nature methods

Article

https://doi.org/10.1038/s41592-023-01955-3

Isolating salient variations of interest in single-cell data with contrastiveVI

Received: 28 June 2022

Accepted: 25 June 2023

Published online: 7 August 2023

Check for updates

Ethan Weinberger[1,2], Chris Lin[1,2] & Su-In Lee[1]

Single-cell datasets are routinely collected to investigate changes in cellular state between control cells and the corresponding cells in a treatment condition, such as exposure to a drug or infection by a pathogen. To better understand heterogeneity in treatment response, it is desirable to deconvolve variations enriched in treated cells from those shared with controls. However, standard computational models of single-cell data are not designed to explicitly separate these variations. Here, we introduce contrastive variational inference (contrastiveVI; https://github.com/suinleelab/contrastiveVI), a framework for deconvolving variations in treatment–control single-cell RNA sequencing (scRNA-seq) datasets into shared and treatment-specific latent variables. Using three treatment–control scRNA-seq datasets, we apply contrastiveVI to perform a variety of analysis tasks, including visualization, clustering and differential expression testing. We find that contrastiveVI consistently achieves results that agree with known ground truths and often highlights subtle phenomena that may be difficult to ascertain with standard workflows. We conclude by generalizing contrastiveVI to accommodate joint transcriptome and surface protein measurements.

Single-cell technologies have emerged as powerful tools for understanding previously unexplored biological diversity. To facilitate investigation of various biological hypotheses, single-cell data are often collected simultaneously from cells in a treatment condition and from control cells. For example, recent studies have profiled cells from cancerous versus healthy tissue[1], cells exposed to drug compounds versus placebos[2] and cells with CRISPR-induced genomic perturbations versus cells with unaltered genomes[3,4]. To better understand a given phenomenon under investigation, it is desirable to isolate the low-dimensional structures and variations enriched in data from the target cells (that is, cells in the treatment condition) and compare them to a corresponding set of background cells (that is, cells in the control condition). With the development of new technologies for measuring cellular responses to large numbers of perturbations in parallel, such as Perturb-seq[3], multiplexed interrogation of gene expression through scRNA-seq (MIX-seq)[2] and multiplexing using lipid-tagged indices for single-cell and single-nucleus RNA sequencing (MULTI-seq)[5] among others, tools for refined understanding of variations unique to target versus background cells will be increasingly critical.

Isolating the variations enriched in a target dataset is the subject of contrastive analysis (CA)[6–11]. Many recent studies proposed probabilistic latent variable models for analyzing single-cell data[12–14]. However, these methods are not suitable for CA because they use a single set of latent variables to model all variations in the data. Because the variations specifically enriched in a target dataset are often subtle compared to overall variations in the data[7], such models will likely strongly entangle the enriched variations of interest with irrelevant latent factors or fail to capture the enriched variations entirely. We note that the goals of CA are fundamentally different from the related problems of batch effect correction (Supplementary Note 1) and differential abundance testing (Supplementary Note 2). Despite the many potential use cases for CA methods with single-cell data, little work has been done on the subject. One recent work[8] designed a probabilistic model for analyzing scRNA-seq count data in the CA setting. Although this method has provided new insights into variations enriched in target scRNA-seq datasets, it assumes that a linear model is sufficiently expressive to model scRNA-seq data, despite previous work that demonstrated substantial improvements by using more expressive nonlinear methods[12].

[1]Paul G. Allen School of Computer Science & Engineering, University of Washington, Seattle, WA, USA. [2]These authors contributed equally: Ethan Weinberger, Chris Lin. e-mail: suinlee@cs.washington.edu

https://doi.org/10.1038/s41592-023-01955-3

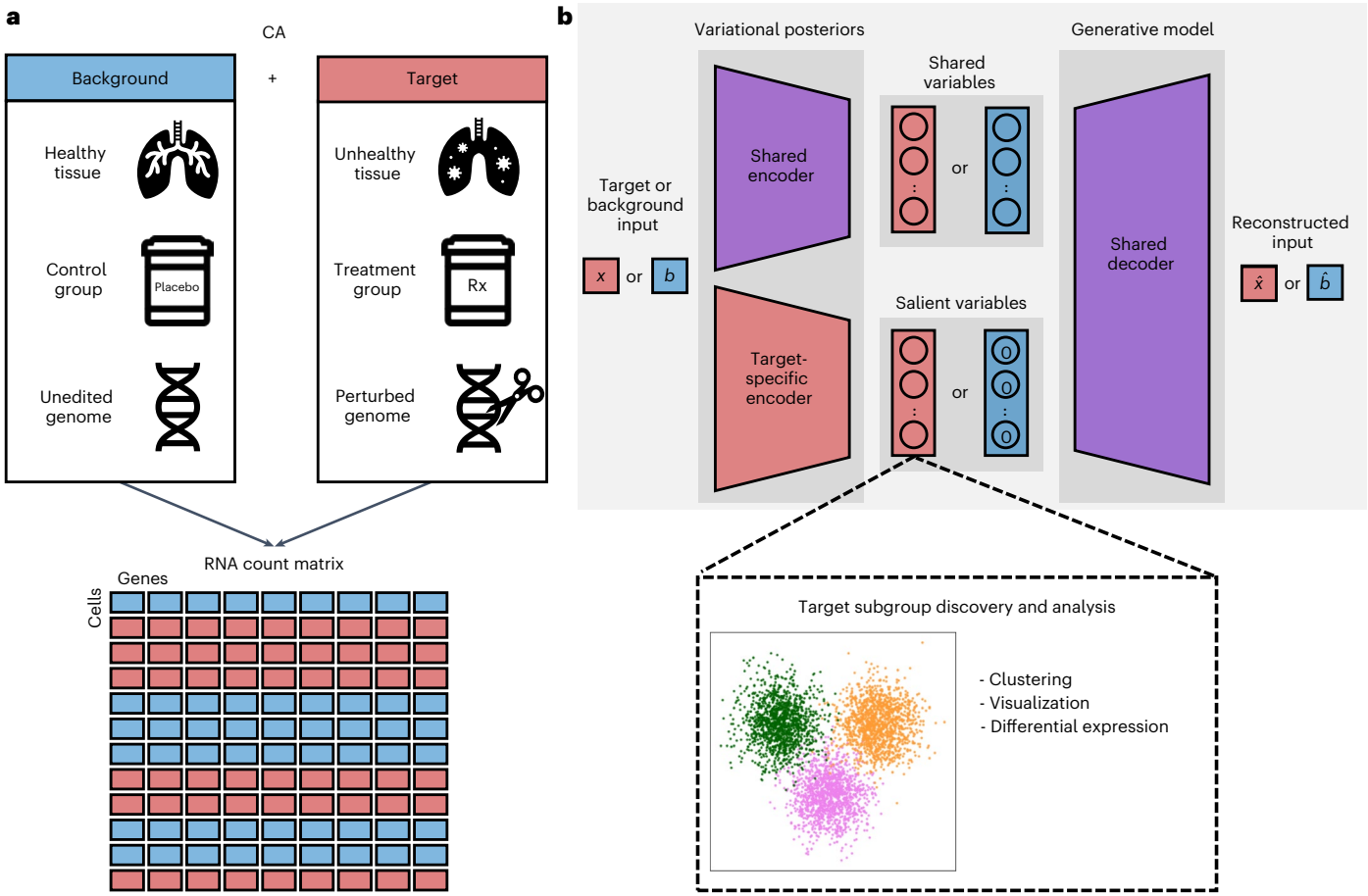

Fig. 1 | Overview of contrastiveVI. For a target dataset of interest and the corresponding background dataset, contrastiveVI separates the variations shared between the two datasets and the variations enriched in the target dataset. **a**, Example background and target data pairs. Samples from both conditions produce an RNA count matrix with each cell labeled as background or target. Rx, prescription. CA, contrastive analysis. **b**, Schematic of the contrastiveVI model. A shared encoder network embeds a cell, whether target (red) or background (blue), into the model's shared latent space, which captures variations common to target and background cells. A second target cell-specific encoder embeds target cells into the model's salient latent space, which captures variations enriched in the target data and not present in the background. For background cells, the values of the salient latent factors are fixed to be a zero vector. Both target and background cells' latent representations are transformed back to the original gene expression space using a single shared decoder network.

Furthermore, this method was not designed to incorporate information from other modalities beyond scRNA-seq.

To address these limitations, we developed contrastiveVI, a deep generative model for analyzing scRNA-seq data in the CA setting (Fig. 1). contrastiveVI models the variations underlying scRNA-seq data using two sets of latent variables: the first, called 'shared variables', are common to background and target cells, while the second, called 'salient variables', are used to model variations specific to target data. contrastiveVI can be used for many analysis tasks, including dimensionality reduction, clustering and differential gene expression testing. To highlight this functionality, we applied our model to analyze three real-world scRNA-seq datasets. We also generalized our framework to multimodal datasets by developing totalContrastiveVI, a model for analyzing joint RNA and surface protein measurements in the CA setting, and applied it to analyze CRISPR-induced variations in an expanded CRISPR-compatible cellular indexing of transcriptomes and epitopes by sequencing (ECCITE-seq)[15] dataset.

Results

The contrastiveVI model

contrastiveVI uses a probabilistic latent variable model to represent uncertainty in observed RNA counts as a combination of biological and technical factors. Model input consists of an RNA count matrix and labels denoting whether each cell is a background or a target cell (Fig. 1a). contrastiveVI encodes each cell as the parameters of a distribution in a low-dimensional latent space. This latent space is divided into two parts, each with its own encoding function. The first set of latent variables, called shared variables, captures factors of variation common to both background and target data. The second set, denoted as salient variables, captures variations unique to the target dataset. Only target data points are assigned salient latent variable values; background data points are instead assigned a zero vector for the salient variables to represent their absence. contrastiveVI also provides a means to estimate the parameters of the distributions underlying the observed RNA measurements given a cell's latent representation. Such distributions explicitly account for technical factors in the observed data, such as sequencing depth, batch effects and dropouts. All distributions are parameterized by neural networks.

The contrastiveVI model is based on the variational autoencoder (VAE) framework[16]. As such, its parameters can be learned using efficient stochastic optimization techniques, easily scaling to large scRNA-seq datasets consisting of tens or hundreds of thousands of cells. Following optimization, we can make use of the various contrastiveVI model components for downstream analyses. For example, the salient latent representations of target samples can be used as inputs to clustering or visualization algorithms to discover subgroups of

target points. Moreover, the distributional parameters can be used for additional tasks such as imputation or differential gene expression analysis. A more detailed description of the contrastiveVI model can be found in the Methods.

Analyzing cell line responses to a small-molecule therapy

After initially validating contrastiveVI on a simulated dataset (Supplementary Note 3), we next applied contrastiveVI to analyze a real-world scRNA-seq dataset collected using the recently developed MIX-seq[2] platform. MIX-seq measures the transcriptional responses of up to hundreds of cancer cell lines in parallel after being treated with one or more small-molecule compounds. Here, our target dataset contained measurements collected by McFarland et al.[2] from 24 cell lines treated with idasanutlin. The small molecule idasanutlin is an antagonist of MDM2, a negative regulator of the tumor-suppressor protein p53, hence offering cancer therapeutic opportunities[17]. It is known[17] that idasanutlin induces activation of the p53 pathway in cell lines with wild-type *TP53*, while transcriptionally inactive mutant *TP53* cell lines do not respond to the compound. For the background dataset, we used measurements from the same cell lines treated with the control compound dimethyl sulfoxide (DMSO).

We began our analysis of this dataset by confirming that contrastiveVI's representations agreed with prior knowledge. As variations that distinguished cell lines were shared across treatment and control cells, we would expect contrastiveVI's shared latent space to clearly separate cells by cell line. Moreover, we would expect increased mixing between DMSO- and idasanutlin-treated cells compared to the original visualization workflow of McFarland et al.[2] (Supplementary Fig. 1), even for cell lines with a wild-type *TP53* gene. We found that cells indeed clearly separated by cell line in contrastiveVI's shared latent space (Fig. 2a). In addition, for *TP53*-wild-type cell lines (Fig. 2b), we observed stronger mixing across treatments (Fig. 2c) as desired. While some *TP53*-wild-type cell lines (for example, RCC10RGB) were not as uniformly mixed as *TP53*-mutant cell lines, we found that this phenomenon corresponded to differences in the proportions of cells in each phase of the cell cycle for treatment versus control cells (Supplementary Note 4).

We next turned our attention toward contrastiveVI's salient representations of treatment cells. Based on idasanutlin's mechanism of action, we would expect separation between *TP53*-wild-type and *TP53*-mutant cell lines. Moreover, because *TP53*-mutant cell lines all exhibit the same (non-)response to the compound, we would expect strong mixing of the *TP53*-mutant cell lines. Qualitatively, we indeed observed clear mixing of *TP53*-mutant cell lines and a separation of cells by *TP53*-mutation status in contrastiveVI's salient latent space (Fig. 2d,e). In our analysis of contrastiveVI's salient latent space, we also observed separation between the individual idasanutlin-responding *TP53*-wild-type cell lines, potentially reflecting cell line-specific responses to the compound. To better understand which genes drove this separation, we used Hotspot[18], a tool for identifying informative genes in a single-cell dataset by ranking genes in terms of spatial autocorrelation with respect to a given metric of cell–cell similarity (for example, the latent space of a VAE). We found (Fig. 2f; see Supplementary Fig. 2 for additional genes) that the top genes returned by Hotspot when applied to contrastiveVI's salient latent space consisted of those encoding members of the p53 signaling pathway, such as *TP53I3* and *CDKN1A*, as well as well-known targets of p53, such as *SUGCT* and *FDXR*. Moreover, we found qualitatively that idasanutlin-induced overexpression of these genes appeared specific to the cell line, with some genes, such as *SUGCT*, only upregulated in a subset of *TP53*-wild-type cell lines. We confirmed these findings quantitatively by using contrastiveVI to impute denoised expression values and perform a differential expression test similar to that of single-cell variational inference (scVI) (Supplementary Note 5).

We compared contrastiveVI's representations with those learned by previously proposed linear contrastive models: the contrastive latent variable model (CLVM) of Severson et al.[10] as well the contrastive Poisson latent variable model (CPLVM) and the contrastive generalized latent variable model (CGLVM) proposed by Jones et al.[8] (Methods). Qualitatively, we found that baseline contrastive models' representations all disagreed with prior knowledge. For example, cells exhibit substantially worse separation by cell line in baseline contrastive models' shared latent spaces than contrastiveVI's shared latent space (Supplementary Fig. 3). Moreover, despite not responding to the treatment, some *TP53*-mutant cell lines clearly separate in CLVM's and CPVLM's salient latent spaces, which could result in misleading conclusions (Supplementary Fig. 4).

We also compared contrastiveVI's embeddings to those returned by non-contrastive scRNA-seq analysis workflows. In particular, we applied principal-component analysis (PCA) as well as scVI[12] and deep count autoencoder (DCA)[19], two deep learning models for scRNA-seq data. We found (Supplementary Fig. 5) that these methods primarily separated cells by cell line with additional visible shifts in *TP53*-wild-type cell lines as a result of idasanutlin treatment. We also trained an scVI model with target versus background as a batch label. Similar to the scVI model trained without this label, we found (Supplementary Fig. 6) that cells primarily separated by cell line with some shifts between treatment and control cells for *TP53*-wild-type cell lines. However, because these methods do not explicitly deconvolve shared and perturbation-specific variations, it is not clear whether the changes in expression driving these shifts for *TP53*-wild-type cell lines were shared across cell lines or whether they were cell line specific. On the other hand, contrastiveVI's salient space immediately highlighted cell line-specific effects.

