PNAS RESEARCH ARTICLE | STATISTICS BIOPHYSICS AND COMPUTATIONAL BIOLOGY

Quantifying common and distinct information in single-cell multimodal data with Tilted Canonical Correlation Analysis

Kevin Z. Lin[a,1] and Nancy R. Zhang[a]

Edited by David Weitz, Harvard University, Cambridge, MA; received March 6, 2023; accepted June 24, 2023

Multimodal single-cell technologies profile multiple modalities for each cell simultaneously, enabling a more thorough characterization of cell populations. Existing dimension-reduction methods for multimodal data capture the "union of information," producing a lower-dimensional embedding that combines the information across modalities. While these tools are useful, we focus on a fundamentally different task of separating and quantifying the information among cells that is shared between the two modalities as well as unique to only one modality. Hence, we develop Tilted Canonical Correlation Analysis (Tilted-CCA), a method that decomposes a paired multimodal dataset into three lower-dimensional embeddings—one embedding captures the "intersection of information," representing the geometric relations among the cells that is common to both modalities, while the remaining two embeddings capture the "distinct information for a modality," representing the modality-specific geometric relations. We analyze single-cell multimodal datasets sequencing RNA along surface antibodies (i.e., CITE-seq) as well as RNA alongside chromatin accessibility (i.e., 10x) for blood cells and developing neurons via Tilted-CCA. These analyses show that Tilted-CCA enables meaningful visualization and quantification of the cross-modal information. Finally, Tilted-CCA's framework allows us to perform two specific downstream analyses. First, for single-cell datasets that simultaneously profile transcriptome and surface antibody markers, we show that Tilted-CCA helps design the target antibody panel to complement the transcriptome best. Second, for developmental single-cell datasets that simultaneously profile transcriptome and chromatin accessibility, we show that Tilted-CCA helps identify development-informative genes and distinguish between transient versus terminal cell types.

multimodal data | matrix factorization | canonical correlation analysis | multiview data | single-cell genomics

High-dimensional multimodal data, where features belonging to two or more modalities are simultaneously profiled, are becoming increasingly widespread across disciplines. In this paper, we focus on paired multimodal data arising in the field of single-cell biology, where technological advances have recently enabled simultaneous profiling of multiple types of features, such as RNA expression, protein abundance, and chromatin accessibility within the same cell and across many cells in parallel (1–4). This type of data is invaluable because cellular processes operate on multiple molecular modalities, and observation of any single modality offers only a partial view of an interconnected system (5–9). When analyzing such data, a basic question is how to separate and quantify the cell-separation geometric information shared across modalities and those unique to a particular modality. We address this question and demonstrate how this quantification can provide scientific insight. Despite our explicit focus on single-cell genomics, the questions and proposed method here broadly apply to paired multimodal data. For concreteness and clarity, we will refer to the individual data points as "cells."

Existing dimension-reduction methods are useful to analyze paired multimodal single-cell data by estimating a low-dimensional space spanning both modalities that captures the "union of information" across both modalities. That is, subpopulations of cells are separable in the low-dimensional space if they are separable in either modality. These methods include JIVE (10), WNN (11), MOFA+ (12), scAI (13), and JSNMF (14), all of which have helped identify nuanced cell types by combining information across both modalities.

In contrast, there is yet no rigorous way to quantify the information shared between both modalities. Compared to the task performed by the aforementioned methods, this is a fundamentally different task that can be thought of as learning the "intersection of information" between modalities. We motivate this from a geometric perspective—two sets of cells are separated in the "common" embedding if both modalities agree that the

Significance

Emerging single-cell multimodal technologies that simultaneously profile two biological modalities are rapidly being incorporated into biomedical research to obtain a more comprehensive understanding of a biological system. However, many existing analyses focus on aggregating the axes of variation across both modalities. What is missing is a computational method to separate and quantify the axes of variation shared between the two modalities from those unique to a particular modality. Hence, we propose a matrix factorization called Tilted-CCA and demonstrate its effectiveness on various multimodal datasets across many biological systems. Our method aids in selecting targeted surface antibody panels that best complement the transcriptome and provides insights into a cell's developmental status by leveraging relations between the transcriptome and chromatin accessibility.

Author contributions: K.Z.L. and N.R.Z. designed research; K.Z.L. and N.R.Z. performed research; K.Z.L. and N.R.Z. contributed new reagents/analytic tools; K.Z.L. analyzed data; and K.Z.L. and N.R.Z. wrote the paper.

The authors declare no competing interest.

This article is a PNAS Direct Submission.

[1] To whom correspondence may be addressed. Email: kevinL1@wharton.upenn.edu.

This article contains supporting information online at https://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2303647120/-/DCSupplemental.

Published July 31, 2023.

sets of cells are separable. Our matrix factorization method, Tilted-CCA, extends the Canonical Correlation Analysis (CCA) (15) framework by decomposing the canonical score vectors in a principled way to uncover an embedding of the cells that abides by the aforementioned geometric perspective. We show through examples that Tilted-CCA provides meaningful quantifications of the overlap (or lack thereof) in information between the two modalities at both the cell and the feature levels. However, we note that the "union" and "intersection" embeddings complement one another when analyzing multimodal data—while the former provides a complete view of all axes of variation supported by either modality, the latter provides insight into which axes of variation are supported by both modalities. Additionally, Tilted-CCA's decomposition quantifies two additional "distinct" embeddings representing the axes of variation unique to either modality after removing the intersection.

We further illustrate the scientific insight enabled by Tilted-CCA through two case studies. In our first case study, we consider the problem of antibody-panel design in paired single-cell profiling of RNA expression and surface antibody abundance (16, 17). Such data are becoming standard in immunology research since combining both modalities enables accurate labeling of cell identity (18–20). However, large antibody panels are expensive for large cohorts. Toward this end, we demonstrate that by quantifying the common and distinct information between the RNA and protein modalities, Tilted-CCA helps to design small antibody panels that most effectively separate immune cell types when paired with transcriptomic data.

In our second case study, we investigate the coordination between chromatin accessibility and gene expression during tissue development and show that Tilted-CCA provides natural metrics distinguishing between transient versus terminal cell states and identifying development-associated genes (21–24). Many existing pseudotime-estimation methods address these questions, but they rely solely on gene expression (25–31) or chromatin accessibility alone (32), and typically require the developmental trajectory estimate. In contrast, the relation between Tilted-CCA's common and distinct embeddings between both modalities provides an alternate and more flexible approach to answer these questions.

Results

Tilted-CCA and the Intersection of Information. We introduce Tilted-CCA through an example of a single-cell CITE-seq dataset, where $n = 30{,}672$ human bone marrow cells are simultaneously profiled along $p_1 = 2{,}000$ genes (RNA modality) and a panel of $p_2 = 25$ surface antibody markers (protein modality) (18). The modality-specific Uniform Manifold Approximation and Projection plots (UMAPs) demonstrate that while the major immune subtypes, such as myeloid, B cells, and T cells, are separated in both modalities, the protein modality better separates the T cell subtypes (Fig. 1*A*). This is expected since the 25 protein antibodies are chosen to separate many T cell subtypes (11). We use this multimodal dataset to exemplify matrix factorization methods, which strive to factorize the data matrices $X^{(1)}$ and $X^{(2)}$ of both modalities into the product of modality-specific loading matrices ($L^{(1)}$ or $L^{(2)}$) and score matrices that decompose into a common embedding C and a modality-specific distinct embedding ($D^{(1)}$ or $D^{(2)}$) (Fig. 1*B*). However, their mathematical properties and biological interpretations can vary dramatically depending on which axes of variation are represented by C and D.

Broadly speaking, existing unsupervised methods define C in Fig. 1*B* to capture the "union of information" (10). A prototypical method is Consensus Principal Component Analysis (PCA) (33–35), which finds a low-dimensional embedding to best approximate both modalities simultaneously. This embedding aggregates the axes of variation in each modality, regardless of whether or not those axes are common to both modalities or unique to a specific modality. This is demonstrated in our CITE-seq dataset, where Consensus PCA simultaneously separates all of the progenitor cell types, CD4+, CD8+, and NK T cells from each other (Fig. 1*C*), a quality that neither modality had alone. To explain how Consensus PCA yields a decomposition in the form of Fig. 1*B*, consider the leading principal components of each modality, each having information to differentiate a different set of cell types (Fig. 1*D*). The common embedding C defined by Consensus PCA is the linear subspace that bisects both principal components, hence retaining the cell-type separation patterns in each modality. Then, the modality-specific distinct embeddings $D^{(1)}$ and $D^{(2)}$ are the residuals orthogonal to C (Fig. 1*E*). Other linear methods such as JIVE (10), MOFA+ (12), scAI (13) and JSNMF (14), and nonlinear methods such as WNN (11) share similar qualities to Consensus PCA (see *SI Appendix*, Fig. S4, for more details).

In contrast to the aforementioned methods, Tilted-CCA is an unsupervised method that estimates a common embedding C that quantifies axes of variation supported by both modalities, i.e., the "intersection of information." This is a fundamentally different goal and yields different insights. This embedding aims to represent the geometric relations shared between the two modalities—two sets of cells should be separable in C if and only if they were separable in each modality alone. For example, in this CITE-seq data, Tilted-CCA's common embedding C separates the myeloid, B-, and T cells from one another, since this separation is supported by both modalities but does not separate the CD4+ and CD8+ T cell subtype, since their separation is unique to the protein modality (Fig. 1*F*). Then, by imposing appropriate mathematical properties onto the decomposition in Fig. 1*B*, Tilted-CCA could then be also able to estimate axes of variation unique to each modality via the distinct embeddings $D^{(1)}$ and $D^{(2)}$.

