SONATA: Disambiguated manifold alignment of single-cell data

Han Zhou[1] Kai Cao[2]
Yang Young Lu[1*]
[1]Cheriton School of of Computer Science, University of Waterloo
[2]Eric and Wendy Schmidt Center, Broad Institute of MIT and Harvard

Abstract

Rapid advancements in single-cell sequencing technologies have empowered biologists to collect high-resolution omics profiles at the cellular level. Integrating such heterogeneous single-cell data, essential for gaining deeper insights into biological processes, is a formidable challenge. The difficulty arises from the lack of shared cells or features across distinct measurements, compounded by biological and technical variations. To address this, manifold alignment methods have emerged as a crucial tool to align heterogeneous single-cell measurements in a shared space. However, current methods recklessly report an arbitrary alignment solution without considering whether the data can be aligned. In this paper, we introduce SONATA, a novel computational method that utilizes a novel cell-cell ambiguity measurement to gauge the ambiguity of cross-modality correspondences. SONATA identifies mutually ambiguous cell groups, characterized by similar geometric structures, and exhaustively enumerates alignment solutions between datasets. This approach distinguishes between cases with unique solutions or multiple alternatives, ensuring robust and interpretable integration of single-cell data. Finally, we validate the effectiveness of SONATA through extensive experiments on simulated and real datasets.

1 Introduction

Rapidly evolving single-cell sequencing technologies have enabled biologists to progressively collect unprecedented high-resolution omics profiles at cellular level. The existing sequencing assays aim to capture different aspects of the cell, including but not limited to, scRNA-seq for gene expression [1], scATAC-seq for chromatin accessibility [2], scMethyl-seq for DNA methylation [3], scHi-C for chromatin 3D conformation [4], and scChIC-seq for histone modifications [5]. Therefore, integrative analysis of such a wealth of single-cell measurements has attracted growing interests, potentially offering better understanding towards fundamental biological processes.

Despite the great promise, integrating single-cell datasets is nontrivial. One reason roots in the difficulty to obtain multiple measurements simultaneously from the same individual cell [6]. Due to the destructive measurement technology, the sets of measurements span disjoint sets of cells. For example, scATAC-seq and scHi-C both operate on the genomic DNA, thus each individual cell can only be measured once subject to the genome cleavage. The lack of connection between features acquired by multiple measurements impedes the integration, especially for measures capturing orthogonal properties of the genome. Another reason lies in variations between datasets, even for identical measurements collected differently [7]. These variations can be biological when the samples

*Corresponding Author. Email: yanglu@uwaterloo.ca

are collected under various conditions across different tissues, organs, individuals, or species. The variations can also be technical when the samples are profiled from multiple laboratories using distinct instruments and experimental protocols. In summary, the lack of shared cells or features across heterogeneous single-cell measurements makes integrating these data extremely challenging.

Among existing data integration solutions, *manifold alignment* is tailored for the setting where different data modalities are profiled in different cell subsets. Widely used manifold alignment methods [8, 9, 10] generally aim to identify a low-dimensional manifold that captures covariation across different data modalities. For example, the low-dimensional manifold can represent a differentiation trajectory from which cells are sampled. Manifold alignment methods attempt to elucidate their underlying manifold structures, identify the correspondence relationships between cells across data modalities, and consequently generate a joint cellular profile integrated from different datasets (Figure 1(A)). In this way, manifold alignment has played an important role in single-cell studies, such as creating large-scale cellular references of human organs and tissues [11].

Despite disparate implementation, existing manifold alignment methods all follow an implicit yet critical assumption that a shared latent manifold structure is preserved between data modalities. Following the assumption, these methods consistently report an alignment solution by identifying the cross-modality correspondence relationships between cells, even when two data modalities may not be unambiguously alignable. As a motivating example, consider cell cycle [12], a periodic biological process in human cells. The periodicity of the cycling cells at different transcriptomic states forms a circular trajectory, where the cell location at the trajectory indicates its timing in the cell cycle. In this case, countless (but wrong) alignments exist between two cycles because rotating a cycle by arbitrary degrees naturally leads to a new solution. Thus, without such a sanity check, blindly aligning single-cell datasets might lead to unfaithful and misleading interpretations [13].

Specifically, the alignability examination attempts to answer the following two questions: (1) "Can we determine if two datasets can possibly be aligned?", and (2) "Given the existence of alignability, can we ensure the alignment is unique to induce an unambiguous correspondence between datasets? " The former question discerns the possible failure of manifold alignment with no valid solution, which can be addressed by statistical testing [14]. The latter question tackles the likely ambiguity of manifold alignment with potentially multiple valid solutions, which, to our knowledge, has largely been ignored. From a practitioner's perspective, a desirable manifold alignment method is only scientifically valuable if a systematic strategy exists to distinguish whether the alignment has no solution, one unique solution, or multiple alternative solutions, rather than reporting an arbitrary alignment solution recklessly.