Finally, to systematically compare across methods, we computed a suite of metrics quantifying the quality of baseline models' salient and shared latent representations (Fig. 2g). These metrics were chosen to capture how well each model's representations agreed with prior knowledge for this dataset: that is, for contrastive models' salient representations, we quantified the separation of *TP53*-mutant and *TP53*-wild-type cells (*TP53* adjusted Rand index (ARI), *TP53* normalized mutual information (NMI), *TP53* silhouette) and mixing of the *TP53*-mutant cell lines (entropy of mutant cell line mixing, mutant cell line mixing silhouette); for shared representations, we measured the separation of individual cell lines (cell line ARI, cell line NMI, cell line silhouette) and mixing across treatments (entropy of treatment mixing, treatment mixing silhouette). As a further comparison, we also computed these same metrics for the latent spaces returned by our non-contrastive baseline workflows. While baseline models all performed poorly on at least one metric, we found that contrastiveVI consistently achieved strong performance across all metrics.

Uncovering cell type-specific responses to pathogens

We next applied contrastiveVI to a more complex dataset with multiple perturbations collected by Haber et al.[20]. This dataset consists of gene expression measurements of intestinal epithelial cells from mice infected with either *Salmonella enterica* (*Salmonella*) or *Heligmosomoides polygyrus*. As a background for this dataset, we used measurements collected from healthy control cells released by the same authors.

We began our analysis by confirming that contrastiveVI's salient and shared representations agreed with high-level prior findings from Haber et al.[20]. As variations that distinguished cell types were shared across treatment and control cells, we would expect cells to separate primarily by cell type and mix across perturbations in the shared latent space. Qualitatively, we found that cells indeed separated by cell type (Fig. 3a) and generally mixed across treatments (Fig. 3b) in contrastiveVI's shared latent space. As noted by Haber et al.[20], *Salmonella* and *H. polygyrus* both induced substantial pathogen-specific changes in

https://doi.org/10.1038/s41592-023-01955-3

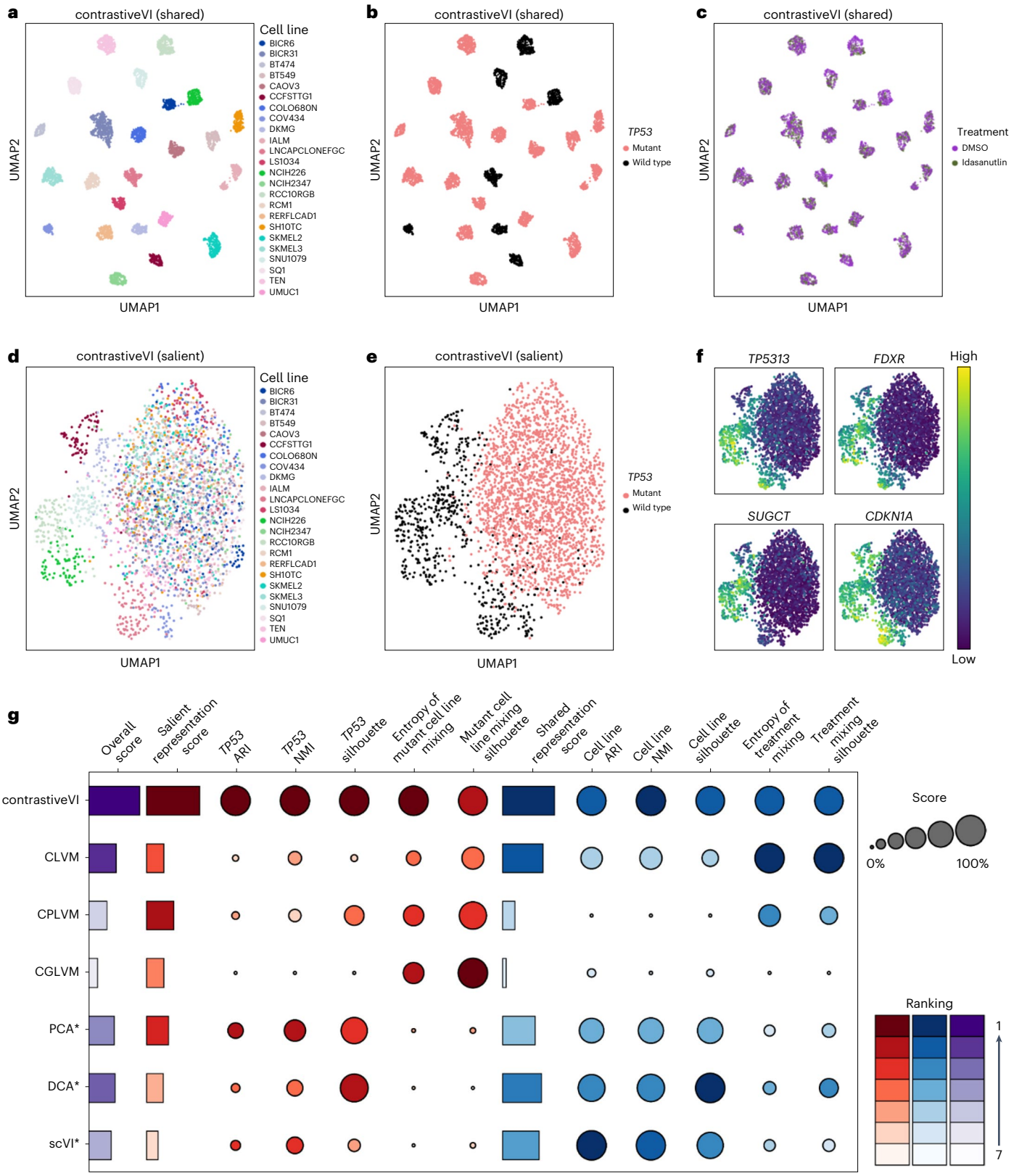

Fig. 2 | contrastiveVI isolates idasanutlin-induced variations in cancer cell lines. a–**c**, UMAP plots of contrastiveVI's shared latent representations for idasanutlin-treated and control cells from McFarland et al.[2] colored by cell line (**a**), *TP53*-mutation status (**b**) and treatment (**c**). **d**–**f**, UMAP plots of contrastiveVI's salient latent space colored by cell line (**d**), *TP53*-mutation status (**e**) and expression levels of the top four genes returned by Hotspot[18] (**f**). RNA expression values depicted in **f** were denoised using contrastiveVI and then log library size transformed (Methods). **g**, Quantitative comparison of salient and shared representation quality for contrastiveVI and baseline methods. *Non-contrastive baselines, for which metrics were computed on the method's single latent space. Individual metrics were scaled to lie between 0 and 1, and overall scores were computed by averaging salient and shared representation scores. Raw metric values are available in Supplementary Tables 1 and 2. Higher values for all metrics indicate better performance.

https://doi.org/10.1038/s41592-023-01955-3

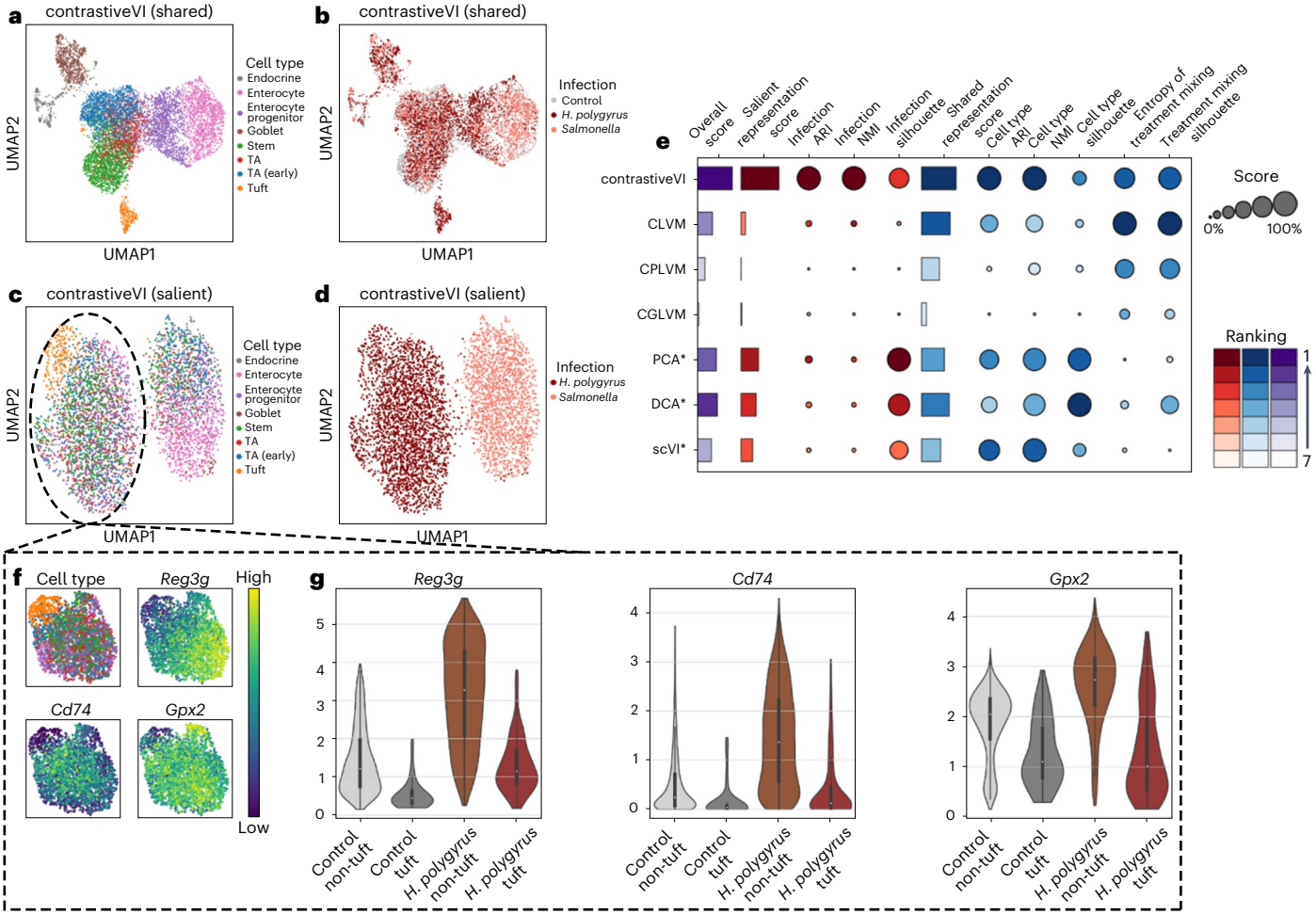

Fig. 3 | contrastiveVI uncovers cell type-specific responses to pathogen infections in mouse intestinal epithelial cells. a,**b**, UMAP plots of contrastiveVI's shared latent representations of treatment (that is, *H. polygyrus*- or *Salmonella*-infected) cells and control cells colored by cell type (**a**) and infection type (**b**). TA, transit amplifying. **c**,**d**, UMAP plots of contrastiveVI's salient representations of *Salmonella*- and *H. polygyrus*-infected epithelial cells, colored by cell type (**c**) and infection (**d**). **e**, Quantitative evaluation of contrastiveVI and baseline models' latent salient and shared representations' agreement with high-level prior knowledge. *Non-contrastive baseline methods, for which metrics were computed on the given method's single latent space. Metrics were normalized as in Fig. 2. Raw values for salient and shared space metrics are available in Supplementary Tables 3 and 4, respectively. **f**,**g**, Further analysis of contrastiveVI's salient representations of *H. polygyrus*-infected cells. RNA expression values depicted in **f**,**g** were denoised using contrastiveVI and then log library size transformed (Methods). Centers of box plots represent median expression values, and upper (lower) box bounds denote the third (first) quartile; upper (lower) whiskers represent the third quartile + 1.5 × interquartile range (first quartile − 1.5 × interquartile range). Minimum and maximum values are denoted by the ends of the corresponding violin plots. Violin plots depict expression values for non-tuft control cells (n = 3,180), control tuft cells (n = 60), *H. polygyrus*-infected non-tuft cells (n = 2,494) and *H. polygyrus*-infected tuft cells (n = 217).

gene expression. Moreover, while a small proportion of these changes were noted to be cell type specific (for example, enterocyte-specific *Salmonella*-induced gene expression changes), most were shared across all cell types. We would thus expect cells to separate primarily by pathogen in contrastiveVI's salient space with increased mixing across cell types. We found (Fig. 3c,d) that cells indeed primarily separated by pathogen with substantially increased mixing across cell types in contrastiveVI's salient latent space. We then qualitatively (Supplementary Note 6) and quantitatively (Fig. 3e) benchmarked contrastiveVI's embeddings against those of baseline models, and we found that baselines' representations frequently disagreed with prior knowledge.

We proceeded to further investigate the additional patterns revealed in contrastiveVI's salient latent space. In particular, we considered the notable separation of *Salmonella*-infected enterocytes from the broader *Salmonella* cluster and of *H. polygyrus*-infected tuft cells from the broader *H. polygyrus* cluster. We first confirmed that these separations were indeed due to cell type-specific changes in gene expression and not simply due to changes in cell type proportions (Supplementary Note 7). As enterocyte-specific *Salmonella*-induced gene expression patterns were already analyzed by Haber et al.[20], we focused the remainder of our analysis on the separation of *H. polygyrus*-infected tuft cells from the broader cluster of *H. polygyrus* cells. To accomplish this, we used Hotspot[18] to uncover the most strongly spatially autocorrelated genes for contrastiveVI's salient representations of *H. polygyrus*-infected cells. For this analysis, we excluded the known tuft marker genes provided by Haber et al.[20], which would exhibit high autocorrelation due to the separation of tuft cells even without any infection-induced changes.

We found that the top ten most spatially autocorrelated genes returned by Hotspot included a number of genes, such as *Reg3b*, *Cd74* and *Gpx2*, associated with the inflammatory response in the intestinal epithelium[21–23]. Upon further inspection, we found that these genes exhibited substantially lower expression in the separated tuft cells than in *H. polygyrus*-infected cells from other cell types (Fig. 3f). Moreover, we found that these genes were significantly upregulated in *H. polygyrus*-infected non-tuft cells compared to non-tuft controls

and yet were not upregulated or were upregulated to a much lesser degree in *H. polygyrus*-infected tuft cells than in control tuft cells (Fig. 3g). This muted upregulation of inflammatory response genes in tuft cells may reflect their distinct role in the type 2 immune response[24]. We note that these tuft-cell-specific patterns in the expression of inflammatory response genes were not discussed by Haber et al.[20] and could potentially have been obscured by the standard analysis workflow employed in that work. For example, Haber et al.[20] found that *Reg3g* was differentially expressed between *H. polygyrus*-infected cells and controls for each individual cell type (false discovery rate $< 1 \times 10^{-13}$ for each cell type). However, this result does not indicate whether the magnitudes of these differences were cell type specific. On the other hand, our Hotspot analysis of contrastiveVI's salient latent space clearly highlighted the presence of cell type-specific effects for *Reg3g* and other inflammation response genes.

Exploring CRISPR-induced variations in a Perturb-seq screen

We next applied contrastiveVI to reanalyze a Perturb-seq dataset originally collected by Norman et al.[4]. In that study, the authors assessed the effects of 284 different CRISPR-mediated perturbations on the growth of K562 cells, where each perturbation induced overexpression of a single gene or a pair of genes. Here, we focused on a subset of these perturbations, which the authors found grouped into stable clusters as determined by applying the HDBSCAN[25] algorithm to the mean expression profile of each perturbation. After obtaining these clusters, Norman et al.[4] then labeled each cluster as expressing a corresponding gene program. In our reanalysis of this dataset, we sought to understand whether analyzing the data at the resolution of individual cells, as opposed to perturbations' mean expression profiles, could provide additional insights beyond those noted in the original analysis of Norman et al.[4].