To achieve these goals, Tilted-CCA builds upon CCA, a method that finds linear transformations for each modality to yield the canonical score matrices $Z^{(1)}$ and $Z^{(2)} \in \mathbb{R}^{n\times r}$ that have high cross-correlation. Here, the leading pair of canonical vectors in $Z^{(1)}$ and $Z^{(2)}$ exemplifies the shared pattern between both modalities that separate the most major cell types (Fig. 1*G*). However, CCA, by itself, only provides two canonical score matrices $Z^{(1)}$ and $Z^{(2)}$ and does not imply an explicit decomposition of the common and distinct axes of variation in the framework of Fig. 1*B*. Tilted-CCA fills this gap by starting from CCA and then decomposing $Z^{(1)}$ and $Z^{(2)}$,

$$Z^{(1)} = C + D^{(1)}, \quad \text{and} \quad Z^{(2)} = C + D^{(2)},$$

where 1) the common embedding C encapsulates the geometric relations of the "intersection of information" and 2) the two distinct embeddings $(D^{(1)})^\top D^{(2)} = 0$. The latter orthogonality constraint was first proposed in D-CCA (36), and enables us to interpret $D^{(1)}$ and $D^{(2)}$ as capturing modality-specific variation. This constraint restricts the common vector in each latent dimension of C to lie along a semicircle defined by the canonical score vectors in $Z^{(1)}$ and $Z^{(2)}$ (Fig. 1*H*). Along

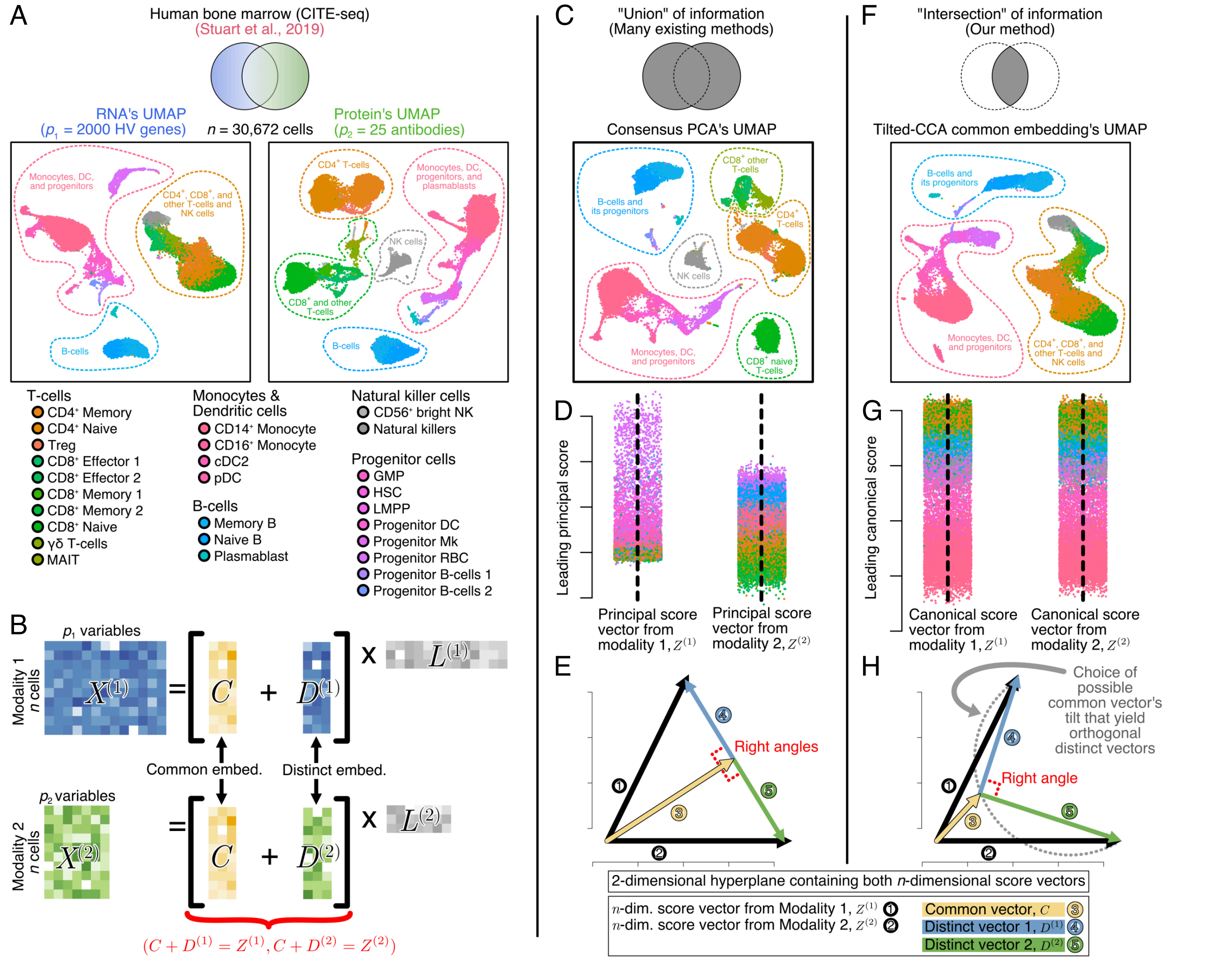

Fig. 1. Embedding methods for multimodal data that learn the "intersection" of information differ from those that learn the "union". (*A*) Summary of the bone marrow CITE-seq dataset, showing either the UMAP of the RNA or protein modality, where cells are colored by the annotated cell types. Note that the coloring scheme of cells is shared among (*A*, *C*, *D*, *F*, and *G*). (*B*) Schematic of a matrix factorization of single-cell paired multiomic data, where $X^{(1)} \in \mathbb{R}^{p_1 \times n}$ denotes the n cells measured on p_1 features and $X^{(1)} \in \mathbb{R}^{p_2 \times n}$ denotes the same n cells measured on p_2 features, already preprocessed to be low-rank. Here, $Z^{(1)}$ and $Z^{(2)}$ denote matrices of score vectors, where the specific definition of these score vectors depends on the context. (*C*) UMAP of the global embedding between the RNA and protein modalities estimated by Consensus PCA, showing the "union of information." (*D*) The pair of leading principal score vectors, one from each modality. (*E*) Mathematical illustration of the decomposition of $Z^{(1)}$ and $Z^{(2)}$ (the two sets of principal score vectors) to achieve a decomposition as shown in (*B*) as used by methods like JIVE. (*F*) UMAP of the common embedding C between the RNA and protein modality estimated by Tilted-CCA, showing the "intersection of information." (*G*) The pair of leading canonical score vectors, one from each modality. (*H*) Mathematical illustration of the decomposition of $Z^{(1)}$ and $Z^{(2)}$ (the two sets of canonical score vectors) used by Tilted-CCA to achieve a decomposition as shown in (*B*).

this semicircle, if a latent dimension of C tilts in the direction of $Z^{(1)}$, the common embedding would resemble Modality 1, leading to the interpretation that Modality 2 has more distinct information represented by large magnitudes in $D^{(2)}$, and vice versa (Fig. 2*A*). Thus, Tilted-CCA optimizes for the appropriate "tilt" of C along this semicircle for each latent dimension to yield the desired geometric relations among the cells. Specifically, Tilted-CCA first computes the nearest-neighbor graph of each modality where each node is a cell (Fig. 2*B*). Then, a target common manifold is constructed from both graphs in order to encapsulate the "intersection of information." Tilted-CCA optimizes the tilt of the common vector for each latent dimension so that the resulting common embedding's nearest neighbor graph approximates this target manifold (Fig. 2*C*). Once the appropriate common embedding C is estimated, we recover the estimated matrix decomposition of both distinct embeddings in the framework of Fig. 1*B* (*Methods*). (Details in *SI Appendix*, along with simulations in *SI Appendix*, Figs. S1–S3.) As we will demonstrate in later sections, the linear framework in Tilted-CCA aids in addressing various cell- or feature-level scientific questions.

Titled-CCA Quantifies the Overlapping versus Distinct Information Each Modality Contributes Toward the Separation of Cell Types. As proof of concept, we start with a pervasive question for multimodal single-cell data: Which cell types are separable in both modalities or only one modality? We use the provided cell-type annotations to investigate this. For example, the UMAPs of the RNA and protein modalities for the aforementioned CITE-seq bone marrow dataset suggest that the protein modality better

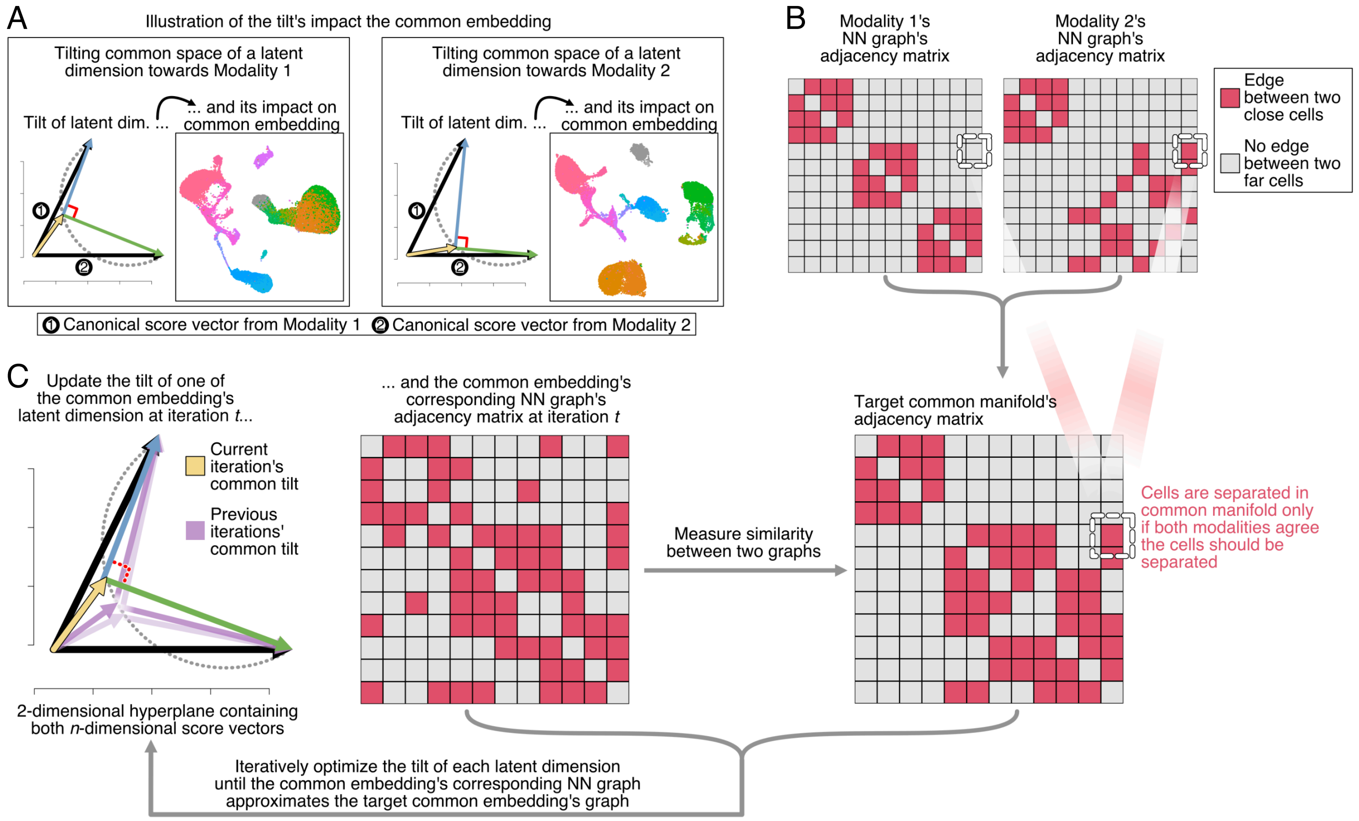
A
Illustration of the tilt's impact the common embedding
Tilting common space of a latent dimension towards Modality 1
Tilt of latent dim. ...
... and its impact on common embedding
Tilting common space of a latent dimension towards Modality 2
Tilt of latent dim. ...
... and its impact on common embedding
❶ Canonical score vector from Modality 1
❷ Canonical score vector from Modality 2
B
Modality 1's NN graph's adjacency matrix
Modality 2's NN graph's adjacency matrix
Edge between two close cells
No edge between two far cells
C
Update the tilt of one of the common embedding's latent dimension at iteration t...
Current iteration's common tilt
Previous iterations' common tilt
2-dimensional hyperplane containing both n-dimensional score vectors
... and the common embedding's corresponding NN graph's adjacency matrix at iteration t
Measure similarity between two graphs
Target common manifold's adjacency matrix
Cells are separated in common manifold only if both modalities agree the cells should be separated
Iteratively optimize the tilt of each latent dimension until the common embedding's corresponding NN graph approximates the target common embedding's graph

