PNAS RESEARCH ARTICLE | STATISTICS
BIOPHYSICS AND COMPUTATIONAL BIOLOGY OPEN ACCESS

# Principled and interpretable alignability testing and integration of single-cell data

Rong Ma[a,1], Eric D. Sun[b], David Donoho[c], and James Zou[b,1]

Edited by David Weitz, Harvard University, Cambridge, MA; received August 9, 2023; accepted January 23, 2024

**Single-cell data integration can provide a comprehensive molecular view of cells, and many algorithms have been developed to remove unwanted technical or biological variations and integrate heterogeneous single-cell datasets. Despite their wide usage, existing methods suffer from several fundamental limitations. In particular, we lack a rigorous statistical test for whether two high-dimensional single-cell datasets are alignable (and therefore should even be aligned). Moreover, popular methods can substantially distort the data during alignment, making the aligned data and downstream analysis difficult to interpret. To overcome these limitations, we present a spectral manifold alignment and inference (SMAI) framework, which enables principled and interpretable alignability testing and structure-preserving integration of single-cell data with the same type of features. SMAI provides a statistical test to robustly assess the alignability between datasets to avoid misleading inference and is justified by high-dimensional statistical theory. On a diverse range of real and simulated benchmark datasets, it outperforms commonly used alignment methods. Moreover, we show that SMAI improves various downstream analyses such as identification of differentially expressed genes and imputation of single-cell spatial transcriptomics, providing further biological insights. SMAI's interpretability also enables quantification and a deeper understanding of the sources of technical confounders in single-cell data.**

single-cell omics | data alignment | random matrix theory | spectral method | Procrustes analysis

The rapid development of single-cell technologies has enabled the characterization of complex biological systems at unprecedented scale and resolution. On the one hand, diverse and heterogeneous single-cell datasets have been generated, enabling opportunities for integrative profiling of cell types and deeper understandings of the associated biological processes (1–3). On the other hand, the widely observed technical and biological variations across datasets also impose unique challenges to many downstream analyses (4–6). The variations between datasets can originate from different experimental protocols, laboratory conditions, sequencing technologies, etc. It may also arise as biological variations when the samples come from distinct spatial locations, times, tissues, organs, individuals, or species.

Several computational algorithms have been developed to remove the unwanted variations and integrate heterogeneous single-cell datasets. To date, the most widely used data integration methods, such as Seurat (7), LIGER (8), Harmony (9), fastMNN (10), and Scanorama (11), are built upon the key assumption that there is a shared latent low-dimensional structure between the datasets of interest. These methods attempt to obtain an alignment of the datasets by identifying and matching their respective low-dimensional structures. As a result, the methods would output some integrated cellular profiles, commonly represented as either a corrected feature matrix, or a joint low-dimensional embedding matrix, where the unwanted technical or biological variations between the datasets have been removed. These methods have played an indispensable role in current single-cell studies such as generating large-scale reference atlases of human organs (12, 13), inferring lineage differentiation trajectories (14, 15), and multiomic characterization of COVID-19 pathogenesis and immune response (16, 17).

Despite their popularity, the existing integration methods also suffer from several fundamental limitations, which makes it difficult to statistically assess findings drawn from the aligned data. First, there is a lack of statistically rigorous methods to determine whether two or more datasets should be aligned. Without such a safeguard, existing methods are used to align and integrate single-cell datasets that do not have a meaningful shared structure, leading to problematic and misleading interpretations (18, 19). Global assessment methods such as k-nearest neighbor batch effect test (kBET) (20), guided PCA (gPCA) (21), probabilistic principal component and covariates analysis (PPCCA) (22), and metrics such as local inverse Simpson's index (LISI) (9), average silhouette width

## Significance

Aligning and integrating different datasets is a key challenge in single-cell research. However, existing methods suffer from several fundamental and under-appreciated limitations. First, we do not have a rigorous statistical test for determining whether two single-cell datasets should even be integrated. Moreover, popular methods substantially distort the data during alignment, making the downstream analysis subject to bias and difficult to interpret. We address both challenges with a unified spectral manifold alignment and inference (SMAI) framework. SMAI is a flexible and interpretable method for aligning datasets with the same type of features, equipped with an alignability test justified by statistical theory. It preserves within-data structures and improves downstream analyses, such as identification of differentially expressed genes and imputation of spatial transcriptomics.

Author contributions: R.M. and J.Z. designed research; R.M., E.D.S., D.D., and J.Z. performed research; R.M., E.D.S., D.D., and J.Z. contributed new reagents/analytic tools; R.M. analyzed data; and R.M., E.D.S., and J.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: rongma@hsph.harvard.edu or jamesz@stanford.edu.

This article contains supporting information online at https://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2313719121/-/DCSupplemental.

Published February 28, 2024.

