negative synergy between these two insults was assessed by calculating cell-gene synergy scores (Figure 4G, H). Among the most significant synergistic effects, we found that interferon stimulated genes (ISGs) showed a global reduction during viral infection in the presence of cigarette smoke compared to viral infection alone (Figure 4I), consistent with previous mechanistic studies from our group showing that the antioxidant defense response induced by CSE suppresses signaling pathways required for induction of interferon-stimulated genes in response to viral RNA in airway epithelial stem cells [43].

In addition to a global attenuation of the interferon response, we discovered that pre-ciliated cells, in particular, exhibit pronounced synergistic expression of a distinct set of genes when co-exposed to RV and CSE (Figure 4G, 4I). Pre-ciliated cells, sometimes referred to as "deuterosomal" cells, are developing multiciliated cells with the marker genes CCNO and CDC20B [44]. Pre-ciliated cells co-exposed to both viral infection and CSE show synergistic induction of genes encoding secreted proteins that are typically associated with secretory cells in resting cultures, including SCGB3A1, LCN2, BPIFB1, SLPI, and WFDC2 (Figure 4I). This pattern could arise from pre-ciliated cells adopting a more secretory phenotype during co-exposure, or secretory cells adopting a pre-ciliated phenotype. These findings highlight the use of CINEMA-OT to identify synergistic effects on gene expression induced by co-exposure to viral infection and cigarette smoke.

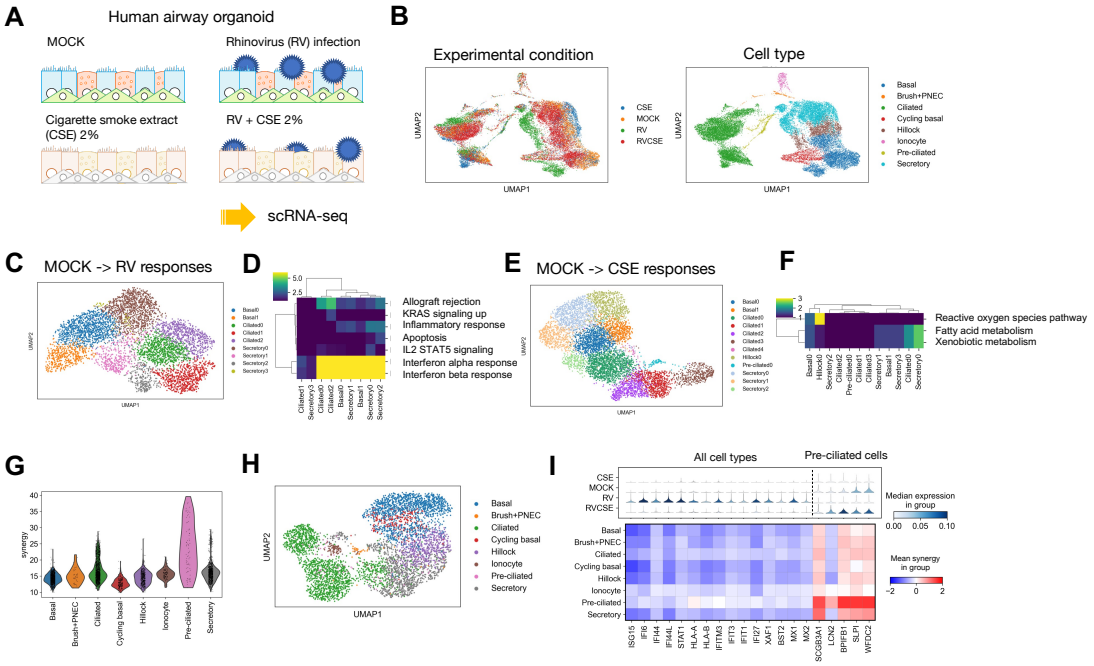

Figure 4: CINEMA-OT identifies a heterogeneous defensive response in human airway epithelial cells to rhinovirus (RV) and cigarette smoke extract (CSE). **A.** Overview of experimental design. Differentiated airway epithelial organoids are challenged with mock (control) or RV 1A infection, with or without CSE exposure. **B.** UMAP projection of expression data labeled by perturbations and cell types. **C.** UMAP projection of the individual treatment effect matrix obtained by CINEMA-OT from the RV response without CSE exposure, colored by response clusters. **D.** Gene set enrichment analysis of MOCK to RV response clusters identified by CINEMA-OT. **E.** UMAP projection of the individual treatment effect matrix obtained by CINEMA-OT from the CSE response without RV exposure, colored by response clusters. **F.** Gene set enrichment analysis of MOCK to CSE response clusters identified by CINEMA-OT. **G.** Cell-wise synergy score visualization. **H.** UMAP visualization of CINEMA-OT synergy embedding, colored by cell types. **I.** Two diverging patterns among CINEMA-OT identified significantly synergistic genes, visualized by stacked violin plots and dotplots. The first pattern mainly consists of interferon stimulated genes globally attenuated in the RVCSE condition compared with the RV condition, leading to overall negative synergy. Genes in the second patterns are activated in pre-ciliated cells from MOCK and RVCSE conditions, leading to significant positive synergy.

2.7 CINEMA-OT reveals principles of innate immune response modulation from combinatorial interferon stimulation

Type I, type II, and type III interferons (IFNs) act as central regulators of immune responses during intracellular pathogen infection, cancer, and in auto-immunity. However, despite the identification and adoption within the literature of a core set of interferon-stimulated genes (ISGs), IFN responses can vary widely by cell type, by individual, by IFN stimulus type, by chronicity of exposure, and by combination with signals delivered by other cytokines. In other words, the interferon response is highly context dependent. This complexity, heterogeneity, and context-specificity of IFN signaling can lead to counter-intuitive results. For example, IFNγ is proposed to play both stimulatory and suppressive roles in cancer, and type I IFNs are used both as an immunosuppressant to treat multiple sclerosis and as immunostimulatory adjuvant treatments for cancer (e.g. melanoma) and chronic viral infection (e.g. HCV) [45–47]. To characterize the complexity of IFN signaling, we conducted acute (2 days) and chronic (7 days) stimulations of peripheral blood mononuclear cells (PBMCs) from multiple healthy donors with type I, type II, and type III IFNs, separately as well as in combination with other cytokines such as TNFα and IL-6 (Figure 5A-B).

In order to understand the underlying structure of PBMC cellular response to interferon stimulation, we use CINEMA-OT to match treatment conditions to the untreated (control) condition. This analysis highlights the underlying hierarchical structure of cellular responses. As the hierarchical structure of cytokine response can vary with cell types, we pool $CD4^+$ T cells and monocytes across experimental batches and run CINEMA-OT analysis. In this case, a confounder is defined by each different experimental batch. In $CD4^+$ T cells, the response can be characterized by four meta-perturbation clusters: No stimulation, IFNγ, IL-6, TNFα; IFNα2, IFNβ, IFNβ+TNFα; IFNβ+IL-6; IFNβ+IFNγ (Figure 5C). In monocytes, a similar structure is observed, except that IFNγ in monocytes represents a distinct response cluster in the phenotypic space (Figure 5D).