Fig. 2. Schematic of Tilted-CCA's optimization of common embedding's tilt across each latent dimension. (*A*) Schematic illustrating the tilt's impact on the common embedding using the bone marrow CITE-seq dataset, where the tilt for the common vector could either lean toward Modality 1 (meaning Modality 2 has more distinct information in this latent dimension) or vice versa. (*B*) Schematic of the nearest-neighbor graphs for each modality. (*C*) Flowchart of Tilted-CCA's optimization procedure. The target common manifold (represented as an adjacency matrix) is calculated-based both modalities' respective nearest-neighbor graphs. Then, an iterative optimization procedure is used where the tilt for each latent dimension is updated based on how similar its corresponding common embedding's nearest-neighbor graph is similar to the target common manifold.

differentiates the T cell subtypes than the RNA modality (Fig. 1*A*). Can we rigorously quantify how much the cell types are separated based on the overlapping (i.e., shared) information between both modalities or distinct information unique to a specific modality?

We start by qualitatively exploring the UMAPs of $D^{(1)}$ and $D^{(2)}$, the distinct components derived from the orthogonal remainder terms after removing the common component C. For the bone marrow CITE-seq data, we see that T cells are mixed in the RNA's distinct component $D^{(1)}$, while the progenitor cell types are well separated (Fig. 3*A*). On the other hand, the progenitor cell types mixed in the protein's distinct component $D^{(2)}$, while the T cells are well separated. These observations match our initial intuitions from a side-by-side comparison of each modality's UMAP in Fig. 1*A* alone. The major cell types, such as myeloid cells, B cells, and T cells, are also distinguishable in both distinct UMAPs, indicating that there are remaining axes of variation in both modalities that separate these cell types but are orthogonal to each other.

We design enrichment scores for each cell type and modality pairing to formally quantify the amount of distinct information that modality contributes toward separating the cell type. These scores do not rely on UMAPs. First, we compute the nearest-neighbor graph for either $D^{(1)}$ and $D^{(2)}$. Then, we define the enrichment score of a cell type by selecting all cells of said type and computing the proportion of their neighbors that share their type, normalized by the baseline proportion (*Methods*). A higher enrichment score signifies that the modality contains more distinct information to define said cell type. We see that for the CITE-seq data, the RNA modality has roughly 3 times more distinct information for the progenitors than the protein modality, while the protein modality has roughly 2.5 more distinct information for the CD4+ and CD8+ T cells than the RNA modality (Fig. 3*B*). This makes biological sense since progenitor cell populations involve transcriptome-level changes not expected to be captured by the 25 antibody markers. We also define the enrichment score for the common embedding, which demonstrates that the overlapping information between both modalities clearly defines the B cells (*SI Appendix*, Fig. S5 *A* and *B*).

Tilted-CCA Quantifies the Degree of Cross-Modality Alignment of Features. Now consider the features in each modality, such as the genes in the RNA modality and the surface antibody markers in the protein modality, for the bone marrow CITE-seq data. How much of each feature's variation lie in the common space versus its modality's distinct space? Compared to standard PCA analysis, this is analogous to asking which features are most aligned with each principal component, which can be addressed via the principal loadings. Here, we propose an intuitively similar metric for Tilted-CCA: For each feature, we quantify its alignment score, defined as the R^2 of regressing its observed expression onto its common component. A higher R^2 implies that the feature's variation lies mainly in the shared common space.

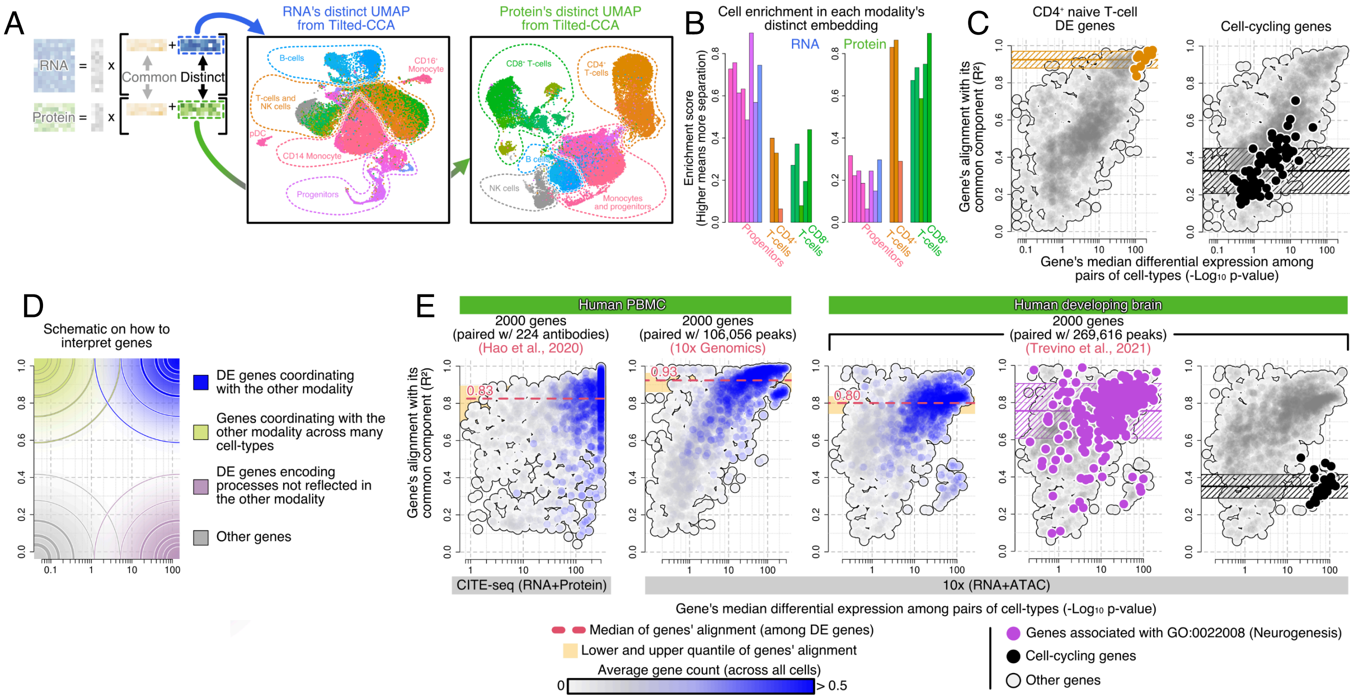

Fig. 3. Tilted-CCA quantifies shared/distinct information on both the cell and feature levels, enabling within-dataset and between-dataset comparisons. (*A*) Schematic of Tilted-CCA's matrix factorization and UMAP of each modality's distinct embedding in the bone marrow CITE-seq dataset. (*B*) Enrichment score of various cell subtypes among progenitors, CD4+ T cells, and CD8+ T cells based on either the RNA's or protein's distinct embedding. (*C*) Alignment-differentiability plot of the 2,000 genes for the bone marrow CITE-seq dataset, marking either marker genes for CD4+ naive T cells, CD8+ naive T cells, or cells associated with cell cycling. The median and quantiles of the alignment are shown along the y-axis for differentially expressed genes (i.e., with a differentiability score more than 10). (*D*) Schematic of how different regions of the alignment-differentiability plot are informative of different features' qualities. (*E*) Alignment-differentiability plots of the 2,000 genes for various datasets measuring RNA+protein or RNA+ATAC modalities, where genes are colored based on their average count, and the median alignment among DE genes for each dataset is marked by a red dashed line. Additionally, the genes associated with neurogenesis and cell cycling are marked in purple and black respectively, with the median and quantiles among these sets of genes shown along the y-axis.

As proof of principle, we hypothesize that cell-type marker genes should be highly aligned with the common space since these genes' expressions should correlate with antibody markers. To investigate this, we plot the common-alignment R^2 of each gene against its differentiability, a meta-statistic that summarizes its differential expression across all pairs of a priori defined cell types. A higher differentiability implies that the gene is a strong marker for certain cell type(s). We see that highly differentiable genes, such as the markers for CD4+ T cells, are indeed highly aligned with the protein modality (Fig. 3*C*). In contrast, genes belonging to processes which do not differentiate cell types, such as cell cycling, have low alignment with common space (Fig. 3*C* and *SI Appendix*, Fig. S5*C*). The feature-level alignments to the common space follow our expectations. See *SI Appendix* for analogous analyses, but focusing on the antibodies.

The alignment-differentiability plots also provide a bird's-eye view of the amount of overlapping information between modalities when comparing across different technologies and biological systems (Fig. 3*D*). Genes in the *Upper-Right* quadrant (blue) of the alignment-differentiability plot are cell-type markers that coordinate with the other modality. In contrast, genes in the *Lower-Right* quadrant (purple) are cell-type markers that complement the other modality. As an example, we examine two multiomic experiments profiling PBMC: a CITE-seq experiment pairing full-transcriptome RNA-seq with 224 antibody markers (11), and a 10x Multiome experiment pairing full-transcriptome RNA-seq with ATAC (37), which measures chromatin accessibility across the genome (Fig. 3*E*). The median alignment of differentially expressed genes with the protein modality is remarkably lower than their alignment with the ATAC modality (0.83 and 0.93, respectively). These results demonstrate that for PBMC, the transcriptome provides additional cell-type separation patterns not present among the antibodies, but the transcriptome and chromatin accessibility predominately capture the same axes of variation.

Next, we compare two tissues, PBMC and the developing brain, sequenced using 10x Multiome. The cross-modality alignment is much higher in PBMC compared to the developing brain (0.93 and 0.8, respectively). Although this difference could be partially explained by slight differences in ATAC sequencing coverage (*SI Appendix*, Fig. S5*D*), we hypothesize that the main driver is biological—the developing brain contains mostly differentiating cell populations, where RNA expression is less in sync with chromatin-level changes as cells are changing cell states. This contrasts PBMC, a terminally differentiated population where RNA-chromatin relations have stabilized. Our observation is supported by examining neurogenesis-related genes (purple) in the developing brain. Many of these genes have low alignment with the common space, which suggests that the degree of feature alignment with Tilted-CCA's common embedding may contain developmental information. We examine this in more detail in a later section through the construction of synchrony scores.