In this paper, we focus on the problem of manifold alignment ambiguity and propose a novel computational method named SONATA (diSambiguated manifOld aligNment of single-cell dATA), as illustrated in Figure 1(B). At a high level, SONATA first defines a cell-cell substitution matrix for a given dataset, describing the likelihood that one cell in a cross-modality correspondence relationship is substituted for another. The key to the substitution matrix is a novel cell-cell ambiguity measurement, where two cells are defined to be ambiguous if they share very similar geometric contexts [15], even though they might be very far apart in the data manifold. For each cell, SONATA identifies a handful of ambiguous cells with high ambiguity. Next, SONATA partitions the cells into mutually ambiguous cell groups such that (1) cells within each group have high feature similarities, and (2) ambiguous cells only exist between different groups rather than within the same group. Given the fact that existing manifold alignment methods seek cross-modality correspondence by relying on cell-specific geometric structures [16, 9, 10], mutually ambiguous

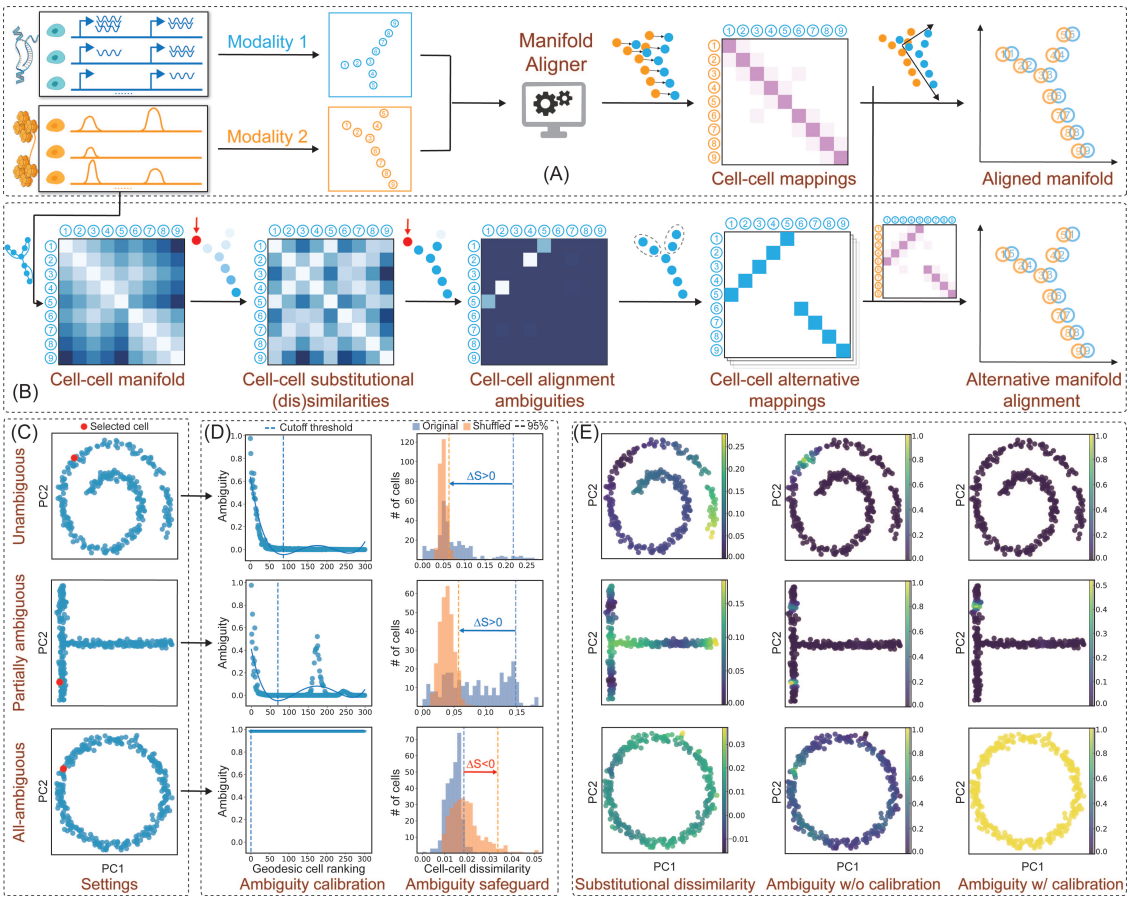

Figure 1: **Workflow of SONATA applied to single-cell multi-omics data alignment.** (A) A common manifold alignment method takes two datasets (*e.g.,* gene expression and chromatin accessibility) as input, elucidates their underlying manifold structures, identifies the cell correspondence relationships across data modalities, and consequently generates a joint cellular profile integrated from different datasets. (B) SONATA first defines a cell-cell substitution matrix which measures the likelihood that one cell can be substituted for another in the alignment. SONATA then partitions the cells into mutually ambiguous groups such that these groups can be mutually considered as alternative alignment candidates. Finally, SONATA enumerates the correspondence between groups and reports all alignment solutions between two datasets exhaustively. (C) SONATA can deal with unambiguous, partially ambiguous, and all-ambiguous settings. (D) SONATA fits a univariate spline to mitigate the distance-related biases and uses a "virtual" cell to protect exceptional cases when all cells are ambiguous. (E) The calibrated cell-cell ambiguities measure whether cells from the same dataset are geometrically similar to obscure the alignment.

cell groups share similar geometric structures and thus can be mutually considered as alternative alignment candidates. Finally, SONATA enumerates the correspondence between substitutional groups and reports all alignment solutions between two datasets exhaustively. We have applied SONATA to both simulated and real datasets to demonstrate its empirical utility.