Next, to demonstrate CINEMA-OT's power in general single-cell level treatment analysis, we focus on analyzing the treatment effects of IFNβ in a single experimental batch (H3D2). CINEMA-OT analysis highlights the induction of coordinated immune responses across cell types as well as cell-type specific responses using a confounder-specific effect volcano plot. For example, despite a global change in interferon stimulated genes (Supplementary Figure 9), monocytes demonstrate a unique program characterized by increased APOBEC3A and IL10 and decreased A2M and CCL24 compared to other cell types (Figure 5E-G, Supplementary Figure 9). CINEMA-OT is also applied to understand the treatment effects in $CD4^+$ T cells of chronic versus acute stimulation and reveals the attenuation of genes involved in the core type I IFN response (Supplementary Figure 10).

To estimate the synergistic effects of acute combinatorial cytokine stimulation, we used CINEMA-OT to calculate cell-gene synergy scores. We next performed gene synergy score analysis by computing the gene-wise synergy score (See Methods). The gene synergy score analysis identified genes that were synergistically induced by each combinatorial perturbation (Figure 5H). Based on selected significant synergy genes, we are able to summarize the cell-wise synergy effect by taking the norm over selected synergy genes. We have found that monocytes exhibit the most significant synergistic regulation compared with other cell types (Figure 5I). Further enrichment analysis identifies a number of chemokines with specific synergistic expression in monocytes with respect to different interferon perturbation (Figure 5J, Supplementary Figure 11). These chemokines exhibit synergistic patterns of response to multi-signal inflammation (e.g. IFNβ + IL-6) in monocytes including inhibition of baseline neutrophil chemotactic signaling and induction of a monocyte chemotactic signaling program. The addition of IFNγ contributes lymphocyte predominant chemokines, while maintaining this core inflammatory programming (Figure 5J, 5K). These results suggest that CINEMA-OT, when applied to combinatorial experiments, is capable of revealing the synergistic logic governing cellular signaling in inflammation and tissue repair.

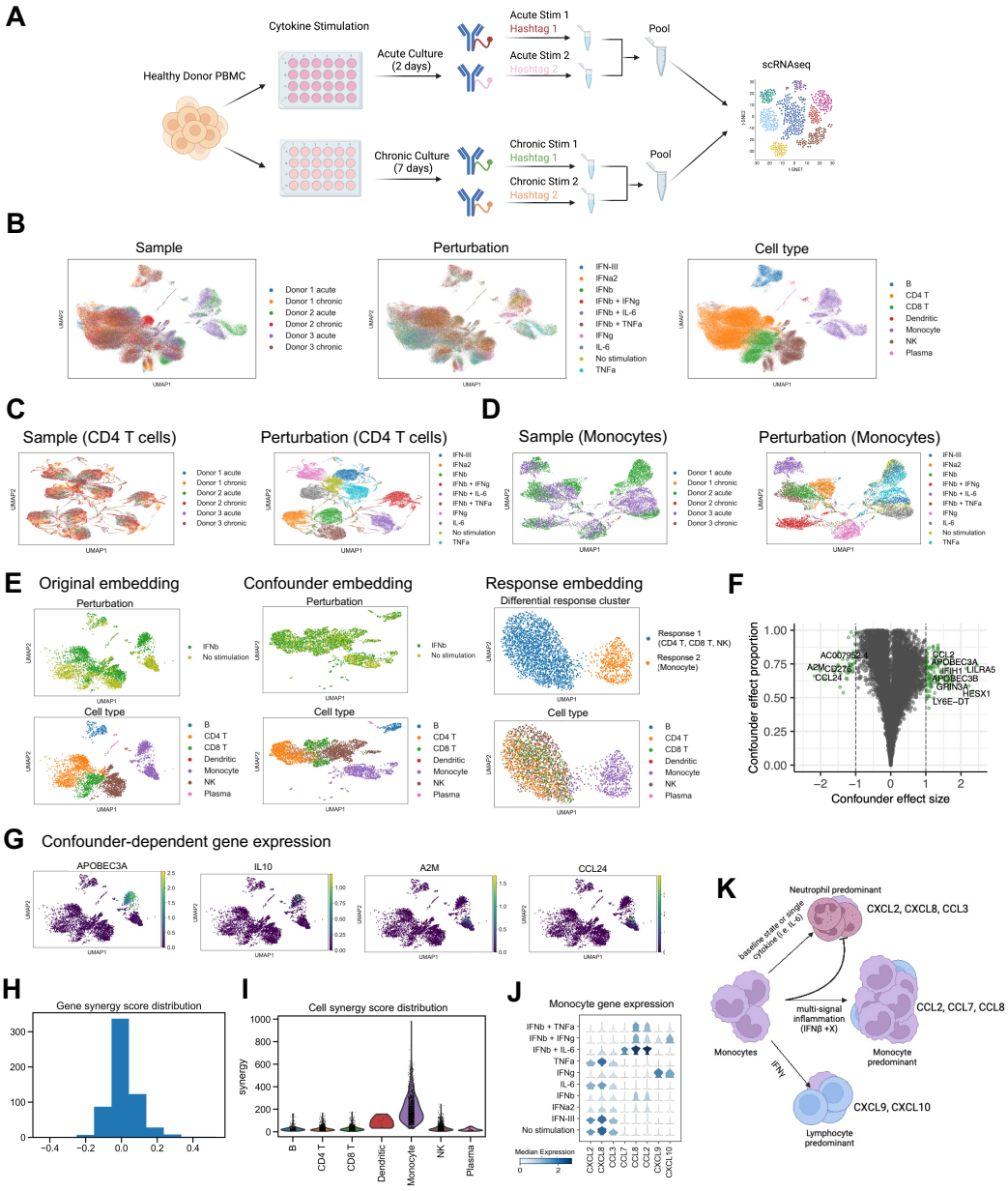

Figure 5: CINEMA-OT reveals combinatorial mechanisms of acute and chronic cytokine stimulation. **A.** Illustration of experimental design. **B.** UMAP projection of expression data colored by samples, perturbations, and cell types. In sample labels, H refers to the donor number, and D refers to the number of days of stimulation. **C.** UMAP projection of the CD4$^+$ T cell counterfactual space from CINEMA-OT. Projections are colored by experimental batch and perturbation type. **D.** UMAP projection of the monocyte counterfactual space from CINEMA-OT. Projections are colored by experimental batch and perturbation type. **E.** UMAP projections of original data, confounder embedding, and individual treatment effects identified by CINEMA-OT respectively after acute stimulation with IFNβ in H3D2, colored by response cluster and cell type. **F.** Volcano plot highlighting genes with significant confounder-specific treatment effect. **G.** Representative confounder-specific treatment-associated gene expression in original UMAP space. **H.** Gene synergy score distribution combining IFNβ + TNFα, IFNβ + IFNγ, and IFNβ + IL-6 combinatorial treatment for in H3D2. **I.** Cell-wise synergy score visualization in acute condition, taking a single experimental batch (H3D2) as an example. **J.** Stacked violin plot of significantly synergistic chemokines identified by CINEMA-OT. **K.** Patterns of chemokine secretion programmed by single or multi-signal cytokine stimulation.