Based on the authors' original findings, we would expect cells to separate based on these gene program labels. However, when examining the perturbed cells using non-CA workflows, we found substantial confounding due to cell cycle stage, leading to poor separation of the labeled gene programs (Fig. 4a and Supplementary Fig. 7). Using measurements from control cells infected with non-targeting guides as a background, we next applied contrastiveVI and our baseline contrastive models to this dataset. We found (Fig. 4b) substantially increased mixing of cells across cycle phases and much stronger separation by labeled gene programs in contrastiveVI's salient latent space as desired. On the other hand, we found that cells continued to mix across gene programs in CPLVM's and CGLVM's salient latent space, and G1 phase cells continued to clearly separate from other cells in CLVM's salient latent space (Supplementary Fig. 8). We also quantified how well each method separated cells by gene program labels, and we found that contrastiveVI achieved significantly better separation than baseline methods (Fig. 4c). Notably, given increased mixing of cells across cell cycle phases in contrastiveVI's salient latent space, the clear separation of cells with perturbations labeled as 'G1 cell cycle arrest' by Norman et al.[4] may at first appear counterintuitive. Upon further investigation (Supplementary Note 8), we found that these cells exhibited an additional unique non-cell cycle-related perturbation effect not discussed by Norman et al.[4] and thus indeed would be expected to separate in contrastiveVI's salient latent space.

During our analysis, we also observed that the cells labeled as expressing an induced granulocyte–apoptosis gene program grouped into multiple distinct subclusters in contrastiveVI's salient latent space. Thus, to further demonstrate how contrastiveVI could provide insights into this dataset not discussed by Norman et al.[4], we investigated this separation in more detail. After rerunning uniform manifold approximation and projection (UMAP) solely on contrastiveVI's salient representations of granulocyte–apoptosis-labeled cells (Fig. 4d), we observed two clear groups of cells perturbed to overexpress *CEBPB* and *SPI1*, respectively, that separated from a larger main cluster. We also noticed that, while most cells perturbed for *CEBPB* could be found in the *CEBPB*-specific cluster, some were also mixed in with the larger cluster. We then proceeded to explore the differences between these two groups of *CEBPB*-perturbed cells. We found (Fig. 4e,f) that some genes, such as the CCAAT enhancer-binding protein β (CEBPB) target *PFN2* (ref. 26–28), were upregulated in both clusters compared to control cells, indicating that the perturbation was successful for both groups. However, we also found that granulocyte marker genes, such as *LST1*, *CEBPE* and *ITGAM*, were overexpressed in the 'mixed' *CEBPB*-perturbed cells compared both to control cells and *CEPBP*-perturbed cells in the 'separate' cluster. This phenomenon indicates a heterogeneous response to the perturbation that could potentially be missed by perturbation-level workflows similar to that of Norman et al.[4].

Extending multimodal models for CA

To better understand the effects of genomic perturbations on cell state, new platforms such as ECCITE-seq[15] and Perturb-CITE-seq[29] have been developed that enable multimodal readouts of CRISPR-perturbed cells. As with unimodal platforms, previous work[30] has reported that the analysis of multimodal perturbation screens may be complicated by the presence of confounding sources of variation shared with control cells. To demonstrate how our framework can be easily extended to multimodal single-cell data, we extended totalVI[31], a deep generative model for CITE-seq data, using the contrastive latent variable modeling techniques employed in contrastiveVI. Our resulting totalContrastiveVI model (Methods) isolates perturbed cell-specific and shared variations present in joint RNA and protein measurements in its salient and shared latent spaces, respectively. Moreover, by building off of totalVI, totalContrastiveVI inherits totalVI's additional downstream analysis capabilities that can be leveraged to better understand the patterns captured by totalContrastiveVI's latent spaces.

To highlight totalContrastiveVI's capabilities, we applied it to analyze an ECCITE-seq dataset from Papalexi et al.[30]. In that work, the authors sought to explore regulatory networks underlying expression of immune checkpoint molecules, such as programmed death ligand (PD-L)1, in THP-1 (ref. 32) cells. To accomplish this, they measured cells' transcriptomes alongside surface protein levels of the proteins PD-L1, PD-L2, CD86 and CD366 for cells perturbed by one of 111 CRISPR guides as well as for a set of control cells infected with non-targeting guide RNA (gRNA). As a baseline, we first applied totalVI[31] to learn a lower-dimensional representation of the perturbed cells. Ideally, the model would capture perturbation-induced variations; however, we found instead that totalVI's latent space was confounded by numerous alternative sources of variation, including transduction replicate identity, cell cycle stage and activation of a gene program relating to the cellular stress response (Fig. 5a and Supplementary Fig. 9).

Using measurements from control cells infected with non-targeting guides as a background, we next applied totalContrastiveVI to this dataset. As expected, we found that the totalContrastiveVI shared latent space was dominated by nuisance variations (Supplementary Fig. 10). By contrast, the totalContrastiveVI salient latent space exhibited a clear clustering structure invariant to replicate identity, cell cycle stage and cellular stress response (Fig. 5b and Supplementary Fig. 11). Of the three clusters revealed in totalContrastiveVI's salient latent space, we found that one consisted of cells perturbed for genes defined by Papalexi et al.[30] as upstream components of the interferon (IFN)-γ pathway, one consisted solely of cells perturbed for *IRF1*, which encodes an IFN-γ mediator, and the remaining cluster consisted of cells from all perturbations (Fig. 5c). We found that these clusters corresponded to distinct RNA expression patterns of immune response-related genes, with strong downregulation in cells perturbed for upstream components of the IFN-γ pathway and weaker but still notable downregulation in cells perturbed for *IRF1* (Supplementary Fig. 12a). We also observed downregulation of the PD-L1 and PD-L2 proteins for cells perturbed

https://doi.org/10.1038/s41592-023-01955-3

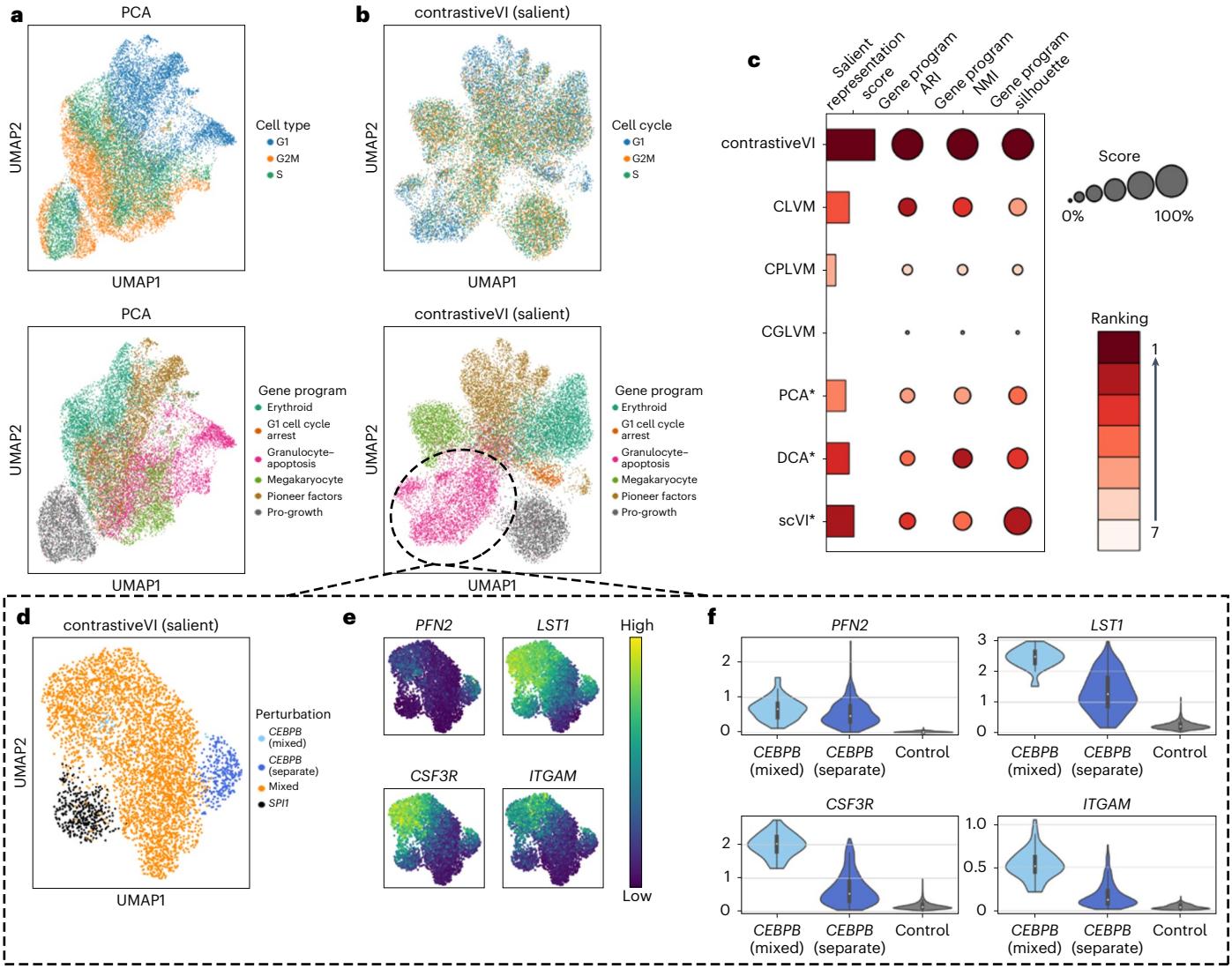

Fig. 4 | contrastiveVI's salient latent space isolates CRISPR perturbation-induced variations in a large-scale Perturb-seq experiment. a,**b**, UMAP plots of a standard scRNA-seq analysis workflow consisting of normalization followed by PCA (**a**) and contrastiveVI's salient latent space (**b**) colored by cell cycle stage (top) and induced gene program identified by ref. 4 (bottom). **c**, Quantitative metrics capturing separation by gene program label in contrastive models' salient latent spaces and non-contrastive models' single latent spaces. *Non-contrastive baseline methods, for which metrics were computed on the given method's single latent space. Metrics were normalized as in Fig. 2. Raw values for metrics are available in Supplementary Table 5. **d**–**f**, Exploration of the granulocyte–apoptosis subclusters revealed in contrastiveVI's salient latent space. RNA expression values depicted in **e**,**f** were denoised using contrastiveVI and then log library size transformed (Methods). Centers of box plots represent median expression values, and upper (lower) box bounds denote the third (first) quartile; upper (lower) whiskers represent the third quartile + 1.5 × interquartile range (first quartile − 1.5 × interquartile range). Minimum and maximum values are denoted by the ends of the corresponding violin plots. Violin plots depict expression values for *CEBPB*-perturbed cells ($n = 311$) that formed a separate cluster in the UMAP plot in **d**, a group of *CEBPB*-perturbed cells that mixed with the larger main cluster with other perturbations ($n = 52$) and control cells ($n = 7{,}275$).

for upstream components of the IFN-γ pathway (Supplementary Fig. 12b). In addition, we compared totalContrastiveVI's results to those obtained with totalVI when provided with covariates reflecting the known unwanted sources of variation in this dataset. Despite requiring prior knowledge of the unwanted sources of variation, we found that this procedure did not fully mitigate the impact of these undesirable variations (Supplementary Note 9).

Similar clusters of perturbed cells were found by Papalexi et al.[30] using a nearest-neighbor-based approach applied to transcriptomic measurements. Thus, to further highlight the merits of our approach over previous workflows, we applied totalContrastiveVI's downstream analysis tools inherited from totalVI to analyze the patterns found in totalContrastiveVI's salient latent space in greater depth. As a case study, we focused on analyzing cells infected with *IFNGR2*-targeting gRNA. While most of these cells clustered with cells perturbed for other members of the IFN-γ pathway in totalContrastiveVI's salient latent space, a substantial number belonged to the larger mixed cluster containing cells infected with all gRNA species (Fig. 5d). This heterogeneity in response to *IFNGR2* perturbation was also noted by Papalexi et al.[30], and, to investigate it, the authors of that study inspected *IFNGR2* sequencing reads overlapping the corresponding gRNA cut site from the two groups of cells infected with *IFNGR2* gRNA. It was found that cells infected by the *IFNGR2* gRNA and that clustered with cells infected by gRNA species targeting other members of the IFN-γ pathway exhibited frameshift indel mutations at the gRNA cut site, indicating successful knockout (KO) of the *IFNGR2*

https://doi.org/10.1038/s41592-023-01955-3

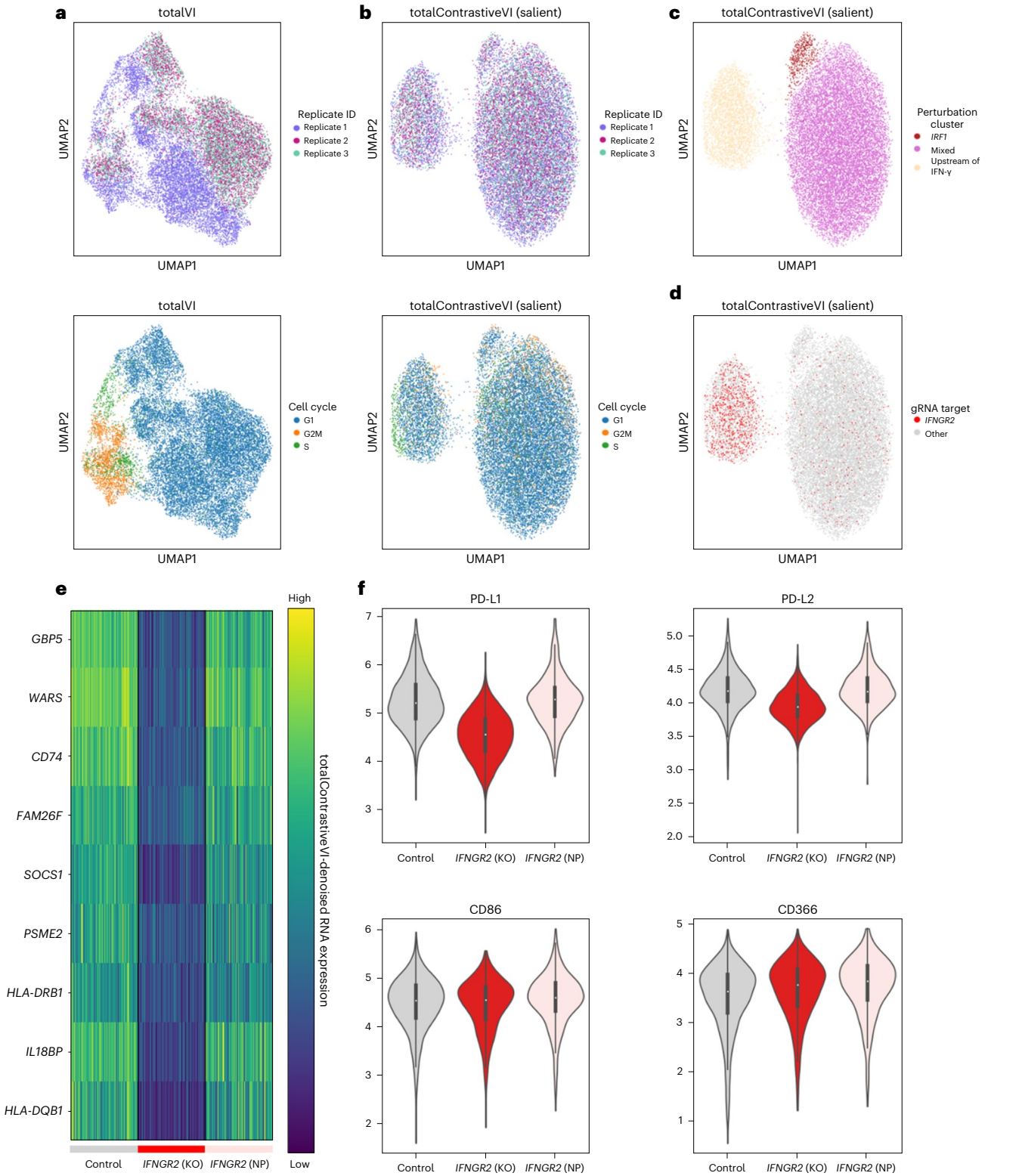

Fig. 5 | totalContrastiveVI isolates perturbation-induced variations in joint RNA and protein measurements. a,**b**, UMAP visualizations of totalVI's embeddings (**a**) and totalContrastiveVI's salient embeddings (**b**) colored by replicate number and cell cycle stage. **c**, Visualization of the three clusters revealed in the totalContrastiveVI salient latent space. **d**, Visualization of cells that expressed *IFNGR2* gRNA in totalContrastiveVI's salient latent space. **e**, totalContrastiveVI-denoised RNA expression levels (log library size normalized; Methods) of immune-related genes for control cells and *IFNGR2*-KO cells as well as cells expressing *IFNGR2* gRNA but that were NP. *WARS* denotes expression of the *WARS1* gene and *FAM26F* denotes expression of the *CALHM6* gene. **f**, Distributions of log (totalContrastiveVI denoised protein + 1) for control (n = 2,386) cells as well as *IFNGR2*-KO (n = 887) and NP (n = 320) cells. Centers of box plots represent median expression values, and upper (lower) box bounds denote the third (first) quartile; upper (lower) whiskers represent the third quartile + 1.5 × interquartile range (first quartile − 1.5 × interquartile range). Minimum and maximum values are denoted by the ends of the corresponding violin plots.

gene. On the other hand, the other set of *IFNGR2* gRNA-infected cells lacked these deleterious mutations, indicating that the perturbation was not successful.