2 Methods

2.1 Problem setup

Let the two datasets of cells be $X = \{x_i\}_{i=1}^{n_x} \in \mathbb{R}^{n_x \times p_x}$ in data modality $\mathcal{X}$ and $Y = \{y_j\}_{j=1}^{n_y} \in \mathbb{R}^{n_y \times p_y}$ in data modality $\mathcal{Y}$, respectively. The numbers of cells in the two modalities are n_x and n_y, and the feature dimensions are p_x and p_y, respectively. Without loss of generality, we assume two datasets can be aligned, where the alignability can be examined by statistical testing [14]. Given X and Y as input, manifold alignment methods aim to identify a low-dimensional manifold that captures covariation between X and Y. As a result, these methods generate a cross-modality correspondence matrix $\Gamma \in \mathbb{R}^{n_x \times n_y}$ explicitly [9, 10] or implicitly [8, 17], where Γ_{ij} represents the probability of the i-th cell in $\mathcal{X}$ aligning with the j-th cell in $\mathcal{Y}$. In this paper, rather than seeking a better Γ, SONATA focuses on the problem of manifold alignment ambiguity and estimates which cells can be aligned confidently and which cannot. The quantified alignment ambiguity enables SONATA to identify alternative manifold alignment solutions that existing methods might neglect, as illustrated in Figure 1(B). Though alternative solutions are induced by geometric resemblance instead of biological validity, SONATA offers a more comprehensive view of single-cell data integration, instead of providing biased and misleading interpretations from incomplete results.

2.2 Measuring the geometric (dis)similarity between cells

SONATA represents the low-dimensional manifold structure of each single-cell dataset using a geodesic distance matrix of the cells. To do this, SONATA first construct a weighted k-nearest neighbor (k-NN) graph of cells based on Euclidean distance [18]. The number of neighbors k should be sufficiently large to connect the corresponding k-NN graph [10]. SONATA then calculates the shortest distance between each node pair on the graph because the shortest distances approximate geodesic distances [19]. We denote the geodesic distance matrices for two datasets X and Y by $\mathcal{K}_x \in \mathbb{R}^{n_x \times n_x}$ and $\mathcal{K}_y \in \mathbb{R}^{n_y \times n_y}$, respectively.

For an arbitrary dataset X and its corresponding manifold $\mathcal{K}_x$ (or Y and $\mathcal{K}_y$), SONATA measures the likelihood that one cell from the dataset can be substituted for another in a cross-modality alignment. To do this, SONATA first represents the geometric context [15] of the i-th cell by its geodesic distance to other cells, in terms of the $\mathcal{K}_i \in \mathbb{R}^n$, *i.e.,* the i-th row of $\mathcal{K}$. Because cells have no natural ordering in the manifold, we sort $\mathcal{K}_i$ into an ordered sequence, denoted as $\mathcal{K}_{(i)} \in \mathbb{R}^{n_x}$, to make the geometric contexts of cells comparable. Then the substitutional likelihood between the i-th and the j-th cells can be measured by a novel cell-cell dissimilarity $S_{ij} = \left\|\mathcal{K}_{(i)} - \mathcal{K}_{(j)}\right\|_1$, where the operator $\|\cdot\|_1$ is the L_1 norm. In this way, a low dissimilarity indicates that two cells share a similar geometric contexts, even though they might be far apart in the underlying manifold.

2.3 Measuring the alignment ambiguity between cells

To quantify which cells can be aligned confidently and which cannot, SONATA defines *ambiguity* to measure whether cells from the same dataset are geometrically similar so that multiple possible solutions might exist to obscure the alignment. The ambiguity between the i-th and the j-th cells is defined as:

$$A_{ij} = \frac{e^{-S_{ij}/\sigma^2}}{\sum_{j'=1}^{n} e^{-S_{ij'}/\sigma^2}} \tag{1}$$

where $\sigma > 0$ is the bandwidth parameter that can be selected in a data-driven way [20].

Equation 1 leads to biased ambiguity estimation even for an unambiguously located cell, as illustrated in Figure 1(C). It is because two cells close to each other naturally have similar geometric contexts, and smaller geodesic distance subsequently induces smaller dissimilarity or equivalently overestimation of ambiguity. To mitigate the distance-related biases, SONATA fits a univariate spline to model the cells' ambiguity as a function of their ranking based on geodesic distance, using the "UnivariateSpline" function from the SciPy package. The inflection point of the spline determines a threshold to exclude the cells with overestimated ambiguity caused by distance-related biases from the consideration of ambiguous cells, as illustrated in Figure 1(D).