3 Discussion

With fast developing high-throughput screening technologies and rising numbers of datasets, single-cell level experimental effect analysis is becoming a critically important task in biological discovery. Current analytical approaches aiming to address this need to face a number of challenges. When treatment effects are confounder specific and do not change relative cell frequencies, differential abundance methods may be unsuitable for extracting the dependence between confounder states and treatment responses. Recent neural network-based methods for characterizing perturbation effects learn nonlinear interactions between confounders and treatment effects, but may be prone to overfitting and have limited interpretability. Moreover, none of the existing approaches employ a causal inference framework for perturbation analysis.

In response to these challenges, we present CINEMA-OT, a framework for single-cell causal treatment effect analysis. By explicitly separating confounder and treatment signals and matching at the single-cell level, CINEMA-OT produces a per-cell view into the effects of experimental perturbations and conditions including disease states.

We applied CINEMA-OT in several use cases including synthetic and real data sets. In benchmarks, CINEMA-OT was able to outperform other methods in experimental perturbation analysis. In human airway organoids, CINEMA-OT revealed how cigarette smoke exposure can interfere with the normal innate immune response to rhinovirus infection. In combinatorial cytokine stimulation of *ex vivo* human peripheral immune cells, CINEMA-OT revealed complex logic which may underlie the specific recruitment of cells from the periphery to tissues responding to various injuries.

Two potential challenges for CINEMA-OT can arise due to bias-variance tradeoffs in optimal transport and the choice of dimensionality for CINEMA-OT's internal representation. For the first challenge, a large smoothness threshold in the entropy regularized method can overly smooth the obtained matching and cause false positives by incorrectly identifying confounder variation as treatment-associated variation. However, too small a threshold would both harm the method's stability and cause high variance. In practice, an adequate threshold can be chosen based on coarse-graining of the matching matrix. For the second challenge, we have applied a rank estimation procedure to quantify the intrinsic dimensionality of the count matrix [48]. For larger datasets, this procedure can be computationally expensive and we have found that a fixed rank of 20 generally yields good performance.

CINEMA-OT is a tool designed for the estimation of gene signatures causally associated with treatment responses. While we have implemented an iterative reweighting procedure to account for differential confounder abundance that may arise in response to treatment, CINEMA-OT is not designed for cases where large changes to confounder distributions are the primary effects of interest, such as proliferation or cell death. In those cases, tools such as MELD, MILO, or DA-seq may be more suitable [17, 14, 15].

Methods

CINEMA-OT

CINEMA-OT is an unsupervised method for separating confounding signals from perturbation signals for matching cells via imputing counterfactuals and computing perturbation effect at a single-cell level (`https://github.com/vandijklab/CINEMA-OT`). The detailed workflow of CINEMA-OT is as follows.

1. **Causal framework and formal proof**

   Here we give an rigorous treatment of the causal framework and underlying assumptions in CINEMA-OT.

   Assumption 1: Independent sources and noise. *Confounding factors and treatment events are pairwise independent random variables.*

   Assumption 2: Linearity of source signal combinations. *Confounding gene signatures can be modeled as a linear combination of confounding sources plus an independent noise term. The measured gene signatures can be modeled as arbitrary functions of confounding factors and treatment events plus an independent noise term.*

If we further assume the effects of confounders and treatments on perturbation-associated gene signatures are additive, then we may consider the problem of separating confounder gene signatures from treatment-associated gene signatures as equivalent to solving the blind source separation problem. However, we can no longer reveal the potential interactions between confounders and treatments in this case. Here we take a different strategy by proving the identifiability of confounding factors in our framework.

Assume the gene count matrix is a gene by cell matrix X, the corresponding principal component matrix is $\hat{X} \in \mathbb{R}^{m \times n}$, the sources are $S = \{s_i\}_{i=1,...,l}$, each source is a vector with size the number of cells and $F = \{f_i\}_{i=1,...,m}$ is a set of arbitrary functions. Then the data generation mechanism is given by

$$\hat{X} = F(S, z) = \begin{bmatrix} f_1(s_1, ..., s_l, z) \\ \vdots \\ f_m(s_1, ..., s_l, z) \end{bmatrix}$$

where z is the treatment indicator.

With Assumption 2, $\hat{X}$ can be defined as a combination of confounding factors and perturbation-associated factors up to arbitrary invertible linear transformations A and B, and where $g_1, ..., g_{m-l}$ are arbitrary functions:

$$\hat{X} = B \begin{bmatrix} AS \\ g_1(s_1, ..., s_l, z) \\ \vdots \\ g_{m-l}(s_1, ..., s_l, z) \end{bmatrix}$$

Then upon any matrix factorization, applying a functional dependence metric, we are able to distinguish if signals are dependent on the treatment variable z. Therefore we have

$$\hat{X} = C\hat{S},$$

where $\hat{f}_1, ...\hat{f}_k$ and $\hat{g}_1, ..., \hat{g}_k$ are defined by:

$$\hat{S} = \begin{bmatrix} \hat{f}_1(s_1, ..., s_l) \\ \vdots \\ \hat{f}_k(s_1, ..., s_l) \\ \hat{g}_1(s_1, ..., s_l, z) \\ \vdots \\ \hat{g}_{m-k}(s_1, ..., s_l, z) \end{bmatrix} = C^{-1}B \begin{bmatrix} AS \\ g_1(s_1, ..., s_l, z) \\ \vdots \\ g_{m-l}(s_1, ..., s_l, z) \end{bmatrix}.$$

We define $\hat{A} = B^{-1}C$, in which case we have

$$S = A^{-1}(\hat{A}\hat{S})_{1:l} = A^{-1}(\sum_j \hat{A}_{ij}\hat{S}_j)_{i=1:l}.$$

Because $\hat{g}$s are pairwise independent functions of z, $\hat{A}_{ij}$ has to be zero for $j > k$, otherwise $A^{-1}(\sum_j \hat{A}_{ij}\hat{S}_j)_{i=1:n}$ would be a function of z, leading to a contradiction.

Therefore we have

$$S = A^{-1}(\sum_{j=1}^{k} \hat{A}_{ij}\hat{S}_j)_{i=1:l} = A^{-1}(\sum_{j=1}^{k} \hat{A}_{ij}\hat{f}_j)_{i=1:l} = A^{-1}\hat{A}^*\hat{f}.$$

Here $\hat{A}^*$ is a $l \times k$ matrix. Therefore because S is of rank l, we must have $k \geq l$, and $\hat{f}$ is a linear transformation of S.

If $k > l$, then because $\hat{f}$ defines a linear transformation on S, $k > l$ indicates $\hat{f}$ is not of full rank. Therefore matrix $\hat{S}$ is not of full rank, contradicting with its a invertible transformation of a full-rank signal matrix. The contradiction leads to the only possibility of $k = l$.

With $k = n$, because A and $\hat{A}^*$ (the sub-matrix of the full-rank matrix $\hat{A}$) are both invertible, according to ICA identifiability theorem, we have

$$S = \hat{S}_{1:l} = (C^{-1}\hat{X})_{1:l}$$

up to a permutation.

In summary, we have proved that upon correctly identifying $\hat{f}$s in our framework, we can identify all underlying confounders up to a permutation, even if there are additional nonlinear treatment-associated signals. The key difficulty we have overcomed is $g_i(s_1, s_2, ..., s_l, z)$ are generally dependent across different is, therefore the ICA identifiability theorem cannot be applied. Note that we cannot guarantee the recovery of treatment-associated signals here, which is why after identifying the confounders, we applied a non-parametric causal matching approach to fully reveal underlying causal processes in the data.