From Fig. 3*E*, it is also worth noting that cell cycle genes generally have high differentiability and low common space alignment. Specifically, this signal is unique to the RNA modality and not shared with ATAC (*SI Appendix*, Fig. S5*E*). This observation is a recurring theme in developing tissues since biologically, the cell cycle is a transient process and is not a permanent aspect of cell identity encoded at the chromatin level.

Designing Minimally Sized Antibody Panels for CITE-seq Data with Tilted-CCA. Multiomic sequencing technologies for joint assay of RNA expression and surface protein abundancy, such as CITE-seq (16) or Abseq (17), require practitioners to select the antibody panel (see *SI Appendix*, Fig. S6 for additional examples). Large antibody panels are expensive, making them impractical for large cohort studies. Hence, we desire to design a small panel of antibodies that provide the most distinct information, complementary to RNA, to separate cell types. We hypothesize that Tilted-CCA is suitable for achieving this since quantifying the intersection, and difference between two modalities can aid in selecting the features that best contribute to their union.

For illustration, we consider an Abseq dataset of 461 genes and 97 surface antibodies of cells from human bone marrow (20) (Fig. 4*A*). Fig. 4*B* previews the panel of 10 antibodies selected by Tilted-CCA (*SI Appendix*, Fig. S7*A*). The 10 antibodies do not separate the cell types well by themselves. However, when these antibodies are paired with the RNA modality via Consensus PCA, the cell types are almost as separable as the original multimodal data (*SI Appendix*, Fig. S7*B*). At a high level, our antibody-panel design strategy greedily finds a set of antibodies whose protein's distinct components are differentially expressed across cell types and are as uncorrelated as possible to each other and to the common embedding (*SI Appendix*, Fig. S7*C*). As a diagnostic, we plot the correlation network among the 97 antibodies, where edges connect pairs of antibodies with highly correlated distinct components (Fig. 4*C*). Here, the node color and size reflect differentiability and alignment with common space, respectively. The 10 selected antibodies are spread out across this network, demonstrating their uncorrelated nature, while having highly differentiable distinct components that are lowly aligned with the common subspace. In contrast, conventional immune markers such as CD11b+, CD14+, and CD16+ are not selected since their protein expressions are already aligned with RNA.

To quantify the benefits of using Tilted-CCA, we also consider two other strategies for antibody-panel design: selecting the 10 antibodies that are most differentially expressed or the 10 antibodies that are least correlated with the RNA modality. These methods yield Consensus PCA embeddings that have poorer separation among cell types, as quantified by using the aforementioned enrichment scores (Fig. 4*D*, *SI Appendix*, Fig. S7 *D* and *E*). This is particularly evident among CD4+/CD8+ T cells and B cells. The former strategy suffers since highly differentiable antibodies might have expression patterns already prevalent in the gene expression and thus provides redundant information when paired with the RNA modality. The latter strategy suffers since antibodies not correlated with gene expressions do not necessarily provide high cell-type separation.

Tilted-CCA Reveals Transient Cell States and Development-Informative Genes in Developing Cell Populations. Joint profiling of gene expression and chromatin accessibility at the single-cell level enables the study of the coordination between chromatin remodeling and transcriptome reprogramming during cellular differentiation (23, 38). Toward this end, we explore the use of the common and distinct embeddings of Tilted-CCA to answer two questions: First, can we identify which cells are in a transient or terminal cell state? Second, can we identify genes associated with development and characterize the temporal coordination between the genes' chromatin activity and RNA expression? Many pseudotime-estimation methods, based on RNA or ATAC alone, have been developed to address these questions. These methods typically start by estimating the underlying cell trajectory and/or RNA velocity fields (25–27, 29, 32, 39, 40). Estimating either cell trajectories or velocity fields requires strong cell differentiation signals and is often difficult to recover with confidence in practice. Hence, complementing existing pipelines, we take an alternative approach to address the above two questions that do not depend on the estimation of cell trajectory or RNA velocity field.

We start with the premise that development is characterized by a coordinated change between the transcriptome and chromatin accessibility. Hence, we posit that the geometry of cells in

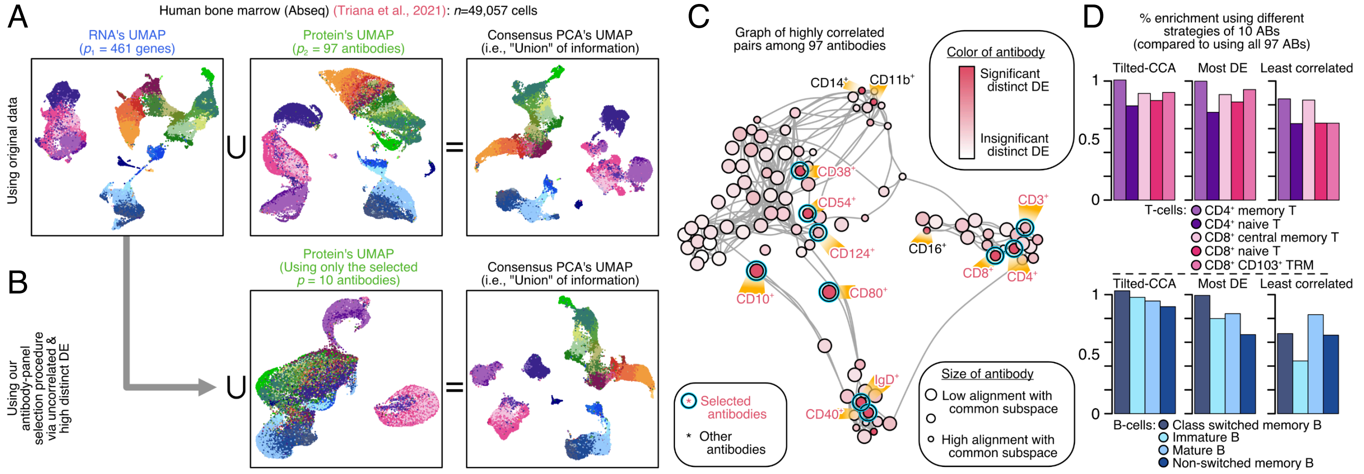

Fig. 4. Tilted-CCA enables targeted antibody panel design for RNA+protein multiome assays. (*A*) UMAP of the RNA and protein modalities, as well as the Consensus PCA capturing the union of information. (*B*) UMAP of the protein modality restricted to the 10 selected antibodies using our procedure showing poor separation among annotated cell types, and the resulting Consensus PCA when combined with the RNA modality showing many of the axes of variations compared to (*A*). (*C*) Correlation graph among the 97 antibodies, where the size of a node denotes the antibody's alignment with Tilted-CCA's common subspace and the color denotes the significance of the antibody's distinct component. All 10 selected antibodies are marked, as well as other conventional immune markers that were not selected by our procedure. (*D*) Percent enrichment of different cell types (either various CD4+/CD8+ T cells or B cells) of the Consensus PCA obtained by combining the RNA modality with different panels of 10 antibodies (obtained either using our Tilted-CCA, the most differentially expressed antibodies, or the least correlated with the transcriptome), when compared to the enrichment of the Consensus PCA obtained by combining the RNA modality with all 97 antibodies. A higher enrichment percent denotes less information is lost when using 10 antibodies.