Additionally, Equation 1 mathematically facilitates only a small portion of cells ambiguous to a particular cell (*i.e.,* ambiguity > 0) and the remaining majority unambiguous (*i.e.,* ambiguity ≈ 0). Though it is generally true that both ambiguous and unambiguous cells exist, Equation 1 fails in exceptional cases when all cells are ambiguous, as illustrated in Figure 1(E). To address this problem, we construct a null distribution of cell-cell dissimilarities as a baseline to highlight whether the dissimilarities used in Equation 1 are reliable. Specifically, we first generate a permuted geodesic distance matrix $\tilde{\mathcal{K}} \in \mathbb{R}^{n_x \times n_x}$ derived from $\mathcal{K}$, which can be regarded as the geometric contexts of n permutations. We then calculate the cell-permutation dissimilarity $\tilde{S}_{ij} = \left\| \mathcal{K}_{(i)} - \tilde{\mathcal{K}}_{(j)} \right\|_1$, serving as the null distribution of cell-cell dissimilarities. Thus intuitively, cell-cell dissimilarity being much smaller than cell-permutation dissimilarity provides evidence that two cells indeed have similar geometric contexts whereas cell-cell dissimilarity being larger than cell-permutation dissimilarity indicates that two cells are far apart in the underlying manifold with distinct geometric contexts. In other words, cell-cell dissimilarities are generally more dispersed than cell-permutation dissimilarities, as illustrated in Figure 1(D). As such, SONATA artificially creates a "virtual" cell, serving as an unambiguity safeguard, of which the cell-cell dissimilarity to the i-th cell is recorded as $\Delta S_i = \text{sgn}(Q_{95}(S_i) - Q_{95}(\tilde{S}_i))Q_{95}(S_i)$, where $\text{sgn}(\cdot)$ is the sign function, $Q_{95}(\cdot)$ is the 95th percentile, and $S_i \in \mathbb{R}^{n_x}$ and $\tilde{S}_i \in \mathbb{R}^{n_x}$ denote the i-th row of S and $\tilde{S}$, respectively. The rationale for the "virtual" cell is two-fold (Figure 1(E)): (1) When $\Delta S_i > 0$, the "virtual" cell is unambiguous as expected (*i.e.,* ambiguity ≈ 0), and (2) In exceptional cases when $\Delta S_i < 0$, the "virtual" cell with the least ambiguity by design, enforces all cells to be ambiguous.

2.4 Generating alternative manifold alignment solutions

Cells participate in ambiguous alignment jointly with their neighbor cells rather than in an isolated fashion. Therefore, SONATA needs to identify mutually ambiguous cell groups so that one group can be substituted for another in a cross-modality alignment, which can be framed as a constrained clustering task [21]. On the one hand, we aim to identify groups of cells close to each other in the low-dimensional manifold, following the primary goal of clustering. On the other hand, we convert the identified cell-cell ambiguity relationships into pairwise constraints, which specify two cells cannot be placed in the same group. Incorporating these constraints, which serve as a degree of belief rather than an assertion, helps to improve the identified groups subject to the belief of ambiguity. Concretely, SONATA uses pairwise constrained K-means algorithm [21], an extension of standard K-means algorithm by adding the pairwise ambiguity-derived constraints. This step partitions a subset of cells involved in at least one cell-cell ambiguity relationship into $K \geq 2$ groups, each corresponding to a mutually ambiguous cell group. The choice of K is determined by the Elbow method by counting the number of violated constraints as a function of the number of clusters. As the number of clusters increases, the number of violated constraints decreases, and the

optimal K corresponds to the cluster number where the decrease diminishes and starts to reach a plateau (See Section 3 for more details).

Finally, SONATA combinatorially permutes identified ambiguous groups and reports alternative alignment solutions exhaustively. The group permutation decomposes into a series of mutual substitutions between group pairs. Specifically, for each group pair $G_s, G_t \subseteq \{1, 2, \cdots, n_x\}$, SONATA first learns the cell-cell correspondence relationships between two groups. The correspondence relationships can be represented by a point matching matrix $\Gamma_{st} = [\gamma_{ij}]_{i,j=1}^{|G_s|,|G_t|} \in \mathbb{R}^{|G_s|\times|G_t|}$ satisfying $\Gamma_{st}\mathbf{1}_{|G_t|} = \mathbf{1}_{|G_s|}$ and $\Gamma_{st}^T\mathbf{1}_{|G_s|} = \mathbf{1}_{|G_t|}$, where $\mathbf{1}_{|G_s|} \in \mathbb{R}^{|G_s|}$ and $\mathbf{1}_{|G_t|} \in \mathbb{R}^{|G_t|}$ are vectors of ones. Γ_{st} can be obtained by re-purposing existing manifold alignment methods for aligning two groups within the same dataset. Thereby, given an existing alignment solution in the form of cross-modality correspondence matrix $\Gamma \in \mathbb{R}^{n_x\times n_y}$ SONATA reports an alternative solution $\mathcal{P}_{st}\Gamma \in \mathbb{R}^{n_x\times n_y}$ where the permutation matrix $\mathcal{P}_{st} = [P_{ij}]_{i,j=1}^{n_x} \in \mathbb{R}^{n_x\times n_x}$ is defined as: $P_{ij} = \begin{cases} \gamma_{ij}, & i \in G_s, j \in G_t \\ \gamma_{ji}, & j \in G_s, i \in G_t \\ 1, & i = j \\ 0, & \text{otherwise} \end{cases}$.