We then applied totalContrastiveVI's downstream analysis workflows to further analyze the two clusters of *IFNGR2* gRNA-infected cells and control cells. We began by considering the non-perturbed (NP) *IFNGR2* gRNA-infected cells. As a first step in analyzing this cluster, we used totalContrastiveVI to obtain denoised RNA expression values (Fig. 5e) and protein counts (Fig. 5f). As expected, we found no notable differences in RNA and protein expression between control and NP cells. We verified this phenomenon using totalContrastiveVI's RNA and protein differential expression workflows, which correctly did not identify any differentially expressed genes or proteins between the NP *IFNGR2* gRNA-infected cells and control cells. We next considered the KO cells. Qualitatively, we observed substantial differences in totalContrastiveVI's denoised RNA and protein expression levels for these cells compared to controls. These results were confirmed using totalContrastiveVI's differential expression workflow, which identified 20 genes (Supplementary Table 6) and the PD-L1 and PD-L2 proteins (Supplementary Table 7) as differentially expressed. This list of genes largely consisted of immune response-related genes, with strong enrichment for immune response pathways (Supplementary Table 8), such as the IFN-γ signaling pathway (adjusted P value $< 1 \times 10^{-9}$) and the PD-1–PD-L1 signaling pathway (adjusted P value $< 1 \times 10^{-8}$). These results are expected, as *IFNGR2* is a known upstream component of the IFN-γ pathway[33], and IFN-γ has been found to have a major effect on PD-1 and PD-L1 expression in cancer cells[34]. We compared totalContrastiveVI's results with a normalization and differential expression workflow similar to that of Papalexi et al.[30]. We found (Supplementary Note 10) that such a workflow led to potentially erroneous differential expression results and was highly sensitive to choice of normalization procedure, thus illustrating the advantages of our deep generative modeling approach.

Discussion

In this work, we introduced contrastiveVI, a framework designed to deconvolve enriched variations in a target single-cell dataset from those shared with a related background dataset using deep CA models. In three different contexts (exposure to drug compounds, response to infection by different pathogens and genomic perturbation via CRISPR guides), we found that contrastiveVI successfully isolated enriched variations in target cells' transcriptomes. By isolating these target cell-specific variations, we found that contrastiveVI highlighted subtle perturbation-induced gene expression patterns in target cells that are more difficult to ascertain using standard single-cell analysis workflows. Similarly, we found that our multimodal totalContrastiveVI model successfully isolated enriched variations in a joint RNA and surface protein dataset.

In our experiments, we found that contrastiveVI's neural network-based modeling approach led to substantially improved performance at recovering known biological ground truths compared to that of linear baseline contrastive models. However, these gains come at the cost of inherent interpretability. While linear baseline models' coefficients can be manually inspected to understand the patterns captured by these models, such an analysis is not feasible for more complex deep neural network models. To mitigate this issue, authors of recent works[18,35,36] have proposed methods to aid in interpreting the representations learned by unsupervised deep learning models. In this work, we specifically employed Hotspot[18] to identify genes that were highly spatially autocorrelated in contrastiveVI's latent spaces, and we validated that these genes reflected meaningful treatment-induced phenomena using contrastiveVI's additional downstream analysis capabilities (for example, imputation and differential expression testing).

A crucial consideration in the use of CA models such as contrastiveVI is choice of an appropriate background dataset. In general, the background dataset should reflect confounding sources of variation that one seeks to mitigate in the analysis of target cells, such as variations due to the cell cycle or differences between cell types that are unrelated to the given treatment. Because single-cell experiments often simultaneously profile cells in both treatment and control conditions, a suitable background dataset is often readily available for a given target dataset. Notably, in our analyses, we found that contrastiveVI was relatively robust to imbalances in the proportions of confounding variations shared between treatment and control cells (for example, significant differences in the proportions of cells in each cell cycle state for target versus background cells). However, if these imbalances are too extreme, then the use of contrastiveVI may not be appropriate. For example, if a cell type is present in the target dataset but missing entirely from the background, then contrastiveVI may produce misleading results. Moreover, caution should be exercised when applying contrastiveVI to analyze datasets from multiple studies with different experimental setups. In this case, beyond issues of background dataset suitability, any results may be confounded by technical differences between studies (that is, study-specific batch effects or differences due to varying sequencing platforms).

The ideas underpinning contrastiveVI admit multiple potential directions for future work. Our contrastive modeling approach could be further extended to handle additional modalities beyond RNA transcript and protein counts by integrating previously proposed modality-specific likelihoods, such as that of peakVI[37] for chromatin accessibility, into the contrastiveVI framework. Moreover, recent work[38,39] has demonstrated success in learning more inherently interpretable representations of single-cell data, where each dimension of the representation corresponds to a known gene pathway. Incorporating such constraints into contrastiveVI could shed further light on the biological phenomena captured in contrastiveVI's salient and shared latent spaces.

Online content

Any methods, additional references, Nature Portfolio reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41592-023-01955-3.

References

1. Zheng, G. X. Y. et al. Massively parallel digital transcriptional profiling of single cells. *Nat. Commun.* **8**, 14049 (2017).
2. McFarland, J. M. et al. Multiplexed single-cell transcriptional response profiling to define cancer vulnerabilities and therapeutic mechanism of action. *Nat. Commun.* **11**, 4296 (2020).
3. Dixit, A. et al. Perturb-seq: dissecting molecular circuits with scalable single-cell RNA profiling of pooled genetic screens. *Cell* **167**, 1853–1866 (2016).
4. Norman, T. M. et al. Exploring genetic interaction manifolds constructed from rich single-cell phenotypes. *Science* **365**, 786–793 (2019).
5. McGinnis, C. S. et al. MULTI-seq: sample multiplexing for single-cell RNA sequencing using lipid-tagged indices. *Nat. Methods* **16**, 619–626 (2019).
6. Zou, J. Y., Hsu, D. J., Parkes, D. C. & Adams, R. P. Contrastive learning using spectral methods. *Adv. Neural Inf. Process. Syst.* **26**, 2238–2246 (2013).
7. Abid, A., Zhang, M. J., Bagaria, V. K. & Zou, J. Exploring patterns enriched in a dataset with contrastive principal component analysis. *Nat. Commun.* **9**, 2134 (2018).
8. Jones, A., Townes, W. F., Li, D. & Engelhardt, B. E. Contrastive latent variable modeling with application to case–control sequencing experiments. *Ann. Appl. Stat.* **16**, 1268–1291 (2022).

9. Li, D., Jones, A. & Engelhardt, B. Probabilistic contrastive principal component analysis. Preprint at *arXiv* https://doi.org/10.48550/arXiv.2012.07977 (2020).
10. Severson, K. A., Ghosh, S. & Ng, K. Unsupervised learning with contrastive latent variable models. In *Proceedings of the AAAI Conference on Artificial Intelligence* **33**, 4862–4869 (AAAI, 2019).
11. Abid, A. & Zou, J. Contrastive variational autoencoder enhances salient features. Preprint at *arXiv* https://doi.org/10.48550/arXiv.1902.04601 (2019).
12. Lopez, R., Regier, J., Cole, M. B., Jordan, M. I. & Yosef, N. Deep generative modeling for single-cell transcriptomics. *Nat. Methods* **15**, 1053–1058 (2018).
13. Risso, D., Perraudeau, F., Gribkova, S., Dudoit, S. & Vert, J.-P. A general and flexible method for signal extraction from single-cell RNA-seq data. *Nat. Commun.* **9**, 284 (2018).
14. Lotfollahi, M. et al. Mapping single-cell data to reference atlases by transfer learning. *Nat. Biotechnol.* **40**, 121–130 (2021).
15. Mimitou, E. P. et al. Multiplexed detection of proteins, transcriptomes, clonotypes and CRISPR perturbations in single cells. *Nat. Methods* **16**, 409–412 (2019).
16. Kingma, D. P. & Welling, M. Auto-encoding variational Bayes. In *2nd International Conference on Learning Representations* (ICLR, 2015).
17. Vassilev, L. T. et al. In vivo activation of the p53 pathway by small-molecule antagonists of MDM2. *Science* **303**, 844–848 (2004).
18. DeTomaso, D. & Yosef, N. Hotspot identifies informative gene modules across modalities of single-cell genomics. *Cell Syst.* **12**, 446–456 (2021).
19. Eraslan, G., Simon, L. M., Mircea, M., Mueller, N. S. & Theis, F. J. Single-cell RNA-seq denoising using a deep count autoencoder. *Nat. Commun.* **10**, 390 (2019).
20. Haber, A. L. et al. A single-cell survey of the small intestinal epithelium. *Nature* **551**, 333–339 (2017).
21. Loonen, L. M. et al. Reg3γ-deficient mice have altered mucus distribution and increased mucosal inflammatory responses to the microbiota and enteric pathogens in the ileum. *Mucosal Immunol.* **7**, 939–947 (2014).
22. Farr, L. et al. Cd74 signaling links inflammation to intestinal epithelial cell regeneration and promotes mucosal healing. *Cell. Mol. Gastroenterol. Hepatol.* **10**, 101–112 (2020).
23. Koeberle, S. C. et al. Distinct and overlapping functions of glutathione peroxidases 1 and 2 in limiting NF-κB-driven inflammation through redox-active mechanisms. *Redox Biol.* **28**, 101388 (2020).
24. Gerbe, F. et al. Intestinal epithelial tuft cells initiate type 2 mucosal immunity to helminth parasites. *Nature* **529**, 226–230 (2016).
25. Campello, R. J., Moulavi, D., Zimek, A. & Sander, J. Hierarchical density estimates for data clustering, visualization, and outlier detection. *ACM Trans. Knowl. Discov. Data* **10**, 1–51 (2015).
26. ENCODE Project Consortium. The ENCODE (Encyclopedia of DNA Elements) Project. *Science* **306**, 636–640 (2004).
27. ENCODE Project Consortium. A user's guide to the Encyclopedia of DNA Elements (ENCODE). *PLoS Biol.* **9**, e1001046 (2011).
28. Rouillard, A. D. et al. The Harmonizome: a collection of processed datasets gathered to serve and mine knowledge about genes and proteins. *Database* **2016**, baw100 (2016).
29. Frangieh, C. J. et al. Multimodal pooled Perturb-CITE-seq screens in patient models define mechanisms of cancer immune evasion. *Nat. Genet.* **53**, 332–341 (2021).
30. Papalexi, E. et al. Characterizing the molecular regulation of inhibitory immune checkpoints with multimodal single-cell screens. *Nat. Genet.* **53**, 322–331 (2021).
31. Gayoso, A. et al. Joint probabilistic modeling of single-cell multi-omic data with totalVI. *Nat. Methods* **18**, 272–282 (2021).
32. Chanput, W., Mes, J. J. & Wichers, H. J. THP-1 cell line: an in vitro cell model for immune modulation approach. *Int. Immunopharmacol.* **23**, 37–45 (2014).
33. Bhat, M. Y. et al. Comprehensive network map of interferon γ signaling. *J. Cell Commun. Signal.* **12**, 745–751 (2018).
34. Garcia-Diaz, A. et al. Interferon receptor signaling pathways regulating PD-L1 and PD-L2 expression. *Cell Rep.* **19**, 1189–1201 (2017).
35. Crabbé, J. & van der Schaar, M. Label-free explainability for unsupervised models. In *International Conference on Machine Learning* 4391–4420 (PMLR, 2022).
36. Lin, C., Chen, H., Kim, C. & Lee, S.-I. Contrastive corpus attribution for explaining representations. In *11th Int. Conf. Learn. Rep. (ICLR 2023)*.
37. Ashuach, T., Reidenbach, D. A., Gayoso, A. & Yosef, N. PeakVI: a deep generative model for single-cell chromatin accessibility analysis. *Cell Rep. Methods* **2**, 100182 (2022).
38. Gut, G., Stark, S. G., Rätsch, G. & Davidson, N. R. pmVAE: learning interpretable single-cell representations with pathway modules. Preprint at *bioRxiv* https://doi.org/10.1101/2021.01.28.428664 (2021).
39. Rybakov, S., Lotfollahi, M., Theis, F. J. & Wolf, F. A. Learning interpretable latent autoencoder representations with annotations of feature sets. Preprint at *bioRxiv* https://doi.org/10.1101/2020.12.02.401182 (2020).

Methods

The contrastiveVI model

Here, we present the contrastiveVI model in more detail. We begin by describing the model's generative process and then the model's inference procedure.

The contrastiveVI generative process. For a target data point x_n, we assume that each expression value x_{ng} for sample n and gene g is generated through the following process:

$$\begin{aligned} z_n &\sim \text{Normal}(0, I) \\ t_n &\sim \text{Normal}(0, I) \\ \ell_n &\sim \text{LogNormal}(\ell_\mu^T s_n, (\ell_\sigma^2)^T s_n) \\ \rho_n &= f_w(z_n, t_n, s_n) \\ w_{ng} &\sim \text{Gamma}(\rho_{ng}, \theta_g) \\ y_{ng} &\sim \text{Poisson}(\ell_n w_{ng}) \\ h_{ng} &\sim \text{Bernoulli}\left(f_h^g(z_n, t_n, s_n)\right) \\ x_{ng} &= \begin{cases} y_{ng} & \text{if } h_{ng} = 0 \\ 0 & \text{otherwise} \end{cases} \end{aligned}$$

In this process, z_n and t_n refer to the two sets of latent variables underlying variations in scRNA-seq expression data. Here, z_n represents variables that are shared across background and target cells, while t_n represents variations unique to target cells, y_{ng} represents RNA expression counts for a gene-cell pair before a potential dropout event, h_{ng} represents the occurence of a dropout event, and w_{ng} is a random variable controlling the rate parameter of the Poisson distribution that generates RNA expression counts for a given gene-cell pair. We place a standard multivariate Gaussian prior on both sets of latent factors, as this specification is computationally convenient for inference in the VAE framework[16]. To encourage deconvolution of shared and target-specific latent factors, for background data points b_n, we assume the same generative process but instead assume that the salient latent factors t_n are drawn from a Dirac delta distribution fixed at zero to represent the absence of salient latent factors in the generative process. Categorical covariates, such as experimental batches, are represented by s_n.