We cannot guarantee theoretically that our threshold procedure selects correct confounders and treatment-associated signals. Accounting for the possible uncertainty, it can still be seen that the matching matrix constructed by CINEMA-OT smoothly interpolates between the matching matrix of Mixscape and null matching matrix used in single-cell differential expression analysis.

Finally, given the confounders, the problem of treatment effect estimation can be solved by standard potential outcome framework. The framework has mainly four assumptions, which are discussed in detail in causal inference textbooks [22]:

1. Stable Unit Treatment Assumption (SUTVA): Samples are independent without interference; 2. Ignorability: there are no unmeasured confounders; 3. Consistency; 4. Positivity: for a given confounder, the probability of perturbation is neither 0 or 1.

2. **Rank initialization**

   In order to perform CINEMA-OT, we first need to initialize the expected matrix rank, representing the total signal number. We here offer two possible approaches for rank initialization in CINEMA-OT.

   Biwhitening [48] is a recently-developed method to remove independent heteroskedastic noise in data with inspirations from random matrix theory. It does diagonal matrix transformation of the data on both sides and thresholding based on the Marchenko-Pastur law [49]. After thresholding, we can get the true matrix rank and the matrix's low dimensional approximation. Mathematical details of biwhitening can be seen in [48]. In CINEMA-OT, we have implemented a version of biwhitening with fixed hyperparameters.

   In large datasets, we suggest using prespecified rank values. Empirically, we have found that CINEMA-OT is robust to rank selection at certain ranges and can give a good performance when $\mathrm{DimSize} = 20$.

3. **Signal selection with independent component analysis**

   Independent component analysis is already a well-addressed method in data analysis and has various implementations. Here we use the FastICA implementation from the package `sklearn.decomposition` [50]. Prior to FastICA, input data is PCA-transformed using scanpy [51].

   In order to identify confounder signals and treatment-associated signals, we have adopted a recently proposed cross rank coefficient [28], which is able to quantify the functional dependence between ICA signals and query signals (in this case, the treatment signals). We use the implementation of this method from the XICOR package in R. The threshold of the cross rank coefficient is set to be 0.5 to 0.75 in this study. We note that tuning the threshold parameter has a practical meaning in the algorithm. High thresholds correspond to less tolerance for false positive treatment signals, which leads to local matching more similar to Mixscape analyses. Meanwhile, setting a low threshold means less tolerance for false positive confounder signals and can lead to lower resolution of matching, which, in the extreme case, coincides with single-cell differential expression testing methods.

4. **Optimal transport matching**

   After selecting confounding signals, we perform matching across treatments via optimal transport, which provides a smooth transport map and does not require neighbor number selection. Here we consider the entropy regularized optimal transport formulation, which

can be efficiently solved by the Sinkhorn-Knopp algorithm [29]. In this formulation of the problem, the penalty coefficient act as a hyper parameter influencing the resolution and smoothness of the transport map. We have empirically determined that the optimal value for the penalty coefficient often lies within the range $(10^{-6}, 10^{-3})$ multiplied by the number of confounding signals.

Algorithm 1 CINEMA-OT

Require: Count matrix $X_0 \in R^{m \times n}$, treatment vector $z \in \{0,1\}^n$, dimension size r, signal filtering threshold d, smoothness s.

$\mathrm{DimSize} \leftarrow r, \mathrm{Thres} \leftarrow d, X \leftarrow X_0$.
unmixing matrix B, source matrix $S \leftarrow \mathrm{ICA}(X, \mathrm{DimSize})$;
$c \leftarrow \mathrm{zeros}(\mathrm{DimSize})$
for $i = 1 : \mathrm{DimSize}$ **do**
    $c_i \leftarrow \mathrm{xicor}(S[i,:], z)$; ▷ Compute Chatterjee cross rank coefficient
end for
$S^c \leftarrow S[c < \mathrm{Thres}, :]$ ▷ Thresholding to separate confounder signals S^c
$M \leftarrow \mathrm{OT}(S^c[:, z=0], S^c[:, z=1], \mathrm{smoothness} = s * S^c.\mathrm{shape}[0])$ ▷ M: Matching matrix
$D \leftarrow X_0[:, z=1]M - X_0[:, z=0]$ ▷ ITE matrix computation
Downstream analysis.

Algorithm 2 OT

Require: Confounder signals S_1, S_2, weights $w_1 = \mathrm{None}, w_2 = \mathrm{None}$, smoothness s.

if w_1 is None **then**
    $r \leftarrow 1/S_1.\mathrm{shape}[0], c \leftarrow 1/S_2.\mathrm{shape}[0]$
else
    $r \leftarrow w_1/w_1.\mathrm{shape}, c \leftarrow w_2/w_2.\mathrm{shape}$
end if
$D \leftarrow \mathrm{PairwiseDistance}(S_1, S_2)$.
$A \leftarrow \exp(-D * D/s)$ ▷ Elementwise multiplication for D here
$M = \mathrm{SinkhornKnopp}(A, \mathrm{setr} = r, \mathrm{setc} = c)$ ▷ Sinkhorn-Knopp algorithm
return M

Iterative reweighting in CINEMA-OT

If confounder signals are not independent of treatment indicators, as in the case of differential abundance, confounder signals and treatment-associated signals may not be completely unmixed by ICA. Even if we select the confounder signals by thresholding as previously described, there may still remain treatment signal within the confounder signals. We have implemented a heuristic approach to enable better separation of these signals through iterative application of ICA and reweighting. If the confounders are independent of treatment labels, then the local abundance of cells in the confounder space should be balanced. If the cells are not balanced in the confounder space, then we attempt to impose balance on the space by reweighting the cells.

In our implementation, we use MELD [17] to estimate each cell's treatment label likelihood. The likelihood is then used for estimation of cell-wise weights.

We note, however, that imbalance in the confounder space may also occur when the threshold parameter d is set too high, causing misidentification of treatment-associated factors as confounding factors. CINEMA-OT has no way of distinguishing between these two scenarios, but our benchmarking suggests that the iterative reweighting procedure assigns correct weights in practice, achieving good separation of confounding and treatment-associated variation. In addition, CINEMA-OT offers users the ability to specify known confounder labels (e.g. cell type, cell cycle), which may be directly used for balancing as an alternative to MELD, without the need for an iterative procedure.

Downstream analysis

1. **Visualization and clustering of the ITE matrix**

Algorithm 3 Iterative reweighting CINEMA-OT

Require: Count matrix $X_0 \in R^{m \times n}$, treatment vector $z \in \{0,1\}^n$, dimension size r, signal filtering threshold d, smoothness s.