3 Results

3.1 Performance on simulated data

3.1.1 Experimental setup

To validate the performance of SONATA, we generated five different simulated datasets, as shown in Figure 2. The simulations can be divided into three categories, including unambiguous, partially ambiguous, and all-ambiguous settings. The unambiguous setting contains a Swiss roll where two manifolds are uniquely aligned without any ambiguity. The partially ambiguous setting contains a T-shaped branch, a Y-shaped branch, and an X-shaped branch to mimic the bifurcating or multifurcating differentiation process. The difference between the T-shaped, Y-shaped, and X-shaped branches lies in the resemblance of two, three, and four branches, respectively. Lastly, the all-ambiguous setting contains a circle to mimic the periodic biological processes, such as the cell cycle. Each simulation contains two modalities with $n_x = n_y = 300$ samples residing in a 2-dimensional manifold, respectively. In particular, partially ambiguous datasets were simulated with branch-level correspondence relationships used to evaluate the manifold alignment performance qualitatively. The 2-dimensional manifold is subsequently mapped to $p_x = 1000$ and $p_y = 2000$ dimensions of features with Gaussian noises ($\sigma = 0.1$), respectively. Because each modality is projected to a different feature dimension, there is no feature-level correspondence relationship. In all simulations, we aligned the datasets using SCOT [9], a commonly-used manifold alignment method using optimal transport. In principle, SONATA is generalizable to any off-the-shelf manifold alignment methods, including MMD-MA [8], SCIM [17], and UnionCom [10].

3.1.2 Simulation results

We first evaluated the performance of the existing manifold alignment method in five simulated datasets. We projected the two aligned modalities to a shared 2D space for visualization using Principal Components Analysis (PCA). As shown in Figure 2(A), as a representative manifold alignment method, SCOT demonstrates its ability to integrate the two modalities with seemingly well-aligned geometric structures in all five simulated datasets. However, in three partially ambiguous datasets

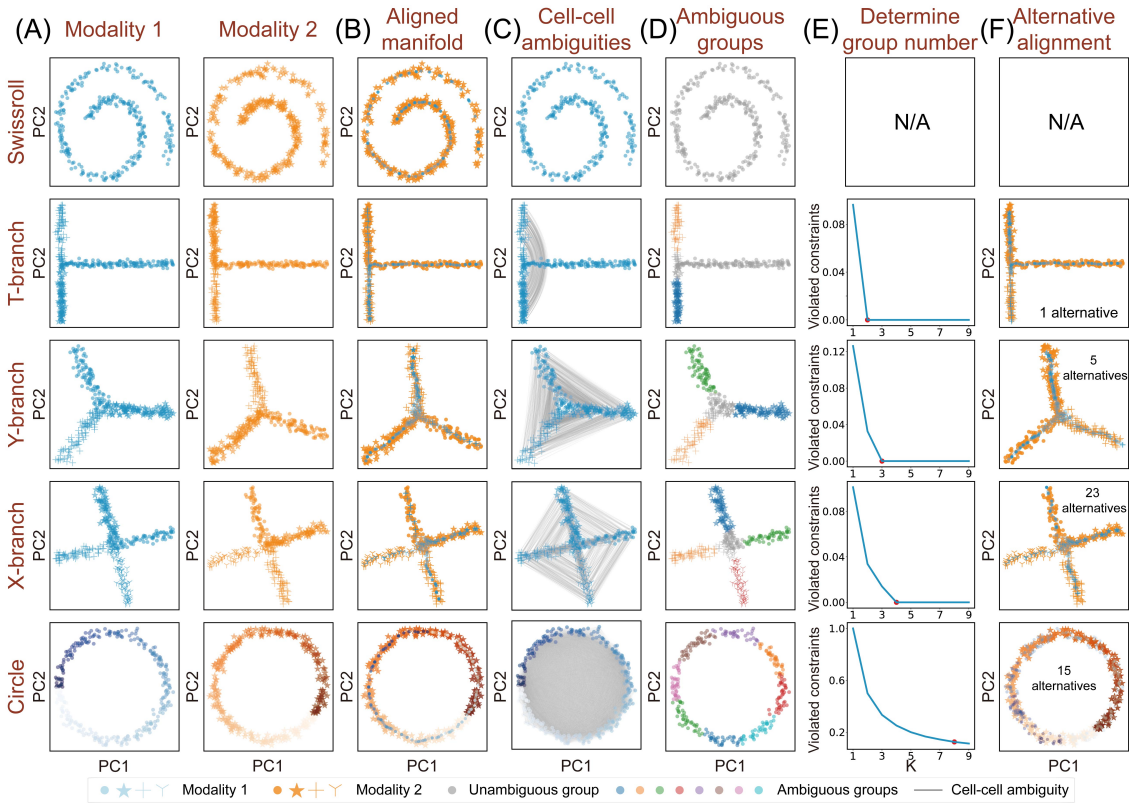

Figure 2: **Evaluating SONATA on simulated datasets.** Each row presents a different simulation dataset, each containing two modalities where cells from different modalities are colored differently. Three partially ambiguous datasets (*i.e.,* T-shaped, Y-shaped, and X-shaped branches) have branch-level correspondence relationships, represented by different markers. To compare alignment results in the all-ambiguous dataset (*i.e.,* circle), we chose a random cell as a reference and thereby colored the remaining cells by following a color gradient. (A) PCA projection of cells from two modalities separately. (B) PCA projection of aligned cells in the shared space after SCOT aligning the two modalities. (C) SONATA identifies cell-cell ambiguity relationships in modality 1, which measures geometrically similar cell pairs to obscure the alignment. (D) SONATA detects mutually ambiguous cell groups in modality 1 based on cell-cell ambiguity relationships, where cells from different groups are colored differently. Depending on its cell-cell ambiguity relationships, each cell is assigned to either an unambiguous or a specific ambiguous cell group. (E) The choice of ambiguous cell group number is determined by the Elbow method by counting the number of violated constraints as a function of the number of clusters. (F) PCA projection of alternative alignment of cells from two modalities suggested by SONATA.