Here, ℓ_μ and $\ell_\sigma^2 \in \mathbb{R}_+^B$, where B denotes the cardinality of the categorical covariate, parameterize the prior for the latent RNA library size scaling factor on a log scale, and s_n is a B-dimensional one-hot vector encoding a categorical covariate index. For each category (for example, experimental batch), ℓ_μ and ℓ_σ^2 are set to the empirical mean and variance of the log library size, respectively. The gamma distribution is parameterized by the mean $\rho_{ng} \in \mathbb{R}_+$ and shape $\theta_g \in \mathbb{R}_+$. Furthermore, following the generative process, θ_g is equivalent to a gene-specific inverse dispersion parameter for a negative binomial distribution, and $\theta \in \mathbb{R}_+^G$ is estimated via variational Bayesian inference. f_w and f_g in the generative process are neural networks that transform the latent space and batch annotations to the original gene space, that is, $\mathbb{R}^d \times \{0,1\}^B \to \mathbb{R}^G$, where d is the size of the concatenated salient and shared latent spaces. The network f_w is constrained during inference to encode the mean proportion of transcripts expressed across all genes using a softmax activation function in the last layer. That is, letting $f_w^g(z_n, t_n, s_n)$ denote the entry in the output of f_w corresponding to gene g, we have $\sum_g f_w^g(z_n, t_n, s_n) = 1$. The neural network f_h encodes whether a particular gene's expression has dropped out in a cell due to technical factors.

Our generative process closely follows that of scVI[12], with the addition of the salient latent factors t_n. While scVI's modeling approach has been shown to excel at many scRNA-seq analysis tasks, our empirical results demonstrate that it is not suited for CA. By dividing the RNA latent factors into shared factors z_n and target-specific factors t_n, contrastiveVI successfully isolates variations enriched in target datasets that were missed by previous methods. We depict the full contrastiveVI generative process as a graphical model in Supplementary Fig. 12.

Inference with contrastiveVI. We cannot compute the contrastiveVI posterior distribution using Bayes' rule because the integrals required to compute the model evidence $p(x_n|s_n)$ are analytically intractable. As such, we instead approximate our posterior distribution using variational inference[40]. For target data points, we approximate our posterior with a distribution q factorized as follows:

$$q_{\phi_x}(z_n, t_n, \ell_n | x_n, s_n) = q_{\phi_z}(z_n | x_n, s_n) q_{\phi_t}(t_n | x_n, s_n) q_{\phi_\ell}(\ell_n | x_n, s_n). \quad (1)$$

Here, ϕ_x denotes a set of learned weights used to infer the parameters of our approximate posterior. Based on our factorization, we can divide ϕ_x into three disjoint sets ϕ_z, ϕ_t and ϕ_ℓ for inferring the parameters of the distributions of z, t and ℓ, respectively. Following the VAE framework[16], we then approximate the posterior for each factor as a deep neural network that takes as input expression levels and outputs the parameters of its corresponding approximate posterior distribution (for example, mean and variance). Moreover, we note that each factor in the posterior approximation shares the same family as its respective prior distribution (for example, $q(z_n|x_n, s_n)$ follows a normal distribution). We can simplify our likelihood by integrating out w_{ng}, h_{ng} and y_{ng}, yielding $p_v(x_{ng}|z_n, t_n, s_n, \ell_n)$, which follows a zero-inflated negative binomial distribution (Supplementary Note 11), and where v denotes the parameters of our generative model. As with our approximate posteriors, we realize our generative model with deep neural networks. For equation (1), we can derive (Supplementary Note 12) a corresponding variational lower bound, where D_{KL} is the Kullback-Leibler divergence term:

$$\begin{aligned} p(x|s) \geq\ & \mathbb{E}_{q(z,t,\ell|x,s)} \log p(x|z,t,\ell,s) - D_{\text{KL}}(q(z|x,s)\,||\,p(z)) \\ & - D_{\text{KL}}(q(t|x,s)\,||\,p(t)) - D_{\text{KL}}(q(\ell|x,s)\,||\,p(\ell|s)). \end{aligned} \quad (2)$$

Similarly, for background data points, we approximate the posterior using the factorization

$$q_{\phi_b}(z_n, t_n, \ell_n | b_n, s_n) = q_{\phi_z}(z_n | b_n, s_n) q_{\phi_t}(t_n | b_n, s_n) q_{\phi_\ell}(\ell_n | b_n, s_n), \quad (3)$$

where ϕ_b denotes a set of learned parameters used to infer the values of z_n and ℓ_n for background samples. Following our factorization, we divide ϕ_b into the disjoint sets ϕ_z, ϕ_t and ϕ_ℓ. Once again, we can simplify our likelihood by integrating out w_{ng}, h_{ng} and y_{ng} to obtain $p_v(b_{ng}|z_n, \mathbf{0}, s_n, \ell_n)$, which follows a zero-inflated negative binomial distribution. We then have the following variational lower bound for our background data points:

$$\begin{aligned} p(b|s) \geq\ & \mathbb{E}_{q(z,t,\ell|b,s)} \log p(b|z,\mathbf{0},\ell,s) - D_{\text{KL}}(q(z|b,s)\,||\,p(z)) \\ & - D_{\text{KL}}(q(t|b,s)\,||\,p(t)) - D_{\text{KL}}(q(\ell|b,s)\,||\,p(\ell|s)). \end{aligned} \quad (4)$$

As specified in our generative model, the prior distribution $p(t)$ for background points is a Dirac delta centered at zero. This constraint enforces the idea that our salient latent factors should capture target cell-specific variations and be uninformative for background cells. Yet, with this constraint, the KL divergence term $D_{\text{KL}}(q(t|b,s)||p(t))$ in equation (4) is not defined, as a Gaussian distribution does not admit a density with respect to a counting measure. To obtain a tractable objective function, previously proposed CA models[8,10] have ignored this term during optimization. However, not explicitly enforcing this constraint may result in t undesirably capturing variations shared with control cells. To work around this issue while still enforcing the desired constraint, we replace our degenerate KL divergence term with a regularization penalty based on the squared Wasserstein distance[41,42].

The Wasserstein distance between a Gaussian random variable and a Dirac distribution has a closed-form solution:

$$W_2^2\left(q\left(t|b,s\right),\,\delta\{t=\mathbf{0}\}\right)=\left\|\mu_t\left(b,s\right)\right\|^2+\left\|\sigma_t\left(b,s\right)\right\|^2, \tag{5}$$

where $\mu_t(b, s)$ and $\sigma_t(b, s)$ denote the mean and standard deviations of the approximate posterior $q(t|b, s)$, and δ is a Dirac distribution centered at $\mathbf{0}$. Substituting this expression for the degenerate KL term in equation (4) yields a new lower bound for background points:

$$\begin{aligned}p(b|s)\geq\ &\mathbb{E}_{q(z,t,\ell|b,s)}\log p(b|z,\ell,s)-D_{\mathrm{KL}}(q(z|b,s)\,||\,p(z))\\&-||\mu_t(b,s)||^2-||\sigma_t(b,s)||^2-D_{\mathrm{KL}}(q(\ell|b,s)\,||\,p(\ell|s)).\end{aligned} \tag{6}$$

We then jointly optimize the parameters of our generative model and inference networks using stochastic gradient descent to maximize the sum of our final bounds for background and target data points. All neural networks used to implement the variational and generative distributions were feedforward and used standard activation functions. We used the same network architecture and hyperparameter values for all experiments, and we refer the reader to Supplementary Note 13 for more details.

We now discuss in more depth how this optimization procedure achieves our specific goal of deconvolving the shared latent factors z and target-specific latent factors t. We begin by considering the shared factors z. First, we note that the latent representations of all cells (that is, both target and background) are decoded using the same decoder network with shared parameters v. When decoding background cells, we only provide the shared latent variable values z to the decoder and hardcode the value of the salient latent variables t to $\mathbf{0}$. By doing so, the shared variables are forced to faithfully capture variations present in background cells, as otherwise the decoder would not be able to accurately reconstruct these cells. Moreover, because the encoder parameters ϕ_z are shared across target and background cells, z will also naturally reflect the presence of these same variations in target cells.

Next, we consider the target cell-specific latent factors t. Because these variables are only used to reconstruct target cells, they are thus implicitly encouraged to capture any remaining target cell-specific variations not already captured by the shared variables z. In addition, to further encourage these variables to reflect only target cell-specific variations (and not also variations shared with control cells), we impose different prior distributions on this variable for target versus background cells. In particular, for target cells, our prior distribution is a relatively flexible isotropic Gaussian distribution, while, for background cells, we impose a much stricter Dirac delta prior centered around $\mathbf{0}$. This stricter prior penalizes our salient variable encoder for mapping background cells to anything beyond an uninformative zero vector. As a result, ϕ_t is further encouraged during optimization to only capture target cell-specific variations as opposed to variations that are shared across target and background cells. In experiments on simulated data with known ground truths (Supplementary Note 3), we found that enforcing this delta prior distribution led to better deconvolution of shared versus target-specific latent factors than what was obtained when ignoring this constraint.

Denoising gene expression values with contrastiveVI. For a given target cell x_n, contrastiveVI can be used to produce a denoised expression profile $\tilde{x}_n$ by inferring x_n's salient and shared latent variables z_n and t_n and then decoding this latent representation back to the full gene expression space. For background cells, the same procedure can be applied but with the salient variables t_n fixed at $\mathbf{0}$. As done in ref. 31, when visualizing these denoised expression values, we applied a log library size transformation. That is, for a given denoised expression value $\tilde{x}_{ng}$ for a gene g for cell n, we computed the value

$$\tilde{x}'_{ng}=\log_e\left(L\times\frac{x_{ng}}{\sum_g x_{ng}}+1\right),$$

where L is a scaling factor. For the results reported in this study, L was set to the median of total RNA counts across cells (the default option in the Scanpy[43] 'normalize_total' function).

Differential gene expression analysis with contrastiveVI. Similar to scVI, contrastiveVI's probabilistic representation of the data admits methods for differential expression testing between two sets of cells. Such tests can be constructed to detect the presence of a differential expression effect without regard to effect size (referred to as the 'vanilla' differential expression test in the scvi-tools[44] package) or to detect a differential expression effect greater than some pre-specified effect size δ (referred to as the 'change' differential expression test in the scvi-tools[44] package). To remove the influence of the effect size parameter δ on the results reported in this study, we used the 'vanilla' test in our experiments. However, for completeness, both tests have been implemented in our Python package, and we describe both tests below.

We begin by describing the 'vanilla' test. For a given gene g and pair of target cells (a, b) with shared latent representations (z_a, z_b), salient latent representations (t_a, t_b), observed gene expression (x_a, x_b) and batch IDs (s_a, s_b), we can formulate two mutually exclusive hypotheses:

$$\begin{aligned}&\mathcal{H}_1^g:=\mathbb{E}_s f_w^g(z_a,t_a,s)>\mathbb{E}_s f_w^g(z_b,t_b,s)\\&\text{versus}\quad\mathcal{H}_2^g:=\mathbb{E}_s f_w^g(z_a,t_a,s)\leq\mathbb{E}_s f_w^g(z_b,t_b,s),\end{aligned}$$

where the expectation $\mathbb{E}_s$ is assessed using empirical frequencies. Evaluating which of these two hypotheses is more likely is equivalent to computing a Bayes factor. The sign of this factor indicates which hypothesis is more likely, and its magnitude indicates a significance level. As done in ref. 12, we consider a Bayes factor K to indicate a significant result if $|K| > 3$, where K is defined as

$$K=\log_e\frac{p(\mathcal{H}_1^g|x_a,x_b)}{p(\mathcal{H}_2^g|x_a,x_b)},$$

and the posterior of these models can be approximated by the variational distribution

$$\begin{aligned}&p\left(\mathcal{H}_1^g\,|x_a,x_b\right)\\&\approx\sum_s\int_{z_a,t_a,z_b,t_b}p\left(f_w^g\left(z_b,t_b,s\right)<f_w^g\left(z_a,t_a,s\right)\right)p\left(s\right)dq\left(z_a,t_a\,|x_a\right)dq\left(z_b,t_b\,|x_b\right),\end{aligned}$$

where $p(s)$ denotes the relative abundance of cells in each batch s, and $dq(\cdot)$ indicates that we are integrating over the distribution q. Here, all our measures are low dimensional; therefore, the integrals can be computed with Monte Carlo sampling. All cells are assumed to be independent; therefore, we can average the Bayes factors across a large set of randomly sampled cell pairs, where one cell in a pair is from each population. The average factor then describes whether cells from one population express g at a higher level. This procedure, with a minor modification, can also be used to test for differentially expressed genes with background cells. For these cells, the salient latent variable values are fixed at $\mathbf{0}$; otherwise, the test is conducted as described previously. For all results reported in this study, 10,000 cell pairs were sampled, and 100 Monte Carlo samples were obtained from the variational posteriors for each cell.

We now describe the effect size-based test (that is, the 'change' test). For two cell groups $A = (a_1, a_2, \ldots, a_n)$ and $B = (b_1, b_2, \ldots, b_m)$ in the target dataset, the posterior probability of gene g being differentially expressed in the two groups can be obtained as proposed by

Boyeau et al.[45]. For any arbitrary cell pair a_i, b_j, we have two mutually exclusive models:

$$\mathcal{M}_1^g : \left|r^g_{a_i,b_j}\right| > \delta \text{ and } \mathcal{M}_0^g : \left|r^g_{a_i,b_j}\right| \leq \delta,$$

where $r^g_{a_i,b_j} = \log_2(\rho^g_{a_i}) - \log_2(\rho^g_{b_j})$ is the log (fold change) of the denoised, library size-normalized expression of gene g, and δ is a predefined threshold for log (fold change) magnitude to be considered biologically meaningful. The posterior probability of differential expression is therefore expressed as $p(\mathcal{M}_1^g | x_{a_i}, x_{b_j})$, which can be obtained by marginalization of the latent variables and categorical covariates:

$$p\left(\mathcal{M}_1^g | x_{a_i}, x_{b_j}\right) = \sum_s \int_{z_{a_i}, t_{a_i}, z_{b_j}, t_{b_j}} p\left(\mathcal{M}_1^g | z_{a_i}, t_{a_i}, z_{b_j}, t_{b_j}\right) p(s)\, dp\left(z_{a_i}, t_{a_i} | x_{a_i}, s\right) dp\left(z_{b_j}, t_{b_j} | x_{b_j}, s\right),$$

where $p(s)$ is the relative abundance of target cells in category s, and the integral can be computed via Monte Carlo sampling using the variational posteriors q_{ϕ_z}, q_{ϕ_t}. Finally, the group-level posterior probability of differential expression is

$$\int_{a,b} p\left(\mathcal{M}_1^g | x_a, x_b\right) dp(a) dp(b),$$

where we assume that cells a and b are independently sampled $a \sim \mathcal{U}(a_1, \ldots, a_m)$ and $b \sim \mathcal{U}(b_1, \ldots, b_m)$, respectively, where $\mathcal{U}$ denotes a uniform distribution. Computationally, this quantity can be estimated by a large random sample of pairs from cell groups A and B.