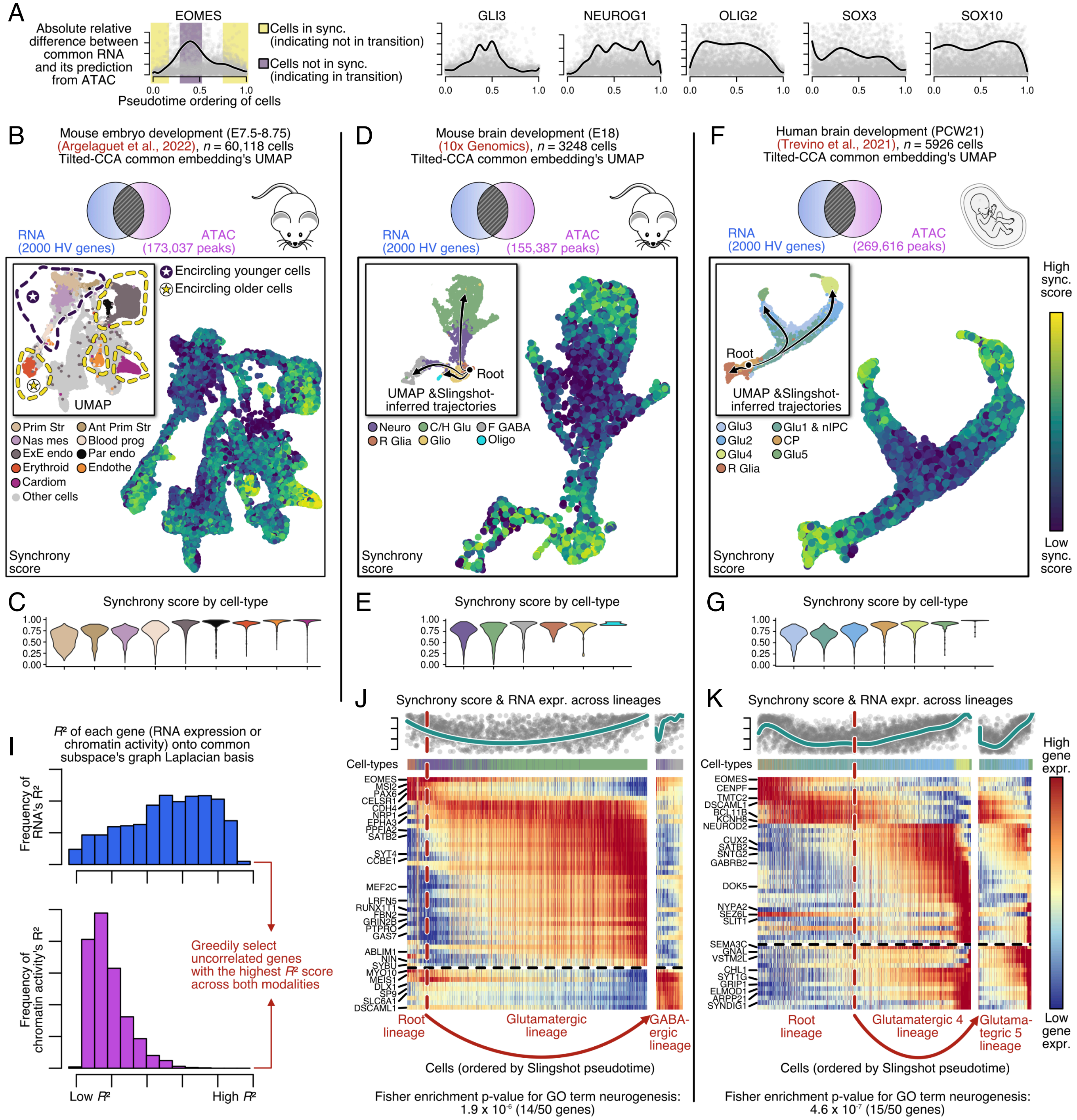

Fig. 5. Tilted-CCA infers the cell's developmental status and development-associated genes based on the common embedding between the RNA and ATAC modality. (*A*) Asynchrony between the ATAC and RNA measured by the residual of predicting the common RNA with the ATAC modality, plotted against Slingshot's pseudotime ordering of the cells in the glutamatergic 4 lineage of the human brain development dataset. (*B*) UMAP of a mouse embryo dataset where cells are colored based on annotated cell types (where regions of young cells are circled in purple, and regions of older cells are circled in yellow) or Tilted-CCA's synchrony scores. (*C*) Corresponding violin plot of the synchrony scores across cell types for the mouse embryo development dataset. (*D*) UMAP of a mouse brain development system shown with the Slingshot trajectories where the cells are colored based on annotated cell types or Tilted-CCA's synchrony scores. (*E*) Corresponding violin plot of the synchrony scores across cell types for the mouse brain development dataset. (*F*) UMAP of a human brain development system, analogous to (*D*). (*G*) Corresponding violin plot of the synchrony scores across cell types for the human brain development dataset. (*H*) Schematic illustrating the selection procedure for development-associated genes. (*I* and *J*) Heatmaps of selected genes and cells in the glutamatergic lineage or the GABAergic lineage in the mouse brain development system or selected genes and cells in the glutamatergic 4 or 5 lineage in the human brain development system. The genes in (*J*) and (*K*) are ordered to visualize the developmental cascade, and the exact Fisher enrichment of the enrichment of the selected 50 genes for the GO term "neurogenesis" is marked. The cells are ordered based on the pseudotimes estimated by Slingshot for visual clarity, and the pseudotimes are not needed when constructing the synchrony scores or selecting the development-associated genes.

the common embedding between RNA and ATAC captures the differentiation trajectory and that large deviations from the common embedding capture the asynchrony between RNA-level and chromatin-level signals. The coordinated yet asynchronous change between chromatin remodeling and gene transcription has been characterized by numerous recent studies on transcription priming (23, 41–45). To measure the synchrony of the RNA and ATAC modalities along a latent developmental trajectory (without knowledge of that trajectory), we design a linear regression to predict the RNA common-space component of each gene using the ATAC modality, where large absolute residuals indicate a lack of synchrony between that gene's RNA and ATAC (*Methods*). The absolute residuals of this prediction are plotted against developmental pseudotime for specific genes relevant to cortical neurogenesis based on a 10x Multiome dataset of developing human brain (37) (Fig. 5*A*). Note that the pseudotime is used only for comparison and not for computing the regression. We see that, indeed, for the neurogenesis-related genes, the residuals are small in terminal cells (i.e., pseudotime close to 1) and large for cells in transition. Encouraged, we design a cell-wise synchrony score, which summarizes the magnitude of these residuals across all genes for each cell, to distinguish between cells in terminal versus transient state without the reliance on trajectory reconstruction and pseudotime estimation (*Methods*).

We apply the synchrony score to three 10x Multiome datasets of developing tissues. First, consider the developing mouse embryo (46), equipped with cell-type labels delineating the youngest cell types such as primitive streak and blood progenitors as well as the terminal cell types such as cardiomyocytes and erythroids. We see that the synchrony scores are indeed low for the former group while high for the latter group (Fig. 5 *B* and *C*). Next, consider the developing mouse brain, where cell-type labels were transferred from an independent RNA reference (47) using SAVERCAT (48). Here, cell lineages originate from the radial glia and differentiate into oligodendrocytes, cortical/hippocampal glutamatergic, and forebrain GABAergic cells. We see that both the radial glia and the cells at terminal fates have high synchrony scores (Fig. 5 *D* and *E*). In contrast, the neuroblast cells and cells in the earlier stages of glutamatergic differentiation have lower synchrony scores. Last, consider the developing human brain (37). Here, development originates from the cycling progenitors and differentiates into either the radial glia or different types of cortical glutamatergic neurons.

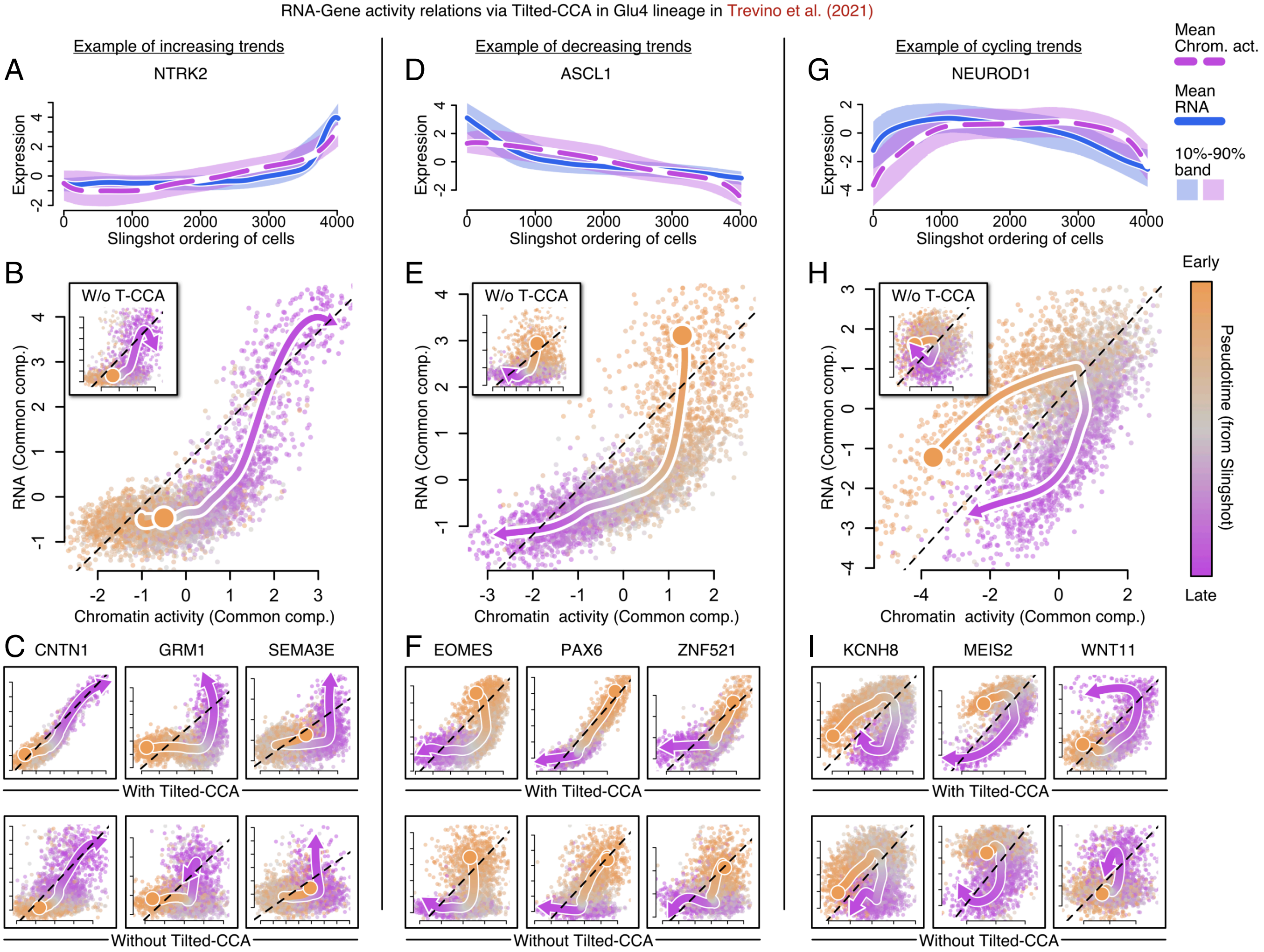

Fig. 6. The relation between a gene's expression and cis-chromatin activity is clarified by Tilted-CCA. (*A*) Time series showing the mean gene expression and cis-chromatin activity across pseudotime after applying Tilted-CCA for NTRK2, illustrating increasing expressions in both modalities. (*B*) Phase portrait showing the mean gene expression plotted against cis-chromatin activity for NTRK2 after applying Tilted-CCA, where cells are colored by the pseudotime. A corresponding inset shows the phase portrait if Tilted-CCA were not performed, illustrating noisier relationships between the two modalities. (*C*) Additional phase portraits for other important genes with increasing trends for both gene expression and cis-chromatin activity, where the phase portraits are shown with (*Top*) and without (*Bottom*) applying Tilted-CCA. (*D, E, F*) Plots for ASCL1, illustrating decreasing expressions in both modalities, analogous to (*A, B, C*) respectively. (*G, H, I*) Plots for NEUROD1, illustrating a cycling relation between both modalities, analogous to (*A, B, C*) respectively. The genes NTRK2, ASCL1 and NEUROD1 in (*A, B, D, E, G, H*) are showcased due to their importance for cortical neurogenesis.

The synchrony scores are high for both the cycling progenitors, radial glia, and the mature terminal cell types (Fig. 5 *F* and *G*) and low for the cells in transition. Thus, the Tilted-CCA synchrony score reliably distinguished between cells in a transient or terminal cell state for all three datasets from developing tissues. Notably, the synchrony score uses RNA-ATAC relationships not used by existing methods and does not require the a priori estimation of cell trajectory or RNA velocity, hence providing orthogonal information that complements existing methods.

We move on to the second question—without a priori knowledge of the developmental trajectory, can we identify the development-associated genes and characterize the relationship between their ATAC-derived chromatin activity and their RNA expression? Here, we define chromatin activity as the total read coverage in peak regions from the ATAC modality that is ±500 base pairs from the gene's transcription start sites. We apply Tilted-CCA to these RNA and chromatin activity modalities, yielding embeddings similar to those in Fig. 5 *D* and *F* (*SI Appendix*, Figs. S8 *C* and *E* and S9*A*). Based on the premise that development-associated processes should be captured in Tilted-CCA's common embedding, we measure how well each gene's expression and chromatin activity conform to the geometry of the common embedding's nearest-neighbor graph. To get a representative list of gene markers for all stages and branches of development, we greedily select genes that highly conform to the common embedding's geometry in both modalities and are uncorrelated with each other (Fig. 5*H*, *Methods*). The 50 genes identified in this way for both the mouse and human developing brain systems are highly enriched for neurogenesis and display varied expression profiles across pseudotime and lineages (Fig. 5 *I* and *J*, and *SI Appendix*, Fig. S9*B*). Notably, the ordering of cells in the shown heatmaps is chosen based on Slingshot's pseudotime and is used only for visualization but not selection. We also show the synchrony score of each cell against its ordered pseudotime, indicating that, as expected, the synchrony is indeed low for the transitioning cells. These findings demonstrate that Tilted-CCA's common embedding provides an alternative method of selecting development-associated genes that do not require the prior estimation of the developmental trajectory (30, 49).