$\text{DimSize} \leftarrow r, \text{Thres} \leftarrow d, X \leftarrow X_0$.
while not converge **do**
  $\text{unmixing matrix } B, \text{source matrix } S \leftarrow \text{ICA}(X, \text{DimSize})$;
  $c \leftarrow \text{zeros}(\text{DimSize})$
  for $i = 1 : \text{DimSize}$ **do**
    $c_i \leftarrow \text{xicor}(S[i,:], z)$; ▷ Compute Chatterjee cross rank coefficient
  end for
  $S^c \leftarrow S[c < \text{Thres}, :]$ ▷ Thresholding to separate confounder signals S^c
  $p \leftarrow \text{MELD}(\text{data} = S^c, \text{SampleLabel} = z)$.
  $w[z=0] \leftarrow p[z=0,1], w[z=1] \leftarrow p[z=1,0]$. ▷ Compute propensity score
  $X \leftarrow \text{sample}(\text{size} = 2n, \text{data} = X, \text{Weight} = w)$
end while
$\hat{S} \leftarrow (BX_0)^T$
$\hat{S}^c \leftarrow \hat{S}[c < \text{Thres}, :]$
$M \leftarrow \text{OT}(\hat{S}^c[:, z=0], \hat{S}^c[:, z=1], w_0 = w[z=0], w_1 = w[z=1], \text{smoothness} = s)$ ▷ M: Matching matrix
$D \leftarrow X_0[:, z=1]M - X_0[:, z=0]$ ▷ ITE matrix computation
Downstream analysis.

With the ITE matrix computed by matching counterfactuals, we are able to numerous standard analyses. We may employ dimensionality reduction techniques such as t-SNE, UMAP, or PHATE [52–54] to visualize clusters in the response space. We may also employ clustering techniques, such as Leiden clustering [23] to group cells by similarity of treatment responses.

2. **Synergy analysis**

For synergy effect, we compare ITE matrices for two treatment conditions against the ITE matrix for the combined treatment. Formal derivation of the synergy score is given as follows.

Consider $D_{A=1,B=0}$ as the ITE matrix for treatment A alone, $D_{A=0,B=1}$ as the ITE matrix for treatment B alone, and $D_{A=1,B=1}$ as the ITE matrix for the combined treatment. We may define a synergy matrix Ψ as:

$$\Psi = D_{A=1,B=1} - (D_{A=1,B=0} + D_{A=0,B=1})$$

Where each entry $\Psi_{g,c}$ represents the synergy score for gene g and cell c. In order to test if a particular gene g has significant synergistic effect, we formulate the problem as if we should reject

$$H_0 : E(\Psi_{g,c}) = 0, \ \forall c.$$

Note here if we apply no normalization, we are aiming for additive synergy; if we instead apply log normalized data, H_0 would test for multiplicative synergy.

We assume that different cells are unlikely to have opposite synergy effects, allowing us to relax H_0 as:

$$H_0 : E(\overline{\Psi}_{g,:}) = 0.$$

Assume the new H_0 holds, then for each gene g, we compute the empirical synergy:

$$\overline{\Psi}_{g,:} = E(\overline{\Psi}_{g,:}) + \epsilon_1 + \epsilon_2 - \epsilon_3 - \epsilon_4.$$

Because here H_0 holds, the expectation of the noise term is zero.

Assume the noise is Poisson, then with the property of Poisson distribution, in the case of log normalization, ϵ_is are averages of i.i.d. scaled log1p Poisson distribution with zero expectation. With the delta method, the variance of the noise term is approximated as:

$$\mathrm{Var}(\epsilon_1+\epsilon_2-\epsilon_3-\epsilon_4)=\frac{1}{n^2}\sum_{c=1}^{n}(\frac{\lambda_{1c}}{(1+\lambda_{1c})^2}+\frac{\lambda_{2c}}{(1+\lambda_{2c})^2}+\frac{\lambda_{3c}}{(1+\lambda_{3c})^2}+\frac{\lambda_{4c}}{(1+\lambda_{4c})^2}).$$

Where $\lambda_{ic}, i \in \{1,2,3,4\}$s are counterfactual cell gene expression expectation for each cell in 4 conditions.

We note the formula $\frac{\lambda}{(1+\lambda)^2}$ is self-standardized as it is a smooth function and is near zero for either large or small λs. Therefore, to simplify the statistical test, we assume

$$\mathrm{Var}(\epsilon_1+\epsilon_2-\epsilon_3-\epsilon_4)=\mathrm{const}.$$

In this case, if we define

$$\text{Synergy score}=|\overline{\Psi}_{g,:}|,$$

identifying most synergistic genes among all genes can be turned into comparing the synergy score over all genes.

3. **GSEA analysis**

   For differential gene expression significance, we have applied the non-parametric Wilcoxon signed-rank test. We apply a p-value threshold (10^{-5}) and expression fold change threshold selected by the user to identify significantly regulated genes. These genes are input into GSEApy for analysis by functional signatures [55, 56].

4. **Attribution analysis**

   By clustering cells both by treatment responses (i.e. using the ITE matrix) and control condition clusters (i.e. cell subtypes), the matching matrix from CINEMA-OT can be coarse grained. The resulting course-grained matching matrix is of shape $\mathrm{ResponseClusterNumber} \times \mathrm{ControlClusterNumber}$. Each column of the matrix gives the likelihood of a control condition cluster to have different modes of response. By reading each row of the matrix, we are able to attribute each response to the underlying control condition cluster.

   Furthermore, to investigate the genes with significant confounder-specific treatment effect, we fit each gene to the causal regression model

   $$X=\alpha z+\beta c+\gamma cz+const$$

   In this case, the confounder-specific effect part is γcz, whose significance can be established by classical linear regression theory. However, in our case, the noise term can stand for latent-factor specific effect thus not satisfying the assumption of classical regression. Therefore, here we instead quantify the ratio of norm between confounder-specific effect and the residual as an indicator of confounder-specific effect significance.

Data simulation and analysis

For the simulation in main figure (Figure 2B-C), we use Scsim to simulate 1000 gene by 5000 cell count matrices with 2-5 underlying cell states with 2 gene regulation programs. For each cell state, we simulate a random discrete distribution to represent the corresponding response distribution of the cell state. Then the response count matrix of shape 500 gene times 5000 cells are simulated and concatenated with the confounder count matrix.

For case I, II, III in Supplementary Figure 2, we use Splatter to simulate 500 gene by 1000 cell count matrices with 2-4 underlying cell states. Then we simulate two orthogonal trajectories, each with 250 genes. After creating confounder genes, we next simulate 500 outcome genes from two groups, enabled by the `group` function in Splatter. Each cell has an equal probability of being assigned into outcome gene cluster 1 (group 1) and cluster 2 (group 2). In the first two scenarios, The confounders are simulated to have random linear effects on the outcome genes, represent potential mixing of confounders and outcomes in the same gene, with or without differential abundance. In the third scenario, the confounders are simulated to have both linear mixing effects and state-specific effects on treatment-associated genes. We simulate 15 times with different random seeds to generate 15

gene count matrices with size 1000×1500 (cells by genes). Then the data is normalized, log transformed, and analyzed by our implemented methods respectively.

For case IV in Supplementary Figure 2, we use Scsim to simulate 500 gene by 1000 cell count matrices with 2-5 underlying cell states with 2 gene regulation programs. In this scenario, the heterogeneity of treatment effects is generated by random sampling independent of the confounders. We simulate 15 times with different random seeds to generate 15 gene count matrices with size 1000×1500 (cells by genes). Then the data is normalized, log transformed, and analyzed by our implemented methods respectively.