This procedure, with a minor modification, can also be used to test for differentially expressed genes between a group of target cells and a group of background cells. Without loss of generality, let A denote a group of cells in the target dataset and B denote a group of cells in the background dataset. When computing the integral in the expression for $p\left(\mathcal{M}_1^g | x_{a_i}, x_{b_j}\right)$, the values of t_{b_j} are fixed at $\mathbf{0}$ to represent their absence in the generative process for background cells. The test then proceeds as previously described for the case of two groups of target cells.

The totalContrastiveVI model

We now present the totalContrastiveVI model in more detail.

The totalContrastiveVI generative process. Formally, for a given cell n, we have gene expression values x_{ng} for each measured gene g and protein expression values $y_{n\tau}$ for each measured protein τ. For gene expression values, we assume the generative process described previously for contrastiveVI.

To account for the technical biases of CITE-seq-based platforms, totalContrastiveVI models protein counts as a mixture of foreground and background components. For target cells, the full generative process for protein measurements is as follows:

$$\begin{aligned}
z_n &\sim \text{Normal}(0, I) \\
t_n &\sim \text{Normal}(0, I) \\
\beta_{n\tau} &\sim \text{LogNormal}(c_\tau^T s_n, (d_\tau^2)^T s_n) \\
\pi_n &= h_\pi(z_n, t_n, s_n) \\
\alpha_n &= g_\alpha(z_n, t_n, s_n) \\
\upsilon_{n\tau} | z_n, s_n &\sim \text{Bernoulli}(\pi_{n\tau}) \\
r_{n\tau} | \upsilon_{n\tau}, \beta_{n\tau}, z_n, t_n, s_n &\sim \text{Gamma}(\phi_\tau, \upsilon_{n\tau}\beta_{n\tau} + (1 - \upsilon_{n\tau})\beta_{n\tau}\alpha_{n\tau}) \\
y_{n\tau} | r_{n\tau} &\sim \text{Poisson}(r_{n\tau})
\end{aligned}$$

Here, $\beta_{n\tau}$ is a protein-specific variable representing a protein-specific background intensity. The parameters $c_\tau \in \mathbb{R}^B$ and $d_\tau^2 \in \mathbb{R}_+^B$ for the prior distribution of $\beta_{n\tau}$ are protein specific and are treated as model parameters to be learned during training. $\upsilon_{n\tau}$ controls whether a given protein's counts are generated from the background or foreground mixture component, with its parameter $\pi_{n\tau}$ being the output of the neural network h_π and representing the probability of the counts being generated due to background alone. g_α is constrained such that its output $\alpha_{n\tau}$ always exceeds 1. Thus, one of the mixture components will always be larger than the other, enabling one to be interpreted as foreground and the other as background. For a given mixture component, $y_{n\tau}|z_n, t_n, s_n, \beta_{n\tau}$ follows a negative binomial distribution, which can be shown by integrating out the rate parameter $r_{n\tau}$ of the final Poisson distribution. Moreover, $y_{n\tau}$ given z_n, t_n and s_n can be shown to follow a negative binomial distribution by integrating out $\upsilon_{n\tau}$, with ϕ_τ acting as a protein-specific inverse dispersion parameter. For background data points, we assume the same generative process but set $t_n = \mathbf{0}$ to represent the absence of salient latent factors.

The generative process of totalContrastiveVI closely follows that of totalVI[31] but with the addition of salient latent factors. We depict the totalContrastiveVI generative process as a graphical model in Supplementary Fig. 14.

Inference with totalContrastiveVI. As with contrastiveVI, for totalContrastiveVI, we approximate our posterior distribution using variational inference. For target data points, we use an approximate posterior factorized as follows:

$$\begin{aligned}
&q_{\phi_{\text{target}}}(z_n, t_n, \ell_n, \beta_n | x_n, y_n, s_n) \\
&= q_{\phi_\beta}(\beta_n | y_n, s_n)\, q_{\phi_z}(z_n | x_n, y_n, s_n)\, q_{\phi_t}(t_n | x_n, y_n, s_n)\, q_{\phi_\ell}(\ell_n | x_n, s_n),
\end{aligned} \tag{7}$$

where ϕ_{target} denotes a set of learned weights for our approximate posterior distribution. We can simplify the gene and protein likelihood as described previously to obtain $p_v(x_{ng} | z_n, t_n, \ell_n, s_n)$, which is a zero-inflated negative binomial distribution, and $p_v(y_{nt} | z_n, t_n, \beta_n, s_n)$, which is a mixture of negative binomials.

For background points, we have the following approximate posterior distribution:

$$\begin{aligned}
&q_{\phi_{\text{background}}}(z_n, t_n, \ell_n, \beta_n | x_n, y_n, s_n) \\
&= q_{\phi_\beta}(\beta_n | y_n, s_n)\, q_{\phi_z}(z_n | x_n, y_n, s_n)\, q_{\phi_t}(t_n | x_n, y_n, s_n)\, q_{\phi_\ell}(\ell_n | x_n, s_n).
\end{aligned} \tag{8}$$

We then jointly optimize the parameters of our generative model and inference networks using stochastic gradient descent to maximize the sum of the ELBOs over our background and target data points. We note that, similar to contrastiveVI, our prior distribution for the salient latent factors t is an isotropic Gaussian for target points and a Dirac delta centered at $\mathbf{0}$ for background points. Moreover, to encourage the variational posterior for t for background points to be close to the Dirac delta prior, we use the Wasserstein distance penalty defined in equation (5) in place of the intractable KL divergence in the ELBO as done for contrastiveVI. All neural networks used to implement the variational and generative distributions were feedforward and used standard activation functions. We used the same network architecture and hyperparameter values for all experiments, and we refer the reader to Supplementary Note 14 for more details.

Differential expression analysis with totalContrastiveVI. Similar to totalVI[31], totalContrastiveVI can be used to compute differentially expressed genes and proteins. To compute differentially expressed genes, we use the same procedures as in contrastiveVI. For proteins, we use a similar framework that can detect differential expression without regard to effect size ('vanilla') or detect effects greater than a pre-specified threshold δ. As for contrastiveVI, for all totalContrastiveVI differential expression results presented in this study, we used

'vanilla' differential expression tests to remove the impact of the choice of δ on our results.

We begin by describing the 'vanilla' test. For this test, we use a similar setup as described for the 'vanilla' RNA differential expression test for contrastiveVI. Let $C_n = \{x_n, y_n, s_n\}$ denote the observed data for a cell n. For a given protein τ and pair of target cells (a, b) with shared latent representations (z_a, z_b), salient latent representations (t_a, t_b) and observed data (C_a, C_b), we can formulate two mutually exclusive hypotheses:

$$\mathcal{H}_1^\tau := \mathbb{E}\left[\tilde{r}_{a_\tau} | \beta_{a_\tau}, v_{a_\tau}, z_a, t_a\right] > \mathbb{E}\left[\tilde{r}_{b_\tau} | \beta_{b_\tau}, v_{b_\tau}, z_b, t_b\right] \text{ versus } \mathcal{H}_2^\tau : \\ = \mathbb{E}\left[\tilde{r}_{a_\tau} | \beta_{a_\tau}, v_{a_\tau}, z_a, t_a\right] \leq \mathbb{E}\left[\tilde{r}_{b_\tau} | \beta_{b_\tau}, v_{b_\tau}, z_b, t_b\right],$$

where our conditional expectation is

$$\mathbb{E}\left[\tilde{r}_{a_i\tau} | \beta_{a_i\tau}, v_{a_i\tau}, z_{a_i}, t_{a_i}\right] = \beta_{a_i\tau}\alpha_{a_i\tau}(1 - v_{a_i\tau}).$$

As done in the contrastiveVI RNA differential expression test, we can then compute a Bayes factor to evaluate which of these two hypotheses is more likely. As done in ref. 31, we consider a log Bayes factor K to indicate a significant result if $|K| > 0.7$, where K is defined as

$$K = \log_e \frac{p(\mathcal{H}_1^\tau | C_a, C_b)}{p(\mathcal{H}_2^\tau | C_a, C_b)},$$

and the posterior of these models can be approximated by the variational distribution as done for the contrastiveVI RNA differential expression test. Moreover, as done for the contrastiveVI RNA differential expression test, $p(\mathcal{H}_1^\tau | C_a, C_b)$ and $p(\mathcal{H}_2^\tau | C_a, C_b)$ can be integrated over the batch variable s_n to compare cells across batches.

Now, we describe the 'change' test. For two groups of cells $A = (a_1, a_2, \ldots, a_n)$ and $B = (b_1, b_2, \ldots, b_m)$ in the target dataset, totalContrastiveVI can be used to compute both differentially expressed genes and differentially expressed proteins. For proteins, we use the same framework, with the following two mutually exclusive hypotheses given two cells a_i and b_j:

$$\mathcal{M}_1^\tau : \left|\gamma_{a_i,b_j}^\tau\right| > \delta \text{ and } \mathcal{M}_0^\tau : \left|\gamma_{a_i,b_j}^\tau\right| \leq \delta,$$

where

$$\gamma_{a_i,b_j}^\tau := \log_2\left(\mathbb{E}\left[\tilde{r}_{a_i\tau} | \beta_{a_i\tau}, v_{a_i\tau}, z_{a_i}, t_{a_i}\right] + \epsilon\right) - \log_2\left(\mathbb{E}\left[\tilde{r}_{b_j\tau} | \beta_{b_j\tau}, v_{b_j\tau}, z_{b_j}, t_{b_j}\right] + \epsilon\right)$$

and our conditional expectation is

$$\mathbb{E}\left[\tilde{r}_{a_i\tau} | \beta_{a_i\tau}, v_{a_i\tau}, z_{a_i}, t_{a_i}\right] = \beta_{a_i\tau}\alpha_{a_i\tau}(1 - v_{a_i\tau}).$$

Based on our previous definition of v, this expression can be interpreted as the foreground mean if the cell was generated from the foreground and zero otherwise. The value ϵ serves as a 'prior count' that ensures that the log (fold change) is defined even when the conditional expectation is zero. We set a default value of $\epsilon = 0.5$ as was done in ref. 31. As described previously, let $C_n = \{x_n, y_n, s_n\}$ denote the observed data for a cell n. The posterior probability of differential expression is therefore expressed as $p(\mathcal{M}_1^\tau | C_{a_i}, C_{b_j})$, which can be obtained by integrating with respect to the distribution

$$\prod_{k \in \{a_i, b_j\}} p(v_{k\tau} | z_k, t_k)\, q(\beta_{k\tau} | z_k, t_k, s_k)\, q(z_k, t_k | C_k).$$

The group-level posterior probability of differential expression can then be estimated by sampling, as previously described for the contrastiveVI gene differential expression test.

Pathway enrichment analysis

Pathway enrichment analysis refers to a computational procedure for determining whether a predefined set of genes (that is, a gene pathway) has statistically significant differences in expression between two biological states. Many tools exist for performing pathway enrichment analysis (see ref. 46 for a review). Our analyses used Enrichr[47], a pathway analysis tool for non-ranked gene lists based on Fisher's exact test, to find enriched pathways from the Reactome 2016 pathway database[48]. Specifically, the Enrichr wrapper implemented in the open-source GSEAPY (https://gseapy.readthedocs.io/en/latest/) Python library was used for our analyses. Pathways enriched at a false discovery rate less than 0.05, adjusted by the Benjamini–Hochberg procedure[49], are reported in this study.

Baseline models

To highlight the merits of contrastiveVI, we compared it to the previously proposed CA methods CLVM, CPLVM and CGLVM. For all these baseline methods, variations shared between the background and target conditions are assumed to be captured by the shared latent variable values $\{z_i^b\}_{i=1}^n$ and $\{z_j^t\}_{j=1}^m$, and target condition-specific variations are captured by the salient latent variable values $\{t_j\}_{j=1}^m$, where n, m are the number of background and target cells, respectively. The CLVM model is trained with a Gaussian likelihood function; therefore, we applied it to log library size-normalized scRNA-seq data. Specifically, each data point is assumed to follow a Gaussian distribution with unit variance and mean given by $S^\top z_i^b + \mu^b$ for a background cell and $S^\top z_j^t + W^\top t_j + \mu^b$ for a target cell, where S, W are model weights that linearly combine the latent variables, and $\mu^b, \mu^t \in \mathbb{R}^G$ are dataset-specific means with G denoting the number of genes. Posterior distributions are fitted using variational inference with mean-field approximation and log-normal variational distributions.

CPVLM and CGLVM instead operate on unnormalized count data. Library size differences between the target and background conditions are modeled by $\{\alpha_i^b\}_{i=1}^n$ and $\{\alpha_j^t\}_{j=1}^m$, whereas gene-specific library sizes are parameterized by $\delta \in \mathbb{R}_+^G$, where G is the number of genes. Each data point is considered Poisson distributed, with the rate parameter determined by $\alpha_i^b \delta \odot (S^\top z_i^b)$ for a background cell i and by $\alpha_j^t \delta \odot (S^\top z_j^t + W^\top t_j)$ for a target cell j, where S, W are model weights that linearly combine the latent variables, and $\odot$ represents an element-wise product. The model weights and latent variables are assumed to have gamma priors, δ has a standard log-normal prior, and α_i^b, α_j^t have log-normal priors with parameters given by the empirical mean and variance of log total counts in each dataset. The CA modeling approaches of CGLVM and CPLVM are similar. In CGLVM, however, the relationships of latent factors are considered additive and relate to the Poisson rate parameter via an exponential link function (similar to a generalized linear modeling scheme). All the priors and variational distributions are Gaussian in CGLVM. As with CLVM, posterior distributions are fitted using variational inference with mean-field approximation and log-normal variational distributions. We present a summary of additional previous work on CA in Supplementary Table 9.

Beyond these CA method baselines, to illustrate the need for models specifically designed for CA, we also consider scVI, a deep generative model for scRNA-seq count data that takes batch effect, technical dropout and varying library size into modeling consideration[12] as well as DCA, an autoencoder neural network for reducing noise in scRNA-seq count data due to technical dropout[19]. We also compare against a typical scRNA-seq analysis workflow in which PCA is applied to library size-normalized, log-transformed counts.

Model optimization details

For all datasets, contrastiveVI or totalContrastiveVI models were trained with 80% of the background and target data; the remaining 20% was reserved as a validation set for early stopping to determine

the number of training epochs needed. Training was stopped early when the validation variational lower bound showed no improvement for 45 epochs, typically resulting in 127–500 epochs of training. All contrastiveVI and totalContrastiveVI models were trained with the Adam optimizer[50], with $\varepsilon = 0.01$, the learning rate at 0.001 and weight decay at 10^{-6}. The same hyperparameters and training scheme were used to optimize the scVI models using only target data, usually with 274–500 epochs of training based on the early stopping criterion. As in the open-source implementation by Eraslan et al., DCA models were trained for a maximum of 500 epochs using the RMSprop optimizer with a learning rate at 0.001 and with early stopping when the validation loss showed no improvement for 15 epochs[19]. As in Jones et al., the CPLVMs were trained via variational inference using all background and target data for 2,000 epochs with the Adam optimizer with $\varepsilon = 10^{-8}$ and a learning rate at 0.05, and the CGLVMs were similarly trained for 1,000 epochs with a learning rate at 0.01 (ref. 8). Finally, as in ref. 10, the CLVMs were trained for 10,000 epochs with the Adam optimizer with $\varepsilon = 10^{-8}$ and a learning rate at 0.01. All models were trained with ten salient and ten shared latent variables five times with different random weight initializations.