Among the development-associated genes, we observe a diverse collection of relations between a gene's expression and its chromatin activity, and importantly, the cross-modal relationship is clarified in Tilted-CCA's common embedding. As expected, for many genes activated during development, their chromatin-level activity precedes their expression activation. An example gene for this along the glutamatergic 4 (Glu4) lineage in the human embryonic brain, such as NTRK2 (Fig. 6 *A* and *B*). This "priming" effect can be visualized as a time series across pseudotime, where chromatin activity increases gradually, preceding and foretelling the steep increase in gene expression much later (Fig. 6*A*). The phase portrait for the common components of RNA expression versus chromatin activity makes this relationship more apparent with a chromatin-activity to RNA curve that "runs before rising" (Fig. 6*B*). Next, consider a gene whose expression decreases over development, such as ASCL1. We find that for these genes, chromatin activity drops much more gradually than RNA expression (Fig. 6 *D* and *E*). The cis-peaks remain open after transcription of the gene has terminated, hinting at short-term cellular memory of the previous state. Last, and perhaps most curiously, we see evidence of genes that do not follow these aforementioned patterns but instead have a cyclical trend (Fig. 6 *G* and *H*). Phase portraits for the other selected genes in Fig. 5*K* are shown in *SI Appendix*, Fig. S8*C*. In almost all cases, the relationship between RNA expression and chromatin activity is clarified in the Tilted-CCA common embedding. This is shown by comparing Fig. 6 *B*, *E*, and *H* to their insets.

Discussion

Tilted-CCA extends CCA to decompose multimodal data into a sum of axes of variations that are shared between modalities and patterns that are distinct to each modality. We demonstrate the utility of Tilted-CCA on single-cell data, where we analyze multiomic datasets measuring the RNA and protein modalities or the RNA and ATAC modalities. For multimodal RNA and protein data, we show that Tilted-CCA aids in the design of targeted antibody panels that are complementary to the RNA modality. For multimodal RNA and ATAC data, we show how Tilted-CCA's common embedding, which captures variation supported by both the RNA and ATAC modalities, can be used to estimate quantities related to cellular differentiation.

Tilted-CCA complements existing dimension-reduction methods for multiomic datasets because Tilted-CCA captures the intersection of information between modalities while existing methods aim to capture the union of information. While these "union"-type methods, such as Consensus PCA, WNN, scAI, and MOFA+, help aggregate information across modalities to improve cell-type discovery, these methods are not suitable for understanding how the two modalities are related to each other. The decomposition of multimodal data into common and distinct components is a fundamentally different question requiring new statistical methodology. Tilted-CCA answers this question by first defining what is common information via the shared geometry of the two high-dimensional matrices and then adapting the linear framework of CCA to encapsulate this information. Importantly, this linear framework enables biologists to address cell-centric or variable-centric questions in downstream analyses.

Materials and Methods

Tilted-CCA consists of five steps: 1) constructing nearest-neighbor graphs from each modality, 2) defining the target common manifold that encodes the common geometry between the two modalities, 3) performing CCA, 4) optimizing the appropriate tilt in the decomposition of CCA's canonical score matrices such that the common space approximates the target common manifold, and 5) deriving the final decomposition to estimate $C, D^{(1)}$, and $D^{(2)}$.

Let $X^{(1)} \in \mathbb{R}^{n\times p_1}$ and $X^{(2)} \in \mathbb{R}^{n\times p_2}$ be the two matrices that form the multiomic dataset, where the same n cells are measured across p_1 features in Modality 1 and p_2 features in Modality 2. We assume that both $X^{(1)}$ and $X^{(2)}$ are preprocessed beforehand to have centered features. We also assume that both $X^{(1)}$ and $X^{(2)}$ are low-rank, i.e., $\text{rank}(X^{(1)}) = \text{rank}(X^{(2)}) = r$, where $r \leq \min\{n, p_1, p_2\}$. See *SI Appendix* for a generalization of this assumption and its discussion.

Step 1: Construct Nearest-Neighbor Graphs. The intent of the first step is to quantify what cell-separation information is contained within each modality. Specifically, Tilted-CCA constructs two nearest-neighbor graphs, one for each $X^{(1)}$ and $X^{(2)}$, the data matrix for each modality. Let k denote the number of nearest neighbors, where $k \ll n$. Let the SVD of $X^{(1)}$ and $X^{(2)}$ be denoted as

$$X^{(1)} = U^{(1)}\Lambda^{(1)}(V^{(1)})^\top, \quad \text{and} \quad X^{(2)} = U^{(2)}\Lambda^{(2)}(V^{(2)})^\top, \qquad [1]$$

where, for both $\ell \in \{1, 2\}$, $U^{(\ell)} \in \mathbb{R}^{n\times r_\ell}$ and $V^{(\ell)} \in \mathbb{R}^{p_\ell\times r_\ell}$ are orthonormal matrices, and $\Lambda^{(\ell)} \in \mathbb{R}^{r_\ell\times r_\ell}$ are diagonal matrices. Consider the low-dimensional embeddings $U^{(1)}\Lambda^{(1)} \in \mathbb{R}^{n\times r_1}$ and $U^{(2)}\Lambda^{(2)} \in \mathbb{R}^{n\times r_2}$. We normalize each cell's embedding by constructing matrices $P^{(\ell)} \in \mathbb{R}^{n\times r_\ell}$ where

$$P^{(\ell)}_{i,\cdot} = \left[U^{(\ell)}\Lambda^{(\ell)}\right]_{i,\cdot} / \left\| \left[U^{(\ell)}\Lambda^{(\ell)}\right]_{i,\cdot} \right\|_2$$

for $\ell \in \{1, 2\}$ and $i \in \{1, \ldots, n\}$. Then, we construct two k-nearest-neighbor graphs $G^{(1)}, G^{(2)} \in \{0, 1\}^{n\times n}$, where these graphs are represented as symmetric binary adjacency matrices among all n cells. See *SI Appendix* for its explicit construction.

The benefit of using nearest-neighbor graphs is that they are flexible mathematical objects to represent the cell-separation information in each modality. They also enable a principled procedure to define the information shared between both modalities in the next step.

Step 2: Construct Target Common Manifold. The intent of the second step is to define what information is shared between the two modalities. Here, Tilted-CCA determines the target common manifold based on the two aforementioned nearest-neighbor graphs $G^{(1)}$ and $G^{(2)}$. This target common manifold $G \in \{0, 1\}^{n\times n}$ is also represented as a symmetric binary adjacency matrix. Ideally, the manifold G (represented as a graph) enumerates which groups of cells are far in both modalities (i.e., both modalities agree that two groups of cells are distinct). Stated differently, as an example, if two groups of cells are far in Modality 1 (dictated by $G^{(1)}$) but are close in Modality 2 (dictated by $G^{(2)}$), we would intuit this as information unique to Modality 1 and, hence, would not this cell-separation information be encoded in G. We design the construction of G to operationalize this idea. This construction of G is at the heart of Tilted-CCA since afterward, Tilted-CCA will estimate a linear decomposition that best approximates G's nearest-neighbor structure in Step 3.

We have two different procedures for constructing G, based on whether or not a clustering is provided a priori for each of $X^{(1)}$ and $X^{(2)}$.

- Scenario with no global clustering structure: When no global structure is provided, constructing G is relatively straightforward. Intuitively, we include an edge between two cells in G if there is an edge present between the two cells in either $G^{(1)}$ or $G^{(2)}$.
- Scenario with global clustering structure: When global structure is provided, the procedure to construct G is more involved. For cell i, we use a quadratic optimization program to determine how many edges from $G^{(1)}$ or $G^{(2)}$ to downsample so both modalities contribute the same amount of information.

For example, the former setting is more suitable for developmental datasets where this is a smooth continuum of cells in $X^{(1)}$ and $X^{(2)}$, meaning there are no well-defined clusters. The latter setting is more suitable for PBMCs, where immune cells form prominent clusters of cells. See *SI Appendix* for further motivation and explicit mathematical construction of G. Intuitively, if Modality 2 contains little geometric information in $G^{(2)}$ compared to Modality 1, then G should also not contain much geometric information. We verify this via pseudoreal experiments based on the CITE-seq bone marrow dataset (18). There, we artificially make the protein modality have less and less cell-type separation and demonstrate the impacts on G, the resulting Tilted-CCA embedding, and the downstream alignment scores (*SI Appendix*, Fig. S12).

While G captures the information shared between both modalities, this graph does not address feature-level questions (for example, which genes coordinate the most with the other modality?). Hence, the following steps aim to estimate a low-dimensional embedding whose common component best approximates the cell-separation information in G.

Step 3: Perform CCA. The intent of the third step is to compute the CCA between $X^{(1)}$ and $X^{(2)}$, as the resulting canonical score matrices enable a principled decomposition in the framework of Fig. 1*B*. We briefly review CCA here to lay out the notation. Let $\Sigma^{(1)} = (X^{(1)})^\top X^{(1)}/n$, $\Sigma^{(2)} = (X^{(2)})^\top X^{(2)}/n$ and $\Sigma^{(12)} = (X^{(1)})^\top X^{(2)}/n$. Recall that $\text{rank}(X^{(1)}) = \text{rank}(X^{(2)}) = r$. Then, a rank-$r$ CCA solves the optimization problem:

$$\{A, B\} = \underset{\substack{A' \in \mathbb{R}^{p_1\times r} \\ B' \in \mathbb{R}^{p_2\times r}}}{\arg\max} \; \text{tr}\left((A')^\top \Sigma^{(12)} (B')\right) \qquad [2]$$

$$\text{such that } (A')^\top \Sigma^{(1)} (A') = I_r \quad \text{and} \quad (B')^\top \Sigma^{(2)} (B') = I_r,$$

where I_r is the r-dimensional identity matrix. Equipped with the solution to Eq. **2**, we can compute the canonical score matrices for $\ell \in \{1, 2\}$,

$$Z^{(\ell)} = X^{(\ell)} A/\sqrt{n}. \qquad [3]$$

Note that by the identity constraints of CCA in Eq. **2**, $Z^{(\ell)}$ is an orthonormal matrix, i.e., $\|Z^{(\ell)}_{\cdot,j}\|_2 = 1$ for all $j \in \{1, \ldots, r\}$. See *SI Appendix* for a more thorough review of the explicit solution of Eq. **2**.

Step 4: Optimize the Common and Distinct Scores. The intent of the fourth step is to decompose the canonical scores $Z^{(1)}$ and $Z^{(2)}$ into a common score matrix C (also called the common embedding) and the distinct score matrices $D^{(\ell)} \in \mathbb{R}^{n\times r}$ for Modality $\ell \in \{1, 2\}$ (also called the distinct embeddings), in such a way that the geometric relations among the n cells in C best approximate those in the target common manifold G. Additionally, Tilted-CCA imposes the constraint that for $\ell \in \{1, 2\}$,

$$Z^{(\ell)} = C + D^{(\ell)},$$

where $(D^{(1)})^\top D^{(2)} = 0 \in \mathbb{R}^{r\times r}$.