For the Mixscape analysis, we have implemented a simple version in Python that matches cells across conditions according to the descriptions in [18]. For Harmony-Mixscape analysis, we have used the Python package `harmonypy` (https://github.com/slowkow/harmonypy) with default settings [57]. For OT analysis, we implemented a function that calls entropy-regularized optimal transport with a tunable smoothness parameter. For ContrastiveVI, we use the default setting (provided in its tutorial: https://colab.research.google.com/drive/1_R1YWQQUJzgQ6kz1XqglL5xZn8b8h1TX?usp=sharing) with 10 dimensions of shared embedding and salient embedding respectively. For CINEMA-OT, we run the analysis with default settings and smoothness is set to be 5×10^{-4}.

Benchmarking metrics

ITE Pearson, ITE Spearman

We evaluate the Pearson / Spearman correlation between output ITE matrices and the ground truth treatment effects (derived using the ground truth matching).

ASW, PCR, Graph connectivity, ARI

These metrics are batch mixing and biological preservation metrics used to evaluate batch correction methods performance in the systematic benchmarking paper [37]. CINEMA-OT uses the first three metrics to evaluate mixing in confounder space, as a surrogate for correct matching that can still be measured when ground truth labels are not present. ARI is used to evaluate the preservation of response clusters in ITE matrices estimated from simulated data. We use the implementations of these metrics from package scib [37].

Diffusion-map dependence coefficient

In order to evaluate preservation of underlying confounders, we use diffusion-map dependence coefficients. We calculate these coefficients for both cell state and cell trajectory. As our simulated data do not form a well-defined trajectory, and multiple orthogonal trajectories may be simultaneously present in the data, we approach trajectory preservation differently from [37].

We evaluate the maximum possible functional dependence coefficient, a rank-based measure, between true order and diffusion map eigenvectors. Low values of the index would indicate there are no components that are of high functional dependence with respect to the trajectory, therefore the information of the trajectory is not well preserved.

The Laplacian eigenvectors rely on kernel function selection. In the case of a covariance kernel, the method reduces to computation of the functional dependence between ground truth and confounding covariates identified by each method. In our case, we select the default setting provided by the `scanpy.tl.diffmap` implementation.

For discrete labels, we use one-hot encoding to create n vectors where n equals the number of labels. Then for each vector, we seek the maximum functional dependence score as described above, and then take average over all n coefficients to get an average score as the final output.

Alzheimers scRNA-seq data

The Alzheimer scRNA-seq data is downloaded from https://drive.google.com/uc?id=1R1aN-LWUQ6c_N44n5-xjy2nEPzl7H0Dc. For mixscape and CINEMA-OT, the data is normalized and log transformed. 2000 highly variable genes are selected with Seurat v3 approach implemented in

Scanpy. As Contrastive-VI models the distribution of count matrix, the original count matrix of highly variable genes is used for Contrastive-VI.

Sci-Plex4 data

The Sci-Plex4 data was accessed from https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM4150379 with GEO accession number GSM4150379. The data is preprocessed via protocol https://github.com/manuyavuz/single-cell-analysis/blob/main/single_cell_analysis/datasets/sciplex.py. After preprocessing, we normalized and log transformed the raw count matrix and perform subsequent analysis described in main text.

After estimating all metrics, each metric is rescaled so that the sum across all methods tested equals 1. Then we sum the rescaled score over metrics and cell types to calculate the combined score. The statistical test is performed by Wilcoxon signed-rank test `scipy.stats.wilcoxon` over the rescaled metrics.

Rhinovirus infection data

Primary human bronchial epithelial cells from healthy adult donors were obtained from commercial vendor (Lonza) and cultured at air-liquid interface according to the manufacturers instructions (Stem Cell Technologies) using reduced hydrocortisone. Cells were kept at air-liquid interface for 4 weeks before experiment; maturation of beating cilia and mucus production was confirmed using light microscope. Cells were then infected with mock or 10^5 PFU human rhinovirus 1A per organoid, with or without exposure to 2% cigarette smoke extract (CSE). Single cell suspension is collected by trypsin digestion at 5 days post infection and submitted to single cell RNA sequencing using The 10X Genomics single-cell $3'$ protocol. The final dataset contains 26420 cells in 4 samples (mock, RV, CSE, RVCSE). We performed normalization (by `sc.pp.normalize_total`), log1p transformation, and selected 3608 highly variable genes, scaled their values for PCA and Leiden clustering analysis. We annotated 8 cell clusters based on known cell type markers of airway epithelial cells [58]: cycling basal, basal, hillock, secretory, pre-ciliated, ciliated, ionocyte, and PNEC/brush cells. CINEMA-OT analysis on mock-RV and mock-CSE was run with default parameters with `smoothness=1e-5`. Synergy analysis was performed with `smoothness=3e-5`.

Interferon treatment data

PBMC processing and in vitro culture

The study was approved by Institutional Review Boards at Yale University (following Yale melanoma skin SPORE IRB protocol). Healthy donors consented to donation of peripheral blood for research use.

Human PBMC were isolated using Lymphoprep density gradient medium (STEMCELL). PBMC were plated at 1 million cells per ml and stimulated with 1000U/ml human IFNα2 (R&D systems), 1000U/ml human IFNβ (pbl assay science 11415), 1000U/ml human IFNγ (pbl assay science), 1ug/ml human IFN-III /IL-29 (R&D systems), 100ng/ml human IL-6 (NCI Biological Resources Branch Preclinical Biologics Repository), 20ng/ml human TNFα (R&D systems), and combinatorial cytokines IFNβ + IL-6, IFNβ + TNFα, IFNβ+ IFNγ at indicated concentrations above for up to 48 hours.

Cell enrichment and 10x sample preparation

Cultured cells were collected stained with TotalSeq anti-human hashtags C0251-C0260 (Biolegend), viability dye (zombie red, Biolegend) and anti-human CD45-FITC (clone HI30, Biolegend) and enriched for live CD45$^+$ cells using BD FACS Aria II. Sorted cells were then resuspended to 1200 cells per ul and barcoded for multiplexed single cell sequencing using 10x Genomics 5v2 chemistry (10x Genomics, PN-1000263).

Sequencing and 10x sample alignment

Single cell RNA sequencing libraries were sequenced on Illumina NovaSeq at read length of 150bp pair-end and depth of 300 million reads per sample.

scRNA-seq data analysis

Data from three donors across Day 2 and Day 7 are concatenated together into labeled anndata objects for analysis. For each of the 6 samples, we filtered cells with less than 200 genes and we filtered genes expressed in fewer than 3 cells. For further quality control, cells with a high proportion of mitochondiral reads (> 7%) were excluded. The distribution of genes per cell was visually inspected and upper thresholds selected on a per-sample basis to exclude doublets. For each of the samples, the upper threshold was selected as [6000,3500,4000,3500,4500,3500] respectively. Following filtering, the count data was normalized and log transformed. Highly variable gene selection was performed by `sc.pp.highly_variable_genes(adata, min_mean=0.0125, max_mean=3, min_disp=0.5)`. Highly variable genes were used for subsequent PCA and UMAP projection.

For individual treatment effect analysis, we additionally filter T cell receptor genes, histocompatibility genes, and immunoglobulin genes from the highly variable gene set. Genes to be filtered were obtained from the HUGO database [59]. After filtering, highly variable genes were selected for downstream visualization analysis.

CINEMA-OT analysis was run on each of the samples separately, with `thres=0.5`, `smoothness=1e-4`, `eps=1e-2`, and preweights given by cell types.