Datasets and preprocessing

We now briefly describe all datasets used in this work along with any corresponding preprocessing steps. For our experiments, datasets were chosen that not only had cells in a target and corresponding background condition but that also had ground truth subclasses of target cells. Moreover, to avoid potential confounding effects, datasets collected using a variety of single-cell platforms (Supplementary Table 10) were used in our experiments. All preprocessing steps were performed using the Scanpy Python package[43], and all our code for downloading and preprocessing these datasets is publicly available at https://github.com/suinleelab/contrastiveVI-reproducibility. For all experiments, we retained the top 2,000 most highly variable genes returned from the Scanpy 'highly_variable_genes' function, with the 'flavor' parameter set to 'seurat_v3'. For all datasets, the number of cells in background versus target conditions can be found in Supplementary Table 10.

Kotliar et al. 2019. This dataset was generated using the simulation framework described in ref. 51 and implemented in the scsim (https://github.com/dylkot/scsim) Python package. Eleven gene programs (ten identity programs $P_1, \ldots, P_{10}$ corresponding to simulated cell types and one activity gene program P_a) were simulated as in Splatter[52]. Cells were then randomly assigned to an identity program with an equal probability for each class. Thirty-five percent of cells from three cell types were randomly selected to express the activity program at a usage level ϕ_i, uniformly distributed between 10% and 70%. Using Splatter's notation, the pre-trended mean gene expression profile λ'_i for each cell $i = 1, \ldots, 10{,}000$ was computed as the weighted sum of the identity and the activity program:

$$\lambda'_i = L_i \left(\phi_i P_a + (1 - \phi_i) P_{I(i)}\right),$$

where L_i denotes the simulated library size for cell i, $I(i)$ denotes the cell type identity program for cell i, and $\phi_i = 0$ for cells that do not express the activity program and $\phi_i \sim \text{Uniform}(0.1, 0.7)$ for those that do. For our experiments, we simulated 10,000 genes, 400 of which were associated with the activity program. All additional hyperparameter values for the simulation were set to those used by Kotliar et al.[51].

Cells were then divided into target and background datasets as follows. For cells types that never expressed the activity program, cells were randomly assigned to the target or background dataset. For cell types that did sometimes express the additional program, cells were assigned to the target dataset if $\phi_i > 0$ and the background dataset otherwise.

McFarland et al. 2020. This dataset has measurements of cancer cell lines' transcriptional responses after being treated with various small-molecule therapies. For our target dataset, we used data from cells that were exposed to idasanutlin, and, for our background, we used data from cells that were exposed to a control solution of DMSO. *TP53*-mutation status was determined using the DepMap[53] 19Q3 data release, available at https://depmap.org/. The count data were downloaded from the authors' Figshare repository at https://figshare.com/articles/dataset/MIX-seq_data/10298696. The number of cells for each cell line can be found in Supplementary Table 11. For our analysis, we excluded any cells that were labeled as low quality (that is, a 'cell_quality' metadata value not equal to 'normal') by McFarland et al.[2].

Haber et al. 2017. This dataset (Gene Expression Omnibus (GEO) accession number GSE92332) consists of scRNA-seq measurements used to investigate the responses of intestinal epithelial cells in mice to different pathogens. Specifically, in this dataset, responses to the bacterium *Salmonella* and the parasite *H. polygyrus* were investigated. Our target dataset included measurements of cells infected with *Salmonella* and from cells 10 d after being infected with *H. polygyrus*, while our background consisted of measurements from healthy control cells released as part of the same study. The number of cells of each cell type can be found in Supplementary Table 12.

Norman et al. 2019. This dataset (GEO accession number GSE133344) has measurements of the effects of 284 different CRISPR-mediated perturbations on K562 cells, where each perturbation induced overexpression of a single gene or a pair of genes. As done in the analysis of Norman et al.[4], we excluded cells with the perturbation label 'NegCtrl1_NegCtrl0__NegCtrl1_NegCtrl0' from our analysis. We also excluded any cells from our analysis that were marked as doublets by Norman et al.[4] (that is, a 'number_of_cells' metadata value greater than 1.0). For our background dataset, we used all remaining unperturbed cells; for our target dataset, we used all perturbed cells that had a gene program label provided by the authors.

Papalexi et al. 2021. This dataset (GEO accession number GSE153056) has measurements of the effects of 111 different CRISPR KO perturbations on THP-1 cells. The dataset contains both transcriptomic measurements and measurements of surface protein levels for the proteins CD86, PD-L1, PD-L2 and CD366. Our background dataset consists of measurements from cells infected with non-targeting gRNA species, while our target dataset consists of measurements from perturbed cells.

Evaluation metrics

Here, we describe the quantitative metrics used in this study. All metrics were computed using their corresponding implementations in the scikit-learn Python package[54]. To facilitate visual comparisons of performance of different models across multiple metrics, we produced overview tables similar to those of Lotfollahi et al.[14] and Saelens et al.[55]. In these tables, individual scores are displayed as circles and aggregated scores are shown as bars. For each individual metric, we computed the mean value for each model trained five times with different random weight initializations. These values were then minimum–maximum scaled to facilitate comparisons between metrics, and these scaled scores were then averaged into aggregated scores of salient or shared representation quality. A final overall score was then produced by averaging the aggregate salient and shared representation scores. We report the raw (that is, unscaled) mean value ± standard error for each metric in Supplementary Tables 2–7.

Average silhouette width. We calculate silhouette width using the latent representations returned by each method. For a given sample i, the silhouette width $s(i)$ is defined as follows. Let $a(i)$ be the average distance between i and other samples with the same

ground truth label, and let $b(i)$ be the smallest average distance between i and all other samples with a different label. The silhouette score $s(i)$ is then

$$s(i) = \frac{b(i) - a(i)}{\max(a(i), b(i))}.$$

A silhouette width close to 1 indicates that i is tightly clustered with cells having the same ground truth label, while a score close to −1 indicates that a cell has been grouped with cells having a different label. In our results, we report the average silhouette width (ASW).

We also used the silhouette width to measure the mixing of groups of cells (for example, the 'mutant cell line mixing silhouette' and 'treatment mixing silhouette' metrics from our analysis of the MIX-seq dataset from McFarland et al.[2]). To accomplish this, we follow the procedure described by Lotfollahi et al.[14], which consists of (1) computing the ASW to measure the separation of different groups of cells and then (2) inverting the ASW by subtracting its absolute value from 1. That is, we compute

$$\text{ASW}_{\text{mixing}} = 1 - |\text{ASW}|.$$

A higher $\text{ASW}_{\text{mixing}}$ score thus implies better mixing of the given groups of cells.

Entropy of mixing. For c groups (for example, cell types, different treatment conditions, etc.), the entropy of mixing[12,56] is defined as

$$\sum_{i=1}^{c} p_i \log p_i,$$

where p_i denotes the proportion of cells from group i in a given region, such that $\sum_{i=1}^{c} p_i = 1$. Next, let U denote a uniform random variable over the population of cells. Let B_U then denote the empirical proportions of cells' groups in the 50 nearest neighbors of cell U. We report the entropy of this variable averaged over 100 random cells U. Higher values of this metric indicate stronger mixing of the c groups.

Adjusted Rand index. The ARI measures agreement between reference clustering labels and labels assigned by a clustering algorithm. Given a set of n samples and two sets of clustering labels describing those cells, the overlap between clustering labels can be described using a contingency table, where each entry indicates the number of cells in common between the two sets of labels. Mathematically, the ARI is calculated as

$$\text{ARI} = \frac{\sum_{ij}\binom{n_{ij}}{2} - \left(\sum_i \binom{a_i}{2}\sum_j \binom{b_j}{2}\right)\binom{n}{2}^{-1}}{\frac{1}{2}\left(\sum_i \binom{a_i}{2} + \sum_j \binom{b_j}{2}\right) - \left(\sum_i \binom{a_i}{2}\sum_j \binom{b_j}{2}\right)\binom{n}{2}^{-1}},$$

where n_{ij} is the number of cells assigned to cluster i based on the reference labels and cluster j based on a clustering algorithm, a_i is the number of cells assigned to cluster i in the reference set, and b_j is the number of cells assigned to cluster j by the clustering algorithm. ARI values closer to 1 indicate stronger agreement between the reference labels and labels assigned by a clustering algorithm. In our experiments, we used the k-means clustering algorithm to assign cluster labels to cells. To reflect the fact that ground truth labels are typically not known a priori, we ran k means and computed the ARI for $k \in [\max(1, \text{true number of clusters} - 3), \text{true number of clusters} + 3]$, and we reported the maximum of these ARI scores.

Normalized mutual information. The NMI measures the agreement between reference clustering labels and labels assigned by a clustering algorithm. The NMI is calculated as

$$\text{NMI} = \frac{I(P;T)}{\sqrt{\mathbb{H}(P)\mathbb{H}(T)}},$$

where P and T denote empirical distributions for the predicted and true clusterings, I denotes mutual information, and $\mathbb{H}$ is the Shannon entropy. To reflect the fact that ground truth labels are typically not known a priori, we ran k means and computed the NMI for $k \in [\max(1, \text{true number of clusters} - 3), \text{true number of clusters} + 3]$, and we reported the maximum of these NMI scores.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

All datasets analyzed in this paper are publicly available. The simulated dataset was generated using the scsim package found at https://github.com/dylkot/scsim. The MIX-seq dataset from McFarland et al.[2] was downloaded from the authors' Figshare repository (https://figshare.com/articles/dataset/MIX-seq_data/10298696). The Haber et al.[20], Norman et al.[4] and Papalexi et al.[30] datasets were downloaded from the National Institutes of Health GEO (accession codes GSE92332, GSE133344 and GSE153056, respectively). Our code for downloading and preprocessing these datasets is available at https://github.com/suinleelab/contrastiveVI-reproducibility.

Code availability

Our Python software package with scvi-tools[44] implementations of the contrastiveVI and totalContrastiveVI models is available at https://github.com/suinleelab/contrastiveVI. Code for reproducing the specific results in this paper is available at https://github.com/suinleelab/contrastiveVI-reproducibility.

References

40. Blei, D. M., Kucukelbir, A. & McAuliffe, J. D. Variational inference: a review for statisticians. *J. Am. Stat. Assoc.* **112**, 859–877 (2017).
41. Villani, C. *Optimal Transport: Old and New*, Vol. 338 (Springer, 2009).
42. Weinberger, E., Lopez, R., Hutter, J.-C. & Regev, A. Disentangling shared and group-specific variations in single-cell transcriptomics data with multiGroupVI. Preprint at *bioRxiv* https://doi.org/10.1101/2022.12.13.520349 (2022).
43. Wolf, F. A., Angerer, P. & Theis, F. J. Scanpy: large-scale single-cell gene expression data analysis. *Genome Biol.* **19**, 15 (2018).
44. Gayoso, A. et al. A Python library for probabilistic analysis of single-cell omics data. *Nat. Biotechnol.* **40**, 163–166 (2022).
45. Boyeau, P. et al. Deep generative models for detecting differential expression in single cells. In *Machine Learning in Computational Biology* (MLCB, 2019).
46. Khatri, P., Sirota, M. & Butte, A. J. Ten years of pathway analysis: current approaches and outstanding challenges. *PLoS Comput. Biol.* **8**, e1002375 (2012).
47. Chen, E. Y. et al. Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool. *BMC Bioinformatics* **14**, 128 (2013).
48. Fabregat, A. et al. The reactome pathway knowledgebase. *Nucleic Acids Res.* **46**, D649–D655 (2018).
49. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. B Methodol.* **57**, 289–300 (1995).

50. Kingma, D. P. & Ba, J. Adam: a method for stochastic optimization. In *3rd International Conference on Learning Representations* (ICLR, 2015).
51. Kotliar, D. et al. Identifying gene expression programs of cell-type identity and cellular activity with single-cell RNA-seq. *eLife* **8**, e43803 (2019).
52. Zappia, L., Phipson, B. & Oshlack, A. Splatter: simulation of single-cell RNA sequencing data. *Genome Biol.* **18**, 174 (2017).
53. Tsherniak, A. et al. Defining a cancer dependency map. *Cell* **170**, 564–576 (2017).
54. Buitinck, L. et al. API design for machine learning software: experiences from the scikit-learn project. Preprint at *arXiv* https://doi.org/10.48550/arXiv.1309.0238 (2013).
55. Saelens, W., Cannoodt, R., Todorov, H. & Saeys, Y. A comparison of single-cell trajectory inference methods. *Nat. Biotechnol.* **37**, 547–554 (2019).
56. Haghverdi, L., Lun, A. T., Morgan, M. D. & Marioni, J. C. Batch effects in single-cell RNA-sequencing data are corrected by matching mutual nearest neighbors. *Nat. Biotechnol.* **36**, 421–427 (2018).

Acknowledgements

We thank members of the Lee laboratory for their helpful feedback on this work. This work was funded by NSF DBI-1552309 and DBI-1759487 (E.W., C.L. and S.-I.L.), NIH R35-GM-128638 and R01-NIA-AG-061132 (E.W., C.L. and S.-I.L.). E.W. was supported by the National Science Foundation Graduate Research Fellowship under grant no. DGE-2140004.

Author contributions

E.W. and C.L. contributed equally. E.W. conceived the study with input from S.-I.L. E.W. implemented an initial prototype of contrastiveVI, and C.L. wrote the final refactored scvi-tools implementation and associated tests. E.W. and C.L. both applied the model to analyze the datasets considered in this work with input from S.-I.L. S.-I.L. supervised the work. All authors participated in writing the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41592-023-01955-3.

Correspondence and requests for materials should be addressed to Su-In Lee.

Peer review information *Nature Methods* thanks Natalie Davidson, and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Primary Handling Editors: Lei Tang and Lin Tang, in collaboration with the *Nature Methods* team.

nature portfolio

Corresponding author(s): Su-In Lee

Last updated by author(s): May 18, 2023

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed
☐	☒ The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒ The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒ A description of all covariates tested
☐	☒ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒ For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☐	☒ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☐	☒ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒ Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	scsim (https://github.com/dylkot/scsim) was used to generate the simulated dataset. No other custom software was used to collect the data analyzed in this manuscript.
Data analysis	https://github.com/suinleelab/contrastiveVI-reproducibility Additional Python packages: scvi-tools (version 0.16.1), scanpy (version 1.8.1), gseapy (version 1.0.4), cplvm (version 0.3), diffxpy (version 0.7.4)

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

All datasets analyzed in this paper are publicly available. The simulated dataset was generated using the scsim package found at https://github.com/dylkot/scsim.

The MIX-Seq dataset from McFarland et al. was downloaded from the authors Figshare repository (https://figshare.com/articles/dataset/MIX-seq_data/10298696). The Haber et al., Norman et al., and Papalexi et al. datasets were downloaded from the National Institutes of Health Gene Expression Omnibus (accesssion codes GSE92332, GSE133344, and GSE153056, respectively). Our code for downloading and preprocessing these datasets is available at https://github.com/suinleelab/contrastiveVI-reproducibility.

Research involving human participants, their data, or biological material

Policy information about studies with human participants or human data. See also policy information about sex, gender (identity/presentation), and sexual orientation and race, ethnicity and racism.

Reporting on sex and gender	n/a
Reporting on race, ethnicity, or other socially relevant groupings	n/a
Population characteristics	n/a
Recruitment	n/a
Ethics oversight	n/a

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	No new sample size calculations were performed in this study. All datasets considered in this work were reported in previous studies which used established criteria for determining the number of cells to collect in order to ensure sufficient power when assessing differences among different subgroups of cells.
Data exclusions	No new datasets were collected in this study. Where applicable, when preprocessing the datasets considered in this work we excluded cells from further consideration using the same exclusion criteria established by the authors of the paper presenting a given dataset (e.g. cells previously marked as "Low quality" were filtered out).
Replication	For quantitative comparisons between contrastiveVI and baseline models, all models were trained with five random weight initializations. We report the (min-max scaled) average performance of each method for each metric in the main text figures, and tables with mean metric values and standard errors can be found in our supplementary materials. Attempts at replication were successful (i.e., model performance was largely consistent across different random initializations).
Randomization	No new datasets were collected in this study (i.e., all experimental group calculations were performed by the authors of the original publications presenting the datasets considered in this work).
Blinding	Not applied. The behavior of the computational models tested in this study is not affected by blinding.