(*i.e.,* T-shaped, Y-shaped, and X-shaped branches), SCOT aligns the branches without putting known branch-level correspondence together. For example, in the T-shaped branch dataset shown in Figure 2(B), the branch marked by '$\star$' in modality 1 is aligned with the one marked by '+' in modality 2, and vice versa. Worse still, in the all-ambiguous circle dataset, SCOT quickly reports an alignment solution without realizing potentially countless number of alignment solutions. Note that countless (but wrong) alignments exist between two cycles because rotating a cycle with arbitrary degrees naturally leads to a new solution.

Given the struggling performance of SCOT in ambiguous datasets, we next investigated whether SONATA could help SCOT identify the ambiguity in datasets, thereby improving the alignment results. As shown in Figure 2(C), SONATA accurately identifies cell-cell ambiguity relationships between geometrically similar branches. For example, SONATA identifies geometrically similar cell pairs between two and three branches with resemblance in the T-shaped and Y-shaped branch datasets, respectively. Based on identified cell-cell ambiguity relationships, SONATA accurately detects two, three, and four mutually ambiguous cell groups in the T-shaped, Y-shaped, and X-shaped branch datasets, respectively (Figure 2(D)). The number of ambiguous cell groups is determined by the Elbow method by counting the number of violated constraints as a function of the number of clusters, as shown in Figure 2(E). It is worth mentioning that the Elbow method recommends eight, a practically specified upper bound, rather than infinite ambiguous cell groups in the all-ambiguous circle dataset. Also, an alternative alignment is considered invalid if it distorts the data manifold by changing the mutual nearest neighbors of cells [22]. Finally, as shown in Figure 2(F), SONATA consistently identifies alternative alignment solutions that put the known branch-level correspondence together.

3.2 Real data analysis

3.2.1 Application to sc-GEM data

We first applied SONATA to integrate a dataset generated using the sc-GEM technology [23], a sequencing technology that simultaneously profiles gene expression and DNA methylation states within the same cell. The dataset was derived from human somatic cell samples undergoing conversion to induced pluripotent stem (iPS) cells [23], which contains $n_x = n_y = 177$ cells with $p_x = 34$ dimensions of gene expression features and $p_y = 27$ dimensions of gene expression features, respectively. We initialized by projecting the gene expression and the DNA methylation data to a 2D space separately using PCA. Given the cell type annotations shown in Figure 3(A), both data exhibit similar trajectories with the same orders of cell types, that is, from the BJ (human foreskin fibroblast) cells at one end to the iPS cells at the other end through the underlying process.

We then investigated the performance of SCOT in aligning sc-GEM datasets. As shown in Figure 3(B), SCOT seemingly aligns the cells between two modalities well by superimposing two trajectories together. However, based on the cell type annotations, iPS cells, located initially at only one end of the trajectory, now reside in both ends of the aligned trajectories. The observation implies that SCOT reversely aligns one modality to another during the alignment process. It is worth noting that we run SCOT multiple times and report the reversely alignment. We investigated whether SONATA could help SCOT identify the ambiguity in sc-GEM datasets, thereby correctly aligning two trajectories. As shown in Figure 3(C), SONATA identifies cell-cell ambiguity relations and subsequently detects two mutually ambiguous cell groups between two ends of the reprogramming trajectory. Finally, SONATA successfully identifies an alternative alignment between two modalities

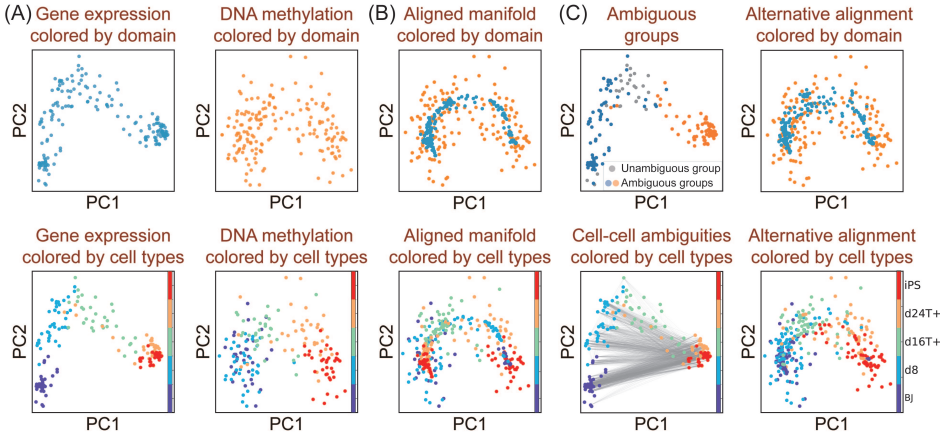

Figure 3: **Applying SONATA to sc-GEM datasets consisting of gene expression and DNA methylation measures.** The cells in each modality are colored by their annotated cell types before or after the alignment. (A) PCA projection of cells from two modalities separately. (B) PCA projection of aligned cells in the shared space after SCOT aligning the two modalities. We ran SCOT multiple times and reported the alignment that flips one modality and reversely maps it to another. (C) SONATA first identifies cell-cell ambiguity relationships in gene expression data, then detects two mutually ambiguous cell groups between two ends of the underlying trajectory, and finally provides an alternative alignment of cells that agrees with the cell type annotations.

that locates BJ cells at one end and iPS cells at the other end.