To construct C, $D^{(1)}$, and $D^{(2)}$, we first describe how the proposed tilts for each latent dimension's common vector enables us to construct these three matrices. The construction is done column-wise, for latent dimension $j \in \{1, \ldots, r\}$. Consider a tilt $\tau_j \in [0, 1]$, where if $\tau_j = 0$ then $C_{\cdot,j} = Z^{(1)}_{\cdot,j}$ and if $\tau_j = 1$ then $C_{\cdot,j} = Z^{(2)}_{\cdot,j}$. Specifically, there is a unique construction of $C_{\cdot,j}, D^{(1)}_{\cdot,j}, D^{(2)}_{\cdot,j} \in \mathbb{R}^n$ based on $Z^{(1)}_{\cdot,j}, Z^{(2)}_{\cdot,j} \in \mathbb{R}^n$ and $\tau_j \in [0, 1]$, such that

$$\begin{aligned}
&C_{\cdot,j} \in \text{span}\left\{Z^{(1)}_{\cdot,j}, Z^{(2)}_{\cdot,j}\right\} \\
&\cos^{-1}\left((C_{\cdot,j})^\top (Z^{(1)}_{\cdot,j})\right)/\|C_{\cdot,j}\|_2 = \tau_j \cdot \cos^{-1}\left((Z^{(2)}_{\cdot,j})^\top (Z^{(1)}_{\cdot,j})\right), \\
&C_{\cdot,j} + D^{(1)}_{\cdot,j} = Z^{(1)}_{\cdot,j}, \quad \text{and} \quad C_{\cdot,j} + D^{(2)}_{\cdot,j} = Z^{(2)}_{\cdot,j}, \\
&(D^{(1)}_{\cdot,j})^\top (D^{(2)}_{\cdot,j}) = 0, \\
&\|C_{\cdot,j}\|_2 \le 1.
\end{aligned}$$

The first relation ensures that we are only considering common vectors in the same hyperplane as the two canonical score vectors–this ensures that the resulting column vectors C are orthogonal after this construction is complete. The second relation is why we call τ_j the "tilt"–we are ensured that $C_{\cdot,j}$ is a vector that has an angle of τ_j-percent between $Z^{(1)}_{\cdot,j}$ and $Z^{(2)}_{\cdot,j}$. The third and fourth relations ensure a valid decomposition of $Z^{(1)}_{\cdot,j}$ and $Z^{(2)}_{\cdot,j}$. The fifth equation ensures that the two distinct vectors $D^{(1)}_{\cdot,j}$ and $D^{(2)}_{\cdot,j}$ are orthogonal, which is why we call these the "distinct" vectors. The last relation ensures a unique decomposition. The vectors $C_{\cdot,j}$, $D^{(1)}_{\cdot,j}$, and $D^{(2)}_{\cdot,j}$ can be constructed to satisfy these constraints using straightforward geometry and linear algebra. (See Fig. 1*H* for the intuition on the details of the calculation).

Next, we describe how we measure the quality of a proposed tilt τ_j based on its similarity to the target common manifold G. (This is reflected by "Measure similarity between two graphs" in Fig. 2*C*.) Consider the common manifold G described in Step 2 and the k-nearest-neighbor graph constructed from C with tilt τ_j, denoted as $G^{(C;\tau_j)}$. Equipped with G and $G^{(C;\tau_j)}$, our similarity is defined by the Grassmannian distance between two sets of eigenvectors, each derived from the normalized random-walk Laplacian graph basis (50) for either of these graphs. Specifically, considering the common manifold G, let $\widetilde{\Delta} \in \mathbb{R}^{n\times n}$ denote the degree matrix which is a diagonal matrix where

$$\widetilde{\Delta}_{i,i} = \sum_{j=1}^{n} G_{ij}, \quad \text{for all } i \in \{1, \ldots, n\},$$

and define the normalized Laplacian as $L' = \widetilde{\Delta}^{-1} G \widetilde{\Delta}^{-1}$, and then define the normalized random-walk Laplacian as

$$L = (\widetilde{\Delta}')^{-1} L' \in \mathbb{R}^{n\times n},$$

for $\widetilde{\Delta}'$ is a diagonal matrix where

$$\widetilde{\Delta}'_{i,i} = \sum_{j=1}^{n} L'_{ij} \quad \text{for all } i \in \{1, \ldots, n\}.$$

Using this construction, let W and $W^{(C;\tau_j)}$ be the leading-K_L eigenvectors of L and $L^{(C;\tau_j)}$ normalized random-walk Laplacian basis matrices constructed from the graphs G and $G^{(C;\tau_j)}$, respectively, for some tuning parameter K_L. Since both W and $W^{(C;\tau_j)}$ are points along the Grassmannian manifold, we use the Grassmannian distance (51, 52) to measure the distance between W and $W^{(C;\tau_j)}$. Specifically, consider the SVD of

$$W^\top W^{(C;\tau_j)} = UDV^\top \in \mathbb{R}^{K_L \times K_L},$$

where $D = \mathrm{diag}(\sigma_1, \ldots, \sigma_{K_L})$. The Grassmannian distance is then defined as

$$\Bigg(\sum_{i=1}^{K_L} \big(\cos^{-1}(\sigma_i)\big)^2\Bigg)^{1/2}. \qquad [4]$$

A smaller Grassmannian distance implies a higher similarity between G and $G^{(C;\tau_j)}$, and we wish to find the τ_j that minimizes this Grassmannian distance.

We now have all the necessary ingredients for an optimization procedure. After an appropriate initialization of all the tilts $\tau_1, \ldots, \tau_r$, we use a zero-order cyclical (either using Nelder–Mead or a grid-search) optimization over $\tau_j \in [0, 1]$ for each latent dimension $j \in \{1, \ldots, r\}$ to minimize the distance between the target manifold G and the common embedding $G^{(C;\tau_j)}$, and we cycle through each latent dimension iteratively (i.e., many epochs) until we reach convergence or a maximal epoch limit. See *SI Appendix* for additional statistical perspectives on this step.

Step 5: Compute the Final Decomposition. The last step is to determine the decomposition of $X^{(1)}$ and $X^{(2)}$ based on the common and distinct scores. Specifically, for Modality $\ell \in \{1, 2\}$, let

$$X^{(\ell,C)} = \Big[Z^{(\ell)} C^\top\Big] \cdot \Big[U^{(\ell)} \Lambda^{(\ell)} (V^{(\ell)})^\top\Big] \in \mathbb{R}^{n\times p_\ell},$$
$$X^{(\ell,D)} = \Big[Z^{(\ell)} (D^{(\ell)})^\top\Big] \cdot \Big[U^{(\ell)} \Lambda^{(\ell)} (V^{(\ell)})^\top\Big] \in \mathbb{R}^{n\times p_\ell}. \qquad [5]$$

This is a sensible construction, as $Z^{(\ell)}(Z^{(\ell)})^\top \in \mathbb{R}^{n\times n}$ is a projection matrix since $Z^{(\ell)} \in \mathbb{R}^{n\times r}$ is an orthonormal matrix and $C + D^{(\ell)} = Z^{(\ell)}$, while $U^{(\ell)}\Lambda^{(\ell)}(V^{(\ell)})^\top$ is each modality's low-rank expression matrix. Hence, by this construction, we are ensured that

$$X^{(\ell,C)} + X^{(\ell,D)} = X^{(\ell)},$$

which justifies why this is a decomposition and that $X^{(1,C)}$ and $X^{(2,C)}$ share the same row space, which justifies why this is called the common component of the decomposition.

Data, Materials, and Software Availability. Tilted-CCA software package, scripts to reproduce the processing, analysis, and figure-generation data have been deposited in GitHub (https://github.com/linnykos/tiltedCCA (53) and https://github.com/linnykos/tiltedCCA_analysis) (54). Previously published data were used for this work. (All data used in this manuscript are publicly available). Specifically, the human bone marrow CITE-seq data (18) are available in the R package SeuratData under "bmcite". The human bone marrow Abseq data (20) (both, either equipped with the whole transcriptome or of only 461 genes) are available at https://figshare.com/projects/Single-cell_proteo-genomic_reference_maps_of_the_human_hematopoietic_system/94469 (55), under WTA_projected.rds or Healthy.rds, respectively. The human PBMC CITE-seq data (11) are available at https://atlas.fredhutch.org/nygc/multimodal-pbmc/ (56). The human PBMC 10x Multiome data are available at https://support.10xgenomics.com/single-cell-multiome-atac-gex/datasets/1.0.0/pbmc_granulocyte_sorted_10k (57). The human brain development 10x Multiome data (37) are available at https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE162170 (58) and https://github.com/GreenleafLab/brainchromatin (59). The mouse brain development 10x Multiome data are available at https://www.10xgenomics.com/resources/datasets/fresh-embryonic-e-18-mouse-brain-5-k-1-standard-2-0-0 (60). The mouse embryo development 10x Multiome data (46) are available at https://github.com/rargelaguet/mouse_organogenesis_10x_multiome_publication (61). The majority of the computational tools used in this manuscript consisted of Seurat (v4.1.1), Signac (v1.7.0), and slingshot (v2.3.1). We derived the cell cycle genes for humans via Seurat::cc.genes, while we used the lists of human housekeeping genes (62) or mouse cell cycling genes (63) from particular published work.

ACKNOWLEDGMENTS. We thank Bernie Devlin, Paul Hess, Mingyao Li, Zongming Ma, Divij Matthew, Kathryn Roeder, William Townes, and Chi-Yun Wu for helpful comments, suggestions, and feedback on the method development and analyses. This work is supported by the National Human Genome Research Institute Grant 5R01HG006137 and the National Institute of General Medical Sciences Grant 5R01GM125301. K.Z.L. is further supported by the Wharton School Dean's Postdoctoral Fellowship.