Behavioural & social sciences study design

All studies must disclose on these points even when the disclosure is negative.

Study description	*Briefly describe the study type including whether data are quantitative, qualitative, or mixed-methods (e.g. qualitative cross-sectional, quantitative experimental, mixed-methods case study).*
Research sample	*State the research sample (e.g. Harvard university undergraduates, villagers in rural India) and provide relevant demographic information (e.g. age, sex) and indicate whether the sample is representative. Provide a rationale for the study sample chosen. For studies involving existing datasets, please describe the dataset and source.*
Sampling strategy	*Describe the sampling procedure (e.g. random, snowball, stratified, convenience). Describe the statistical methods that were used to predetermine sample size OR if no sample-size calculation was performed, describe how sample sizes were chosen and provide a rationale for why these sample sizes are sufficient. For qualitative data, please indicate whether data saturation was considered, and*

	what criteria were used to decide that no further sampling was needed.
Data collection	*Provide details about the data collection procedure, including the instruments or devices used to record the data (e.g. pen and paper, computer, eye tracker, video or audio equipment) whether anyone was present besides the participant(s) and the researcher, and whether the researcher was blind to experimental condition and/or the study hypothesis during data collection.*
Timing	*Indicate the start and stop dates of data collection. If there is a gap between collection periods, state the dates for each sample cohort.*
Data exclusions	*If no data were excluded from the analyses, state so OR if data were excluded, provide the exact number of exclusions and the rationale behind them, indicating whether exclusion criteria were pre-established.*
Non-participation	*State how many participants dropped out/declined participation and the reason(s) given OR provide response rate OR state that no participants dropped out/declined participation.*
Randomization	*If participants were not allocated into experimental groups, state so OR describe how participants were allocated to groups, and if allocation was not random, describe how covariates were controlled.*

Ecological, evolutionary & environmental sciences study design

All studies must disclose on these points even when the disclosure is negative.

Study description	*Briefly describe the study. For quantitative data include treatment factors and interactions, design structure (e.g. factorial, nested, hierarchical), nature and number of experimental units and replicates.*
Research sample	*Describe the research sample (e.g. a group of tagged Passer domesticus, all Stenocereus thurberi within Organ Pipe Cactus National Monument), and provide a rationale for the sample choice. When relevant, describe the organism taxa, source, sex, age range and any manipulations. State what population the sample is meant to represent when applicable. For studies involving existing datasets, describe the data and its source.*
Sampling strategy	*Note the sampling procedure. Describe the statistical methods that were used to predetermine sample size OR if no sample-size calculation was performed, describe how sample sizes were chosen and provide a rationale for why these sample sizes are sufficient.*
Data collection	*Describe the data collection procedure, including who recorded the data and how.*
Timing and spatial scale	*Indicate the start and stop dates of data collection, noting the frequency and periodicity of sampling and providing a rationale for these choices. If there is a gap between collection periods, state the dates for each sample cohort. Specify the spatial scale from which the data are taken*
Data exclusions	*If no data were excluded from the analyses, state so OR if data were excluded, describe the exclusions and the rationale behind them, indicating whether exclusion criteria were pre-established.*
Reproducibility	*Describe the measures taken to verify the reproducibility of experimental findings. For each experiment, note whether any attempts to repeat the experiment failed OR state that all attempts to repeat the experiment were successful.*
Randomization	*Describe how samples/organisms/participants were allocated into groups. If allocation was not random, describe how covariates were controlled. If this is not relevant to your study, explain why.*
Blinding	*Describe the extent of blinding used during data acquisition and analysis. If blinding was not possible, describe why OR explain why blinding was not relevant to your study.*

Did the study involve field work? ☐ Yes ☐ No

Field work, collection and transport

Field conditions	*Describe the study conditions for field work, providing relevant parameters (e.g. temperature, rainfall).*
Location	*State the location of the sampling or experiment, providing relevant parameters (e.g. latitude and longitude, elevation, water depth).*
Access & import/export	*Describe the efforts you have made to access habitats and to collect and import/export your samples in a responsible manner and in compliance with local, national and international laws, noting any permits that were obtained (give the name of the issuing authority, the date of issue, and any identifying information).*
Disturbance	*Describe any disturbance caused by the study and how it was minimized.*

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☒	☐ Clinical data
☒	☐ Dual use research of concern
☒	☐ Plants

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Antibodies

Antibodies used — *Describe all antibodies used in the study; as applicable, provide supplier name, catalog number, clone name, and lot number.*

Validation — *Describe the validation of each primary antibody for the species and application, noting any validation statements on the manufacturer's website, relevant citations, antibody profiles in online databases, or data provided in the manuscript.*

Eukaryotic cell lines

Policy information about cell lines and Sex and Gender in Research

Cell line source(s) — *State the source of each cell line used and the sex of all primary cell lines and cells derived from human participants or vertebrate models.*

Authentication — *Describe the authentication procedures for each cell line used OR declare that none of the cell lines used were authenticated.*

Mycoplasma contamination — *Confirm that all cell lines tested negative for mycoplasma contamination OR describe the results of the testing for mycoplasma contamination OR declare that the cell lines were not tested for mycoplasma contamination.*

Commonly misidentified lines (See ICLAC register) — *Name any commonly misidentified cell lines used in the study and provide a rationale for their use.*

Palaeontology and Archaeology

Specimen provenance — *Provide provenance information for specimens and describe permits that were obtained for the work (including the name of the issuing authority, the date of issue, and any identifying information). Permits should encompass collection and, where applicable, export.*

Specimen deposition — *Indicate where the specimens have been deposited to permit free access by other researchers.*

Dating methods — *If new dates are provided, describe how they were obtained (e.g. collection, storage, sample pretreatment and measurement), where they were obtained (i.e. lab name), the calibration program and the protocol for quality assurance OR state that no new dates are provided.*

☐ Tick this box to confirm that the raw and calibrated dates are available in the paper or in Supplementary Information.

Ethics oversight — *Identify the organization(s) that approved or provided guidance on the study protocol, OR state that no ethical approval or guidance was required and explain why not.*

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Animals and other research organisms

Policy information about studies involving animals; ARRIVE guidelines recommended for reporting animal research, and Sex and Gender in Research

Laboratory animals — *For laboratory animals, report species, strain and age OR state that the study did not involve laboratory animals.*

Wild animals — *Provide details on animals observed in or captured in the field; report species and age where possible. Describe how animals were caught and transported and what happened to captive animals after the study (if killed, explain why and describe method; if released, say where and when) OR state that the study did not involve wild animals.*

Reporting on sex	*Indicate if findings apply to only one sex; describe whether sex was considered in study design, methods used for assigning sex. Provide data disaggregated for sex where this information has been collected in the source data as appropriate; provide overall numbers in this Reporting Summary. Please state if this information has not been collected. Report sex-based analyses where performed, justify reasons for lack of sex-based analysis.*
Field-collected samples	*For laboratory work with field-collected samples, describe all relevant parameters such as housing, maintenance, temperature, photoperiod and end-of-experiment protocol OR state that the study did not involve samples collected from the field.*
Ethics oversight	*Identify the organization(s) that approved or provided guidance on the study protocol, OR state that no ethical approval or guidance was required and explain why not.*

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Clinical data

Policy information about clinical studies

All manuscripts should comply with the ICMJE guidelines for publication of clinical research and a completed CONSORT checklist must be included with all submissions.

Clinical trial registration	*Provide the trial registration number from ClinicalTrials.gov or an equivalent agency.*
Study protocol	*Note where the full trial protocol can be accessed OR if not available, explain why.*
Data collection	*Describe the settings and locales of data collection, noting the time periods of recruitment and data collection.*
Outcomes	*Describe how you pre-defined primary and secondary outcome measures and how you assessed these measures.*

Dual use research of concern

Policy information about dual use research of concern

Hazards

Could the accidental, deliberate or reckless misuse of agents or technologies generated in the work, or the application of information presented in the manuscript, pose a threat to:

No	Yes	
☐	☐	Public health
☐	☐	National security
☐	☐	Crops and/or livestock
☐	☐	Ecosystems
☐	☐	Any other significant area

Experiments of concern

Does the work involve any of these experiments of concern:

No	Yes	
☐	☐	Demonstrate how to render a vaccine ineffective
☐	☐	Confer resistance to therapeutically useful antibiotics or antiviral agents
☐	☐	Enhance the virulence of a pathogen or render a nonpathogen virulent
☐	☐	Increase transmissibility of a pathogen
☐	☐	Alter the host range of a pathogen
☐	☐	Enable evasion of diagnostic/detection modalities
☐	☐	Enable the weaponization of a biological agent or toxin
☐	☐	Any other potentially harmful combination of experiments and agents

Plants

Seed stocks	*Report on the source of all seed stocks or other plant material used. If applicable, state the seed stock centre and catalogue number. If plant specimens were collected from the field, describe the collection location, date and sampling procedures.*
Novel plant genotypes	*Describe the methods by which all novel plant genotypes were produced. This includes those generated by transgenic approaches, gene editing, chemical/radiation-based mutagenesis and hybridization. For transgenic lines, describe the transformation method, the*

	number of independent lines analyzed and the generation upon which experiments were performed. For gene-edited lines, describe the editor used, the endogenous sequence targeted for editing, the targeting guide RNA sequence (if applicable) and how the editor was applied.
Authentication	*Describe any authentication procedures for each seed stock used or novel genotype generated. Describe any experiments used to assess the effect of a mutation and, where applicable, how potential secondary effects (e.g. second site T-DNA insertions, mosiacism, off-target gene editing) were examined.*

ChIP-seq

Data deposition

☐ Confirm that both raw and final processed data have been deposited in a public database such as GEO.

☐ Confirm that you have deposited or provided access to graph files (e.g. BED files) for the called peaks.

Data access links *May remain private before publication.*	*For "Initial submission" or "Revised version" documents, provide reviewer access links. For your "Final submission" document, provide a link to the deposited data.*
Files in database submission	*Provide a list of all files available in the database submission.*
Genome browser session (e.g. UCSC)	*Provide a link to an anonymized genome browser session for "Initial submission" and "Revised version" documents only, to enable peer review. Write "no longer applicable" for "Final submission" documents.*

Methodology

Replicates	*Describe the experimental replicates, specifying number, type and replicate agreement.*
Sequencing depth	*Describe the sequencing depth for each experiment, providing the total number of reads, uniquely mapped reads, length of reads and whether they were paired- or single-end.*
Antibodies	*Describe the antibodies used for the ChIP-seq experiments; as applicable, provide supplier name, catalog number, clone name, and lot number.*
Peak calling parameters	*Specify the command line program and parameters used for read mapping and peak calling, including the ChIP, control and index files used.*
Data quality	*Describe the methods used to ensure data quality in full detail, including how many peaks are at FDR 5% and above 5-fold enrichment.*
Software	*Describe the software used to collect and analyze the ChIP-seq data. For custom code that has been deposited into a community repository, provide accession details.*

Flow Cytometry

Plots

Confirm that:

☐ The axis labels state the marker and fluorochrome used (e.g. CD4-FITC).

☐ The axis scales are clearly visible. Include numbers along axes only for bottom left plot of group (a 'group' is an analysis of identical markers).

☐ All plots are contour plots with outliers or pseudocolor plots.

☐ A numerical value for number of cells or percentage (with statistics) is provided.

Methodology

Sample preparation	*Describe the sample preparation, detailing the biological source of the cells and any tissue processing steps used.*
Instrument	*Identify the instrument used for data collection, specifying make and model number.*
Software	*Describe the software used to collect and analyze the flow cytometry data. For custom code that has been deposited into a community repository, provide accession details.*
Cell population abundance	*Describe the abundance of the relevant cell populations within post-sort fractions, providing details on the purity of the samples and how it was determined.*
Gating strategy	*Describe the gating strategy used for all relevant experiments, specifying the preliminary FSC/SSC gates of the starting cell population, indicating where boundaries between "positive" and "negative" staining cell populations are defined.*

☐ Tick this box to confirm that a figure exemplifying the gating strategy is provided in the Supplementary Information.

Magnetic resonance imaging

Experimental design

Design type	*Indicate task or resting state; event-related or block design.*
Design specifications	*Specify the number of blocks, trials or experimental units per session and/or subject, and specify the length of each trial or block (if trials are blocked) and interval between trials.*
Behavioral performance measures	*State number and/or type of variables recorded (e.g. correct button press, response time) and what statistics were used to establish that the subjects were performing the task as expected (e.g. mean, range, and/or standard deviation across subjects).*

Acquisition

Imaging type(s)	*Specify: functional, structural, diffusion, perfusion.*
Field strength	*Specify in Tesla*
Sequence & imaging parameters	*Specify the pulse sequence type (gradient echo, spin echo, etc.), imaging type (EPI, spiral, etc.), field of view, matrix size, slice thickness, orientation and TE/TR/flip angle.*
Area of acquisition	*State whether a whole brain scan was used OR define the area of acquisition, describing how the region was determined.*

Diffusion MRI ☐ Used ☐ Not used

Preprocessing

Preprocessing software	*Provide detail on software version and revision number and on specific parameters (model/functions, brain extraction, segmentation, smoothing kernel size, etc.).*
Normalization	*If data were normalized/standardized, describe the approach(es): specify linear or non-linear and define image types used for transformation OR indicate that data were not normalized and explain rationale for lack of normalization.*
Normalization template	*Describe the template used for normalization/transformation, specifying subject space or group standardized space (e.g. original Talairach, MNI305, ICBM152) OR indicate that the data were not normalized.*
Noise and artifact removal	*Describe your procedure(s) for artifact and structured noise removal, specifying motion parameters, tissue signals and physiological signals (heart rate, respiration).*
Volume censoring	*Define your software and/or method and criteria for volume censoring, and state the extent of such censoring.*

Statistical modeling & inference

Model type and settings	*Specify type (mass univariate, multivariate, RSA, predictive, etc.) and describe essential details of the model at the first and second levels (e.g. fixed, random or mixed effects; drift or auto-correlation).*
Effect(s) tested	*Define precise effect in terms of the task or stimulus conditions instead of psychological concepts and indicate whether ANOVA or factorial designs were used.*

Specify type of analysis: ☐ Whole brain ☐ ROI-based ☐ Both

Statistic type for inference (See Eklund et al. 2016)	*Specify voxel-wise or cluster-wise and report all relevant parameters for cluster-wise methods.*
Correction	*Describe the type of correction and how it is obtained for multiple comparisons (e.g. FWE, FDR, permutation or Monte Carlo).*

Models & analysis

n/a	Involved in the study
☐	☐ Functional and/or effective connectivity
☐	☐ Graph analysis
☐	☐ Multivariate modeling or predictive analysis

Functional and/or effective connectivity	*Report the measures of dependence used and the model details (e.g. Pearson correlation, partial correlation, mutual information).*
Graph analysis	*Report the dependent variable and connectivity measure, specifying weighted graph or binarized graph,*

Graph analysis	*subject- or group-level, and the global and/or node summaries used (e.g. clustering coefficient, efficiency, etc.).*
Multivariate modeling and predictive analysis	*Specify independent variables, features extraction and dimension reduction, model, training and evaluation metrics.*