3.2.2 Application to SNARE-Seq data

We next applied SONATA to integrate a dataset generated using the SNARE-seq technology [24], a sequencing technology that simultaneously profiles gene expression and chromatin accessibility within the same cell. The dataset was derived from a mixture of BJ, H1, K562, and GM12878 human cell lines [24], which contains $n_x = n_y = 1047$ cells in both modalities. Following the data preprocessing procedure of SCOT, we carried out dimensionality reduction on the chromatin accessibility information using the topic modeling framework cisTopic [25], resulting in $p_x = 19$ dimensions of chromatin accessibility features. We also performed dimensionality reduction on the gene expression information using PCA, resulting in $p_y = 10$ dimensions of gene expression features. We unit normalized both features and aligned the datasets using SCOT [9], We initialized by projecting the gene expression and the chromatin accessibility data to a 2D space separately using PCA. Given the cell line annotations shown in Figure 4(A), both data exhibit similar patterns from two perspectives: (1) the H1 and GM12878 cells reside in separate clusters, and (2) the BJ and K562 cells are close to each other geometrically.

We then investigated the performance of SCOT in aligning SNARE-Seq datasets. As shown in Figure 4(B), SCOT seemingly aligns the cells between two modalities well by retaining the geometric structure of the chromatin accessibility modality. However, by scrutinizing the cell line annotations, H1, BJ, and K562 cells are mixed together in different clusters. The observation implies that SCOT flips the correspondence between H1 and BJ/K562 clusters during the alignment process. It is worth noting that we run SCOT multiple times and report the flipped alignment. We investigated whether SONATA could help SCOT identify the ambiguity in SNARE-seq datasets,

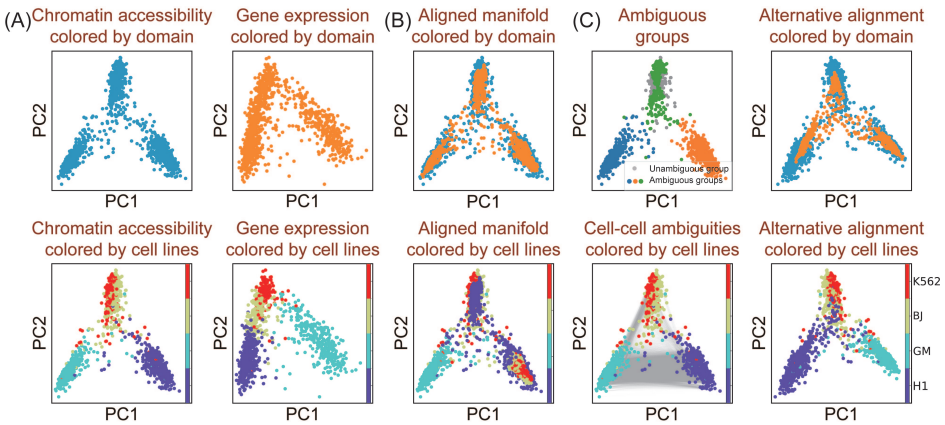

Figure 4: **Applying SONATA to SNARE-Seq datasets consisting of gene expression and chromatin accessibility measures.** The cells in each modality are colored by their annotated cell lines before or after the alignment. (A) PCA (Principal Components Analysis) projection of cells from two modalities separately. (B) PCA projection of aligned cells in the shared space after SCOT aligning the two modalities. We ran SCOT multiple times and reported the alignment that mixes H1, BJ, and K562 cells together in different clusters. (C) SONATA first identifies cell-cell ambiguity relationships in gene expression data, then detects three mutually ambiguous cell groups (*i.e.,* H1, GM12878, and BJ/K562), and finally provides an alternative alignment of cells that agrees with the cell line annotations.

thereby correctly pairing gene expression profiles with the correct chromatin accessibility profiles. As shown in Figure 4(C), after further denoising chromatin accessibility into 2D space, SONATA identifies cell-cell ambiguity relations and subsequently detects three mutually ambiguous cell groups (*i.e.,* H1, GM12878, and BJ/K562). Finally, SONATA successfully identifies an alternative alignment between two modalities that locates (1) H1 and GM12878 cells in separate clusters, and (2) BJ and K562 cells in the same cluster.

4 Discussion and conclusion

In this work, we proposed SONATA, a novel method that can help mitigate the ambiguity of manifold alignment by identifying potentially multiple valid solutions. The alignment ambiguity is achieved by proposing a novel cell-cell ambiguity measurement, where two cells are defined to be ambiguous if they share very similar geometric contexts, even though they might be very far apart in the data manifold. SONATA summarizes cell-cell ambiguities into mutually ambiguous cell groups such that these groups share similar geometric structures and thus can be mutually considered as alternative alignment candidates. The key novelty of SONATA is that it provides a systematic strategy to distinguish whether the manifold alignment has one unique solution or multiple alternative solutions, in comparison to existing methods that report an arbitrary alignment solution recklessly. Our experiments demonstrate that SONATA improves state-of-the-art manifold alignment methods by identifying and resolving the alignment ambiguity. In summary, SONATA promises to enhance the accuracy and reliability of integrative single-cell sequencing studies, unlocking deeper insights into complex biological systems.