For the synergy analysis of donor 3 on day 2 (H3D2), we selected significant synergy genes by a absolute value threshold of 0.15.

References

[1] Maya E Kotas and Ruslan Medzhitov. Homeostasis, inflammation, and disease susceptibility. *Cell*, 160(5):816–827, 2015.

[2] Asif Adil, Vijay Kumar, Arif Tasleem Jan, and Mohammed Asger. Single-cell transcriptomics: current methods and challenges in data acquisition and analysis. *Frontiers in Neuroscience*, 15: 591122, 2021.

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

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

[5] Lin Yang, Yuqing Zhu, Hua Yu, Xiaolong Cheng, Sitong Chen, Yulan Chu, He Huang, Jin Zhang, and Wei Li. scmageck links genotypes with multiple phenotypes in single-cell crispr screens. *Genome biology*, 21(1):1–14, 2020.

[6] Bin Duan, Chi Zhou, Chengyu Zhu, Yifei Yu, Gaoyang Li, Shihua Zhang, Chao Zhang, Xiangyun Ye, Hanhui Ma, Shen Qu, et al. Model-based understanding of single-cell crispr screening. *Nature communications*, 10(1):1–11, 2019.

[7] Xin Jin, Sean K Simmons, Amy Guo, Ashwin S Shetty, Michelle Ko, Lan Nguyen, Vahbiz Jokhi, Elise Robinson, Paul Oyler, Nathan Curry, et al. In vivo perturb-seq reveals neuronal and glial abnormalities associated with autism risk genes. *Science*, 370(6520):eaaz6063, 2020.

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

[9] Michael I Love, Wolfgang Huber, and Simon Anders. Moderated estimation of fold change and dispersion for rna-seq data with deseq2. *Genome biology*, 15(12):1–21, 2014.

[10] Mark D Robinson, Davis J McCarthy, and Gordon K Smyth. edger: a bioconductor package for differential expression analysis of digital gene expression data. *bioinformatics*, 26(1):139–140, 2010.

[11] Davis J McCarthy, Yunshun Chen, and Gordon K Smyth. Differential expression analysis of multifactor rna-seq experiments with respect to biological variation. *Nucleic acids research*, 40(10):4288–4297, 2012.

[12] Yunshun Chen, Aaron TL Lun, and Gordon K Smyth. From reads to genes to pathways: differential expression analysis of rna-seq experiments using rsubread and the edger quasi-likelihood pipeline. *F1000Research*, 5, 2016.

[13] Matthew E Ritchie, Belinda Phipson, DI Wu, Yifang Hu, Charity W Law, Wei Shi, and Gordon K Smyth. limma powers differential expression analyses for rna-sequencing and microarray studies. *Nucleic acids research*, 43(7):e47–e47, 2015.

[14] Emma Dann, Neil C Henderson, Sarah A Teichmann, Michael D Morgan, and John C Marioni. Differential abundance testing on single-cell data using k-nearest neighbor graphs. *Nature Biotechnology*, 40(2):245–253, 2022.

[15] Jun Zhao, Ariel Jaffe, Henry Li, Ofir Lindenbaum, Esen Sefik, Ruaidhrí Jackson, Xiuyuan Cheng, Richard Flavell, and Yuval Kluger. Detection of differentially abundant cell subpopulations discriminates biological states in scrna-seq data. *bioRxiv*, page 711929, 2020.

[16] Maren Buettner, Johannes Ostner, Christian L Mueller, Fabian J Theis, and Benjamin Schubert. sccoda is a bayesian model for compositional single-cell data analysis. *Nature communications*, 12(1):1–10, 2021.

[17] Daniel B Burkhardt, Jay S Stanley, Alexander Tong, Ana Luisa Perdigoto, Scott A Gigante, Kevan C Herold, Guy Wolf, Antonio J Giraldez, David van Dijk, and Smita Krishnaswamy. Quantifying the effect of experimental perturbations at single-cell resolution. *Nature biotechnology*, 39(5):619–629, 2021.

[18] Efthymia Papalexi, Eleni P Mimitou, Andrew W Butler, Samantha Foster, Bernadette Bracken, William M Mauck, Hans-Hermann Wessels, Yuhan Hao, Bertrand Z Yeung, Peter Smibert, et al. Characterizing the molecular regulation of inhibitory immune checkpoints with multimodal single-cell screens. *Nature genetics*, 53(3):322–331, 2021.

[19] Mohammad Lotfollahi, Anna Klimovskaia Susmelj, Carlo De Donno, Yuge Ji, Ignacio L Ibarra, F Alexander Wolf, Nafissa Yakubova, Fabian J Theis, and David Lopez-Paz. Compositional perturbation autoencoder for single-cell response modeling. *BioRxiv*, 2021.

[20] Charlotte Bunne, Stefan G Stark, Gabriele Gut, Jacobo Sarabia del Castillo, Kjong-Van Lehmann, Lucas Pelkmans, Andreas Krause, and Gunnar Ratsch. Learning single-cell perturbation responses using neural optimal transport. *bioRxiv*, 2021.

[21] Romain Lopez, Jeffrey Regier, Michael B Cole, Michael I Jordan, and Nir Yosef. Deep generative modeling for single-cell transcriptomics. *Nature methods*, 15(12):1053–1058, 2018.

[22] Guido W Imbens and Donald B Rubin. *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press, 2015.

[23] Tyler J VanderWeele and Ilya Shpitser. On the definition of a confounder. *Annals of statistics*, 41(1):196, 2013.

[24] Sona Vodenkova, Tomas Buchler, Klara Cervena, Veronika Veskrnova, Pavel Vodicka, and Veronika Vymetalkova. 5-fluorouracil and other fluoropyrimidines in colorectal cancer: Past, present and future. *Pharmacology & therapeutics*, 206:107447, 2020.

[25] Lenka Kubiczkova, Lenka Sedlarikova, Roman Hajek, and Sabina Sevcikova. Tgf-β–an excellent servant but a bad master. *Journal of translational medicine*, 10(1):1–24, 2012.

[26] D.B. Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66(5):688–701, 1974.

[27] Chandler Squires and Caroline Uhler. Causal structure learning: a combinatorial perspective. *arXiv preprint arXiv:2206.01152*, 2022.

[28] Sourav Chatterjee. A new coefficient of correlation. *Journal of the American Statistical Association*, 116(536):2009–2022, 2021.

[29] Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. *Advances in neural information processing systems*, 26, 2013.

[30] Florian Gunsilius and Yuliang Xu. Matching for causal effects via multimarginal optimal transport. *arXiv preprint arXiv:2112.04398*, 2021.

[31] Richard Sinkhorn. A relationship between arbitrary positive matrices and doubly stochastic matrices. *The annals of mathematical statistics*, 35(2):876–879, 1964.

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

[33] Ethan Weinberger, Chris Lin, and Su-In Lee. Isolating salient variations of interest in single-cell transcriptomic data with contrastivevi. *bioRxiv*, 2021.

[34] Luke Zappia, Belinda Phipson, and Alicia Oshlack. Splatter: simulation of single-cell rna sequencing data. *Genome biology*, 18(1):1–15, 2017.