Author affiliations: [a]Department of Statistics and Data Science, University of Pennsylvania, Philadelphia, PA 19104

1. S. Teichmann, M. Efremova, Method of the year 2019: Single-cell multimodal omics. *Nat. Methods* **17**, 2020 (2020).
2. M. Eisenstein, The secret life of cells. *Nat. Methods* **17**, 7–10 (2020).
3. C. Zhu, S. Preissl, B. Ren, Single-cell multimodal omics: The power of many. *Nat. Methods* **17**, 11–14 (2020).
4. Ma. Anjun *et al.*, Integrative methods and practical challenges for single-cell multi-omics. *Trends Biotechnol.* **38**, 1007–1022 (2020).
5. M. Civelek, A. J. Lusis, Systems genetics approaches to understand complex traits. *Nat. Rev. Genet.* **15**, 34–48 (2014).
6. K. A. Hoadley *et al.*, Multiplatform analysis of 12 cancer types reveals molecular classification within and across tissues of origin. *Cell* **158**, 929–944 (2014).
7. Y. V. Sun, Y. J. Hu, Integrative analysis of multi-omics data for discovery and functional studies of complex human diseases. *Adv. Genet.* **93**, 147–190 (2016).
8. K. Yugi, H. Kubota, A. Hatano, S. Kuroda, Trans-omics: How to reconstruct biochemical networks across multiple omic-layers. *Trends Biotechnol.* **34**, 276–290 (2016).
9. Y. Hasin, M. Seldin, A. Lusis, Multi-omics approaches to disease. *Genome Biol.* **18**, 1–15 (2017).
10. E. F. Lock, K. A. Hoadley, J. S. Marron, A. B. Nobel, Joint and individual variation explained (JIVE) for integrated analysis of multiple data types. *Ann. Appl. Stat.* **7**, 523 (2013).
11. Y. Hao *et al.*, Integrated analysis of multimodal single-cell data. *Cell* **184**, 3573–3587 (2021).
12. R. Argelaguet *et al.*, MOFA+: A statistical framework for comprehensive integration of multi-modal single-cell data. *Genome Biol.* **21**, 1–17 (2020).
13. S. Jin, L. Zhang, Q. Nie, scAI: An unsupervised approach for the integrative analysis of parallel single-cell transcriptomic and epigenomic profiles. *Genome Biol.* **21**, 1–19 (2020).
14. Y. Ma, Z. Sun, P. Zeng, W. Zhang, Z. Lin, JSNMF enables effective and accurate integrative analysis of single-cell multiomics data. *Briefings Bioinf.* **23**, bbac105 (2022).

15. H. Hotelling, *Relations Between Two Sets of Variates in Breakthroughs in Statistics* (Springer, 1992), pp. 162–190.
16. M. Stoeckius *et al*., Simultaneous epitope and transcriptome measurement in single cells. *Nat. Methods* **14**, 865–868 (2017).
17. P. Shahi, S. C. Kim, J. R. Haliburton, Z. J. Gartner, A. R. Abate, Abseq: Ultrahigh-throughput single cell protein profiling with droplet microfluidic barcoding. *Sci. Rep.* **7**, 1–12 (2017).
18. T. Stuart *et al*., Comprehensive integration of single-cell data. *Cell* **177**, 1888–1902 (2019).
19. A. Gayoso *et al*., Joint probabilistic modeling of single-cell multi-omic data with totalVI. *Nat. Methods* **18**, 272–282 (2021).
20. S. Triana *et al*., Single-cell proteo-genomic reference maps of the hematopoietic system enable the purification and massive profiling of precisely defined cell states. *Nat. Immunol.* **22**, 1577–1589 (2021).
21. J. Cao *et al*., Joint profiling of chromatin accessibility and gene expression in thousands of single cells. *Science* **361**, 1380–1385 (2018).
22. S. Chen, B. B. Lake, K. Zhang, High-throughput sequencing of the transcriptome and chromatin accessibility in the same cell. *Nat. Biotechnol.* **37**, 1452–1457 (2019).
23. Ma. Sai *et al*., Chromatin potential identified by shared single-cell profiling of RNA and chromatin. *Cell* **183**, 1103–1116 (2020).
24. C. Zhu *et al*., An ultra high-throughput method for single-cell joint analysis of open chromatin and transcriptome. *Nat. Struct. Mole. Biol.* **26**, 1063–1070 (2019).
25. C. Trapnell *et al*., The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nat. Biotechnol.* **32**, 381 (2014).
26. K. Street *et al*., Slingshot: Cell lineage and pseudotime inference for single-cell transcriptomics. *BMC Genomics* **19**, 477 (2018).
27. G. La Manno *et al*., RNA velocity of single cells. *Nature* **560**, 494–498 (2018).
28. W. Saelens, R. Cannoodt, H. Todorov, Y. Saeys, A comparison of single-cell trajectory inference methods. *Nat. Biotechnol.* **37**, 547–554 (2019).
29. M. Lange *et al*., Cell Rank for directed single-cell fate mapping. *Nat. Methods* **19**, 159–170 (2022).
30. K. Van den Berge *et al*., Trajectory-based differential expression analysis for single-cell sequencing data. *Nat. Commun.* **11**, 1–13 (2020).
31. H. R. de Bézieux, K. Van den Berge, K. Street, S. Dudoit, Trajectory inference across multiple conditions with condiments: Differential topology, progression, differentiation, and expression. bioRxiv (2021). https://www.biorxiv.org/content/10.1101/2021.03.09.433671 (Accessed 2 March 2023).
32. M. Tedesco *et al*., Chromatin velocity reveals epigenetic dynamics by single-cell profiling of heterochromatin and euchromatin. *Nat. Biotechnol.* **40**, 235–244 (2022).
33. J. A. Westerhuis, T. Kourti, J. F. MacGregor, Analysis of multiblock and hierarchical PCA and PLS models. *J. Chem. Soc.* **12**, 301–321 (1998).
34. H. Abdi, L. J. Williams, D. Valentin, Multiple factor analysis: Principal component analysis for multitable and multiblock data sets. *Wiley Interdiscip. Rev. Comput. Stat.* **5**, 149–179 (2013).
35. M. Tenenhaus, A. Tenenhaus, P. J. Groenen, Regularized Consensus PCA. arXiv [Preprint] (2015). http://arxiv.org/abs/1504.07005 (Accessed 2 March 2023).
36. H. Shu, X. Wang, H. Zhu, D-CCA: A decomposition-based canonical correlation analysis for high-dimensional datasets. *J. Am. Stat. Assoc.* **115**, 292–306 (2020).
37. A. E. Trevino *et al*., Chromatin and gene-regulatory dynamics of the developing human cerebral cortex at single-cell resolution. *Cell* **184**, 5053–5069 (2021).
38. R. S. Ziffra *et al*., Single-cell epigenomics reveals mechanisms of human cortical development. *Nature* **598**, 205–213 (2021).
39. G. S. Gulati *et al*., Single-cell transcriptional diversity is a hallmark of developmental potential. *Science* **367**, 405–411 (2020).
40. V. Bergen, M. Lange, S. Peidli, F. A. Wolf, F. J. Theis, Generalizing RNA velocity to transient cell states through dynamical modeling. *Nat. Biotechnol.* **38**, 1408–1414 (2020).
41. C. Li, M. Virgilio, K. L. Collins, J. D. Welch, "Single-cell multi-omic velocity infers dynamic and decoupled gene regulation" in *International Conference on Research in Computational Molecular Biology*, (Springer, 2022), pp. 297–299.
42. C. Bonifer, P. N. Cockerill, Chromatin priming of genes in development: Concepts, mechanisms and consequences. *Exp. Hematol.* **49**, 1–8 (2017).
43. A. Wang *et al*., Epigenetic priming of enhancers predicts developmental competence of hESC-derived endodermal lineage intermediates. *Cell Stem Cell* **16**, 386–399 (2015).
44. N. Tiwari *et al*., Stage-specific transcription factors drive astrogliogenesis by remodeling gene regulatory landscapes. *Cell Stem Cell* **23**, 557–571 (2018).
45. V. P. Schulz *et al*., A unique epigenomic landscape defines human erythropoiesis. *Cell Rep.* **28**, 2996–3009 (2019).
46. R. Argelaguet *et al*., Decoding gene regulation in the mouse embryo using single-cell multi-omics. bioRxiv (2022). https://www.biorxiv.org/content/10.1101/2021.03.09.433671 (Accessed 2 March 2023).
47. G. La Manno *et al*., Molecular architecture of the developing mouse brain. *Nature* **596**, 92–96 (2021).
48. M. Huang, Z. Zhang, N. R. Zhang, Dimension reduction and denoising of single-cell RNA sequencing data in the presence of observed confounding variables. bioRxiv (2020). https://www.biorxiv.org/content/10.1101/2021.03.09.433671 (Accessed 2 March 2023).
49. D. Song, J. J. Li, PseudotimeDE: Inference of differential gene expression along cell pseudotime with well-calibrated P-values from single-cell RNA sequencing data. *Genome Biol.* **22**, 1–25 (2021).
50. T. Shnitzer, M. Yurochkin, K. Greenewald, J. Solomon, Log-euclidean signatures for intrinsic distances between unaligned datasets. arXiv [Preprint] (2022). http://arxiv.org/abs/2202.01671 (Accessed 2 March 2023).
51. K. Ye, L. H. Lim, Schubert varieties and distances between subspaces of different dimensions. *SIAM J. Matrix Anal. Appl.* **37**, 1176–1197 (2016).
52. T. T. Cai *et al*., Rate-optimal perturbation bounds for singular subspaces with applications to high-dimensional statistics. *Ann. Stat.* **46**, 60–89 (2018).
53. K. Z. Lin, Tilted-CCA. Github. https://github.com/linnykos/tiltedCCA. Deposited 9 October 2022.
54. K. Z. Lin, Tilted-CCA Analysis. Github. https://github.com/linnykos/tiltedCCA_analysis. Deposited 20 June 2023.
55. L. Velten, S. Triana, D. Vonficht, Single-cell proteo-genomic reference maps of the human hematopoietic system. Figshare. https://figshare.com/projects/Single-cell_proteo-genomic_reference_maps_of_the_human_hematopoietic_system/94469. Accessed 16 May 2022.
56. T. Stuart *et al*., Integrated Analysis Of Multimodal Single-Cell Data. Fred Hutch. https://atlas.fredhutch.org/nygc/multimodal-pbmc/. Accessed 21 December 2021.
57. 10x Genomics, PBMC from a healthy donor - granulocytes removed through cell sorting (10k). Single cell multiome ATAC + Gene expression datasets. https://support.10xgenomics.com/single-cell-multiome-atac-gex/datasets/1.0.0/pbmc_granulocyte_sorted_10k. Accessed 28 December 2020.
58. A. E. Trevino *et al*., Chromatin and gene-regulatory dynamics of the developing human cerebral cortex at single-cell resolution. Gene Expression Omnibus. https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE162170. Accessed 16 March 2022.
59. A. E. Trevino, Brain Chromatin. Github. https://github.com/GreenleafLab/brainchromatin. Accessed 16 March 2022.
60. 10x Genomics, Fresh Embryonic E18 Mouse Brain (5k). 10x Genomics datasets. https://www.10xgenomics.com/resources/datasets/fresh-embryonic-e-18-mouse-brain-5-k-1-standard-2-0-0. Accessed 19 June 2021.
61. R. Argelaguet, Decoding gene regulation in the mouse embryo using single-cell multi-omics. Github. https://github.com/rargelaguet/mouse_organogenesis_10x_multiome_publication. Accessed 8 August 2022.
62. B. W. Hounkpe, F. Chenou, F. de Lima, E. V. De Paula, HRT atlas v1.0 database: Redefining human and mouse housekeeping genes and candidate reference transcripts by mining massive RNA-seq datasets. *Nucleic Acids Res.* **49**, D947–D955 (2021).
63. A. Riba *et al*., Cell cycle gene regulation dynamics revealed by RNA velocity and deep-learning. *Nature Commun.* **13**, 1–13 (2022).