This work points to several promising directions for future research. SONATA is designed as a

post-processing method designed to complement existing manifold alignment methods. We rationalize that the proposed cell-cell ambiguity measurement can be re-purposed to measure the cell-cell similarity between different modalities. Using SONATA directly for both manifold alignment and disambiguation would be interesting directions to pursue. Additionally, several critical single-cell applications, such as batch-effect correction [7] and comparative analysis between case-control studies [26], are closely associated with the manifold alignment. We would like to use SONATA to identify and resolve the ambiguity in wide applications.

References

[1] Tang, F. et al. mRNA-Seq whole-transcriptome analysis of a single cell. Nature Methods **6**, 377–382 (2009).

[2] Buenrostro, J. D. et al. Single-cell chromatin accessibility reveals principles of regulatory variation. Nature **523**, 486–490 (2015).

[3] Smallwood, S. A. et al. Single-cell genome-wide bisulfite sequencing for assessing epigenetic heterogeneity. Nature Methods **11**, 817–820 (2014).

[4] Nagano, T. et al. Single-cell Hi-C reveals cell-to-cell variability in chromosome structure. Nature **502**, 59–64 (2013).

[5] Ku, W. L. et al. Single-cell chromatin immunocleavage sequencing (scChIC-seq) to profile histone modification. Nature Methods **16**, 323–325 (2019).

[6] Argelaguet, R., Cuomo, A. S., Stegle, O. & Marioni, J. C. Computational principles and challenges in single-cell data integration. Nature Biotechnology **39**, 1202–1215 (2021).

[7] Luecken, M. D. et al. Benchmarking atlas-level data integration in single-cell genomics. Nature Methods **19**, 41–50 (2022).

[8] Liu, J., Huang, Y., Singh, R., Vert, J.-P. & Noble, W. S. Jointly embedding multiple single-cell omics measurements. In International Workshop on Algorithms in Bioinformatics, vol. 143, 10:1–10:13 (2019).

[9] Demetci, P., Santorella, R., Sandstede, B., Noble, W. S. & Singh, R. SCOT: Single-cell multi-omics alignment with optimal transport. Journal of Computational Biology **29** (2022).

[10] Cao, K., Bai, X., Hong, Y. & Wan, L. Unsupervised topological alignment for single-cell multi-omics integration. Bioinformatics **36**, i48–i56 (2020).

[11] Elmentaite, R., Conde, C. D., Yang, L. & Teichmann, S. A. Single-cell atlases: shared and tissue-specific cell types across human organs. Nature Reviews Genetics **23**, 395–410 (2022).

[12] Schafer, K. The cell cycle: a review. Veterinary Pathology **35**, 461–478 (1998).

[13] Chari, T. & Pachter, L. The specious art of single-cell genomics. PLOS Computational Biology **19**, e1011288 (2023).

[14] Ma, R., Sun, E. D., Donoho, D. & Zou, J. Is your data alignable? Principled and interpretable alignability testing and integration of single-cell data. bioRxiv 2023–08 (2023).

[15] Ribeiro, L. F. R., Saverese, P. H. P. & Figueiredo, D. R. struc2vec: Learning node representations from structural identity. In ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 385–394 (2017).

[16] Cui, Z., Chang, H., Shan, S. & Chen, X. Generalized unsupervised manifold alignment. In Advances in Neural Information Processing Systems, 2429–2437 (Curran Associates, Inc., Montreal, Canada, 2014).

[17] Stark, S. G. et al. SCIM: Universal single-cell matching with unpaired feature sets. Bioinformatics **36**, i919–i927 (2020).

[18] Costa, J. A. & Hero, A. O. Manifold learning using euclidean k-nearest neighbor graphs [image processing examples]. In 2004 IEEE International Conference on Acoustics, Speech, and Signal Processing, vol. 3, iii–988 (IEEE, 2004).

[19] Tenenbaum, J. B., de Silva, V. & Langford, J. C. A global geometric framework for nonlinear dimensionality reduction. Science **290**, 2319–2323 (2000).

[20] Singh, R. et al. Unsupervised manifold alignment for single-cell multi-omics data. In Proceedings of the 11th ACM International Conference on Bioinformatics, Computational Biology and Healt 1–10 (2020).

[21] Basu, S., Banerjee, A. & Mooney, R. J. Active semi-supervision for pairwise constrained clustering. In International Conference on Data Mining, 333–344 (SIAM, 2004).

[22] Haghverdi, L., Lun, A. T. L., Morgan, M. D. & Marioni, J. C. Batch effects in single-cell RNA-sequencing data are corrected by matching mutual nearest neighbors. Nature Biotechnology **36**, 421–427 (2018).

[23] Cheow, L. et al. Single-cell multimodal profiling reveals cellular epigenetic heterogeneity. Nature Methods **13**, 833–836 (2016).

[24] Chen, S., Lake, B. B. & Zhang, K. High-throughput sequencing of the transcriptome and chromatin accessibility in the same cell. Nature Biotechnology **37**, 1452–1457 (2019).

[25] González-Blas, C. B. et al. cisTopic: cis-regulatory topic modelling on single-cell ATAC-seq data. Nature Methods **16**, 397–400 (2018).

[26] Park, Y. P. & Kellis, M. CoCoA-diff: counterfactual inference for single-cell gene expression analysis. Genome Biology **22**, 1–23 (2021).