[35] Dylan Kotliar, Adrian Veres, M Aurel Nagy, Shervin Tabrizi, Eran Hodis, Douglas A Melton, and Pardis C Sabeti. Identifying gene expression programs of cell-type identity and cellular activity with single-cell rna-seq. *Elife*, 8, 2019.

[36] Alexandra Grubman, Gabriel Chew, John F Ouyang, Guizhi Sun, Xin Yi Choo, Catriona McLean, Rebecca K Simmons, Sam Buckberry, Dulce B Vargas-Landin, Daniel Poppe, et al. A single-cell atlas of entorhinal cortex from individuals with alzheimers disease reveals cell-type-specific gene expression regulation. *Nature neuroscience*, 22(12):2087–2097, 2019.

[37] Malte D Luecken, Maren Büttner, Kridsadakorn Chaichoompu, Anna Danese, Marta Interlandi, Michaela F Müller, Daniel C Strobl, Luke Zappia, Martin Dugas, Maria Colomé-Tatché, et al. Benchmarking atlas-level data integration in single-cell genomics. *Nature methods*, 19(1):41–50, 2022.

[38] Carlo Sala Frigerio, Leen Wolfs, Nicola Fattorelli, Nicola Thrupp, Iryna Voytyuk, Inga Schmidt, Renzo Mancuso, Wei-Ting Chen, Maya E Woodbury, Gyan Srivastava, et al. The major risk factors for alzheimers disease: age, sex, and genes modulate the microglia response to aβ plaques. *Cell reports*, 27(4):1293–1306, 2019.

[39] Tomas Eckschlager, Johana Plch, Marie Stiborova, and Jan Hrabeta. Histone deacetylase inhibitors as anticancer drugs. *International journal of molecular sciences*, 18(7):1414, 2017.

[40] Shijie C Zheng, Genevieve Stein-OBrien, Jonathan J Augustin, Jared Slosberg, Giovanni A Carosso, Briana Winer, Gloria Shin, Hans T Bjornsson, Loyal A Goff, and Kasper D Hansen. Universal prediction of cell-cycle position using transfer learning. *Genome biology*, 23(1): 1–27, 2022.

[41] Nagarjuna R Cheemarla, Timothy A Watkins, Valia T Mihaylova, Bao Wang, Dejian Zhao, Guilin Wang, Marie L Landry, and Ellen F Foxman. Dynamic innate immune response determines susceptibility to sars-cov-2 infection and early replication kinetics. *Journal of Experimental Medicine*, 218(8), 2021.

[42] Neal G Ravindra, Mia Madel Alfajaro, Victor Gasque, Nicholas C Huston, Han Wan, Klara Szigeti-Buck, Yuki Yasumoto, Allison M Greaney, Victoria Habet, Ryan D Chow, et al. Single-cell longitudinal analysis of sars-cov-2 infection in human airway epithelium identifies target cells, alterations in gene expression, and cell state changes. *PLoS biology*, 19(3):e3001143, 2021.

[43] Valia T Mihaylova, Yong Kong, Olga Fedorova, Lokesh Sharma, Charles S Dela Cruz, Anna Marie Pyle, Akiko Iwasaki, and Ellen F Foxman. Regional differences in airway epithelial cells reveal tradeoff between defense against oxidative stress and defense against rhinovirus. *Cell reports*, 24(11):3000–3007, 2018.

[44] Sandra Ruiz García, Marie Deprez, Kevin Lebrigand, Amélie Cavard, Agnès Paquet, Marie-Jeanne Arguel, Virginie Magnone, Marin Truchi, Ignacio Caballero, Sylvie Leroy, et al. Novel dynamics of human mucociliary differentiation revealed by single-cell rna sequencing of nasal epithelial cultures. *Development*, 146(20):dev177428, 2019.

[45] Joseph L Benci, Lexus R Johnson, Ruth Choa, Yuanming Xu, Jingya Qiu, Zilu Zhou, Bihui Xu, Darwin Ye, Katherine L Nathanson, Carl H June, et al. Opposing functions of interferon coordinate adaptive and innate immune responses to cancer immune checkpoint blockade. *Cell*, 178(4):933–948, 2019.

[46] Juan Dubrot, Peter P Du, Sarah Kate Lane-Reticker, Emily A Kessler, Audrey J Muscato, Arnav Mehta, Samuel S Freeman, Peter M Allen, Kira E Olander, Kyle M Ockerman, et al. In vivo crispr screens reveal the landscape of immune evasion pathways across cancer. *Nature Immunology*, 23(10):1495–1506, 2022.

[47] Stephen L Hauser and Bruce AC Cree. Treatment of multiple sclerosis: a review. *The American journal of medicine*, 133(12):1380–1390, 2020.

[48] Boris Landa, Thomas TCK Zhang, and Yuval Kluger. Biwhitening reveals the rank of a count matrix. *arXiv preprint arXiv:2103.13840*, 2021.

[49] Vladimir A Marčenko and Leonid Andreevich Pastur. Distribution of eigenvalues for some sets of random matrices. *Mathematics of the USSR-Sbornik*, 1(4):457, 1967.

[50] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12:2825–2830, 2011.

[51] F Alexander Wolf, Philipp Angerer, and Fabian J Theis. Scanpy: large-scale single-cell gene expression data analysis. *Genome biology*, 19(1):1–5, 2018.

[52] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

[53] Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018.

[54] Kevin R Moon, David van Dijk, Zheng Wang, Scott Gigante, Daniel B Burkhardt, William S Chen, Kristina Yim, Antonia van den Elzen, Matthew J Hirn, Ronald R Coifman, et al. Visualizing structure and transitions in high-dimensional biological data. *Nature biotechnology*, 37 (12):1482–1492, 2019.

[55] Aravind Subramanian, Pablo Tamayo, Vamsi K Mootha, Sayan Mukherjee, Benjamin L Ebert, Michael A Gillette, Amanda Paulovich, Scott L Pomeroy, Todd R Golub, Eric S Lander, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences*, 102(43):15545–15550, 2005.

[56] Maxim V Kuleshov, Matthew R Jones, Andrew D Rouillard, Nicolas F Fernandez, Qiaonan Duan, Zichen Wang, Simon Koplev, Sherry L Jenkins, Kathleen M Jagodnik, Alexander Lachmann, et al. Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. *Nucleic acids research*, 44(W1):W90–W97, 2016.

[57] Ilya Korsunsky, Nghia Millard, Jean Fan, Kamil Slowikowski, Fan Zhang, Kevin Wei, Yuriy Baglaenko, Michael Brenner, Po-ru Loh, and Soumya Raychaudhuri. Fast, sensitive and accurate integration of single-cell data with harmony. *Nature methods*, 16(12):1289–1296, 2019.

[58] Lindsey W Plasschaert, Rapolas Žilionis, Rayman Choo-Wing, Virginia Savova, Judith Knehr, Guglielmo Roma, Allon M Klein, and Aron B Jaffe. A single-cell atlas of the airway epithelium reveals the cftr-rich pulmonary ionocyte. *Nature*, 560(7718):377–381, 2018.

[59] Susan Tweedie, Bryony Braschi, Kristian Gray, Tamsin EM Jones, Ruth L Seal, Bethan Yates, and Elspeth A Bruford. Genenames. org: the hgnc and vgnc resources in 2021. *Nucleic acids research*, 49(D1):D939–D946, 2021.