(ASW) (23), adjusted Rand index (ARI) (24), have been proposed to quantitatively characterize the quality of alignment or the extent of batch-effect removal, based on specific alignment procedures. However, these methods can only provide post-hoc evaluations of the mixing of batches, which may not necessarily reflect the actual alignability or structure-sharing between the original datasets. Moreover, these methods do not account for the noisiness (ASW, ARI, LISI) or the effects of high dimensionality (kBET, gPCA, PPCCA) of the single-cell datasets, resulting in biased estimates and test results. Other methods such as limma (25) and MAST (26) consider linear batch correction, whose focus is restricted to differential testing and does not account for possible covariance shifts.

Moreover, research suggests that existing integration methods may introduce serious distortions to the individual datasets during their alignment process (18, 19). In this study, we systematically evaluate the severity and effects of such distortions for several popular integration methods. Our results confirm that these methods, while eliminating the possible differences between datasets, may also alter the original biological signals contained in individual datasets, causing misleading results such as power loss and false discoveries in downstream analyses. Finally, none of these popular integration methods admits a tractable closed-form expression of the final alignment function, with a clear geometric meaning of its constitutive components. As a result, these methods would suffer from a lack of interpretability, making it difficult to inspect and understand the nature of any removed variations, or to distinguish the unwanted variations from the potentially biologically meaningful variations.

To overcome the above limitations, we present a spectral manifold alignment and inference (SMAI) framework for accountable and interpretable integration of single-cell data with the same type of features. Our contribution is two-fold. First, we develop a rigorous statistical test (SMAI-test) that can robustly determine the alignability between two datasets. Second, motivated by this test, we propose an interpretable spectral manifold alignment algorithm (SMAI-align) that enables more reliable data integration without altering or corrupting the original biological signals. Our systematic experiments demonstrate that SMAI improves various downstream analyses such as the identification of cell types and their associated marker genes, and the prediction of single-cell spatial transcriptomics. Moreover, we show that SMAI's interpretability provides insights into the sources of technical confounders in single-cell data.

## Results

**Overview of SMAI.** SMAI consists of two components: SMAI-test flexibly determines the global or partial alignability between the datasets, whereas SMAI-align searches for the best similarity transformation to achieve the alignment.

SMAI-test evaluates the statistical significance against the null hypothesis that two single-cell datasets are alignable up to some similarity transformation, that is, combinations of scaling, translation, and rotation. In line with several previous works, it leverages random matrix theory to be robust to noisy and high-dimensional single-cell omic data (27–30). As a hypothesis testing framework that determines the presence and the nature of batch effects between two datasets, SMAI-test tells users if the datasets should or should not be aligned by SMAI-test or other integration methods, as enforcing alignment would introduce substantial distortions. To increase flexibility, SMAI-test also allows for testing against partial alignability between the datasets, where the users can specify a threshold $t\%$, so that the null hypothesis states that at least $t\%$ of the samples in both datasets are alignable. Recommended values for $t$ are between 50 and 70 depending on the context, to ensure both sufficient sample size (or power), and flexibility to local heterogeneity (*SI Appendix*, Notes); we used $t = 60$ for the real datasets analyzed in this study. Importantly, the statistical validity of SMAI-test is theoretically guaranteed over a wide range of settings (*Materials and Methods*, Theorem 1), suitable for modeling high-dimensional single-cell data. We support the empirical validity of the test with both simulated data and multiple real-world benchmark datasets, ranging from transcriptomics, chromatin accessibility, to spatial transcriptomics.

SMAI-align incorporates a high-dimensional shuffled Procrustes analysis, which iteratively searches for the sample correspondence and the best similarity transformation that minimizes the discrepancy between the intrinsic low-dimensional signal structures of the datasets. SMAI-align enjoys several advantages over the existing integration methods. First, SMAI-align returns an alignment function in terms of a similarity transformation, which has a closed-form expression (*SI Appendix*, Notes) equipped with a clear geometric meaning. The better interpretability enables quantitative characterization of the source and magnitude of any removed and remaining variations, and may bring insights into the mechanisms underlying the batch effects. Second, due to the shape-invariance property of similarity transformations, SMAI-align preserves the relative distances between the samples within individual datasets throughout the alignment, making the final integrated data less susceptible to technical distortions and therefore more suitable and reliable for downstream analyses. Third, unlike many existing methods (such as Seurat, Harmony, and fastMNN), which require specifying a target dataset for alignment and whose performance is asymmetric with respect to the order of datasets, SMAI-align obtains a symmetric invertible alignment function that is indifferent to such an order, making its output more consistent and robust to technical artifacts. Below we sketch the main ideas of the SMAI algorithm and leave the details to *Materials and Methods* and *SI Appendix*, Notes.

**SMAI-Test.** Suppose that $\mathbf{X} \in \mathbb{R}^{d\times n_1}$ and $\mathbf{Y} \in \mathbb{R}^{d\times n_2}$ are the normalized count matrices generated from two single-cell experiments, with $d$ being the number of features (genes) and $n_1$ and $n_2$ being the respective numbers of cells. To test for the alignability between $\mathbf{X}$ and $\mathbf{Y}$, SMAI-test assumes a low-rank spiked covariance matrix model (*SI Appendix*, Fig. S1) where the low-dimensional signal structures of $\mathbf{X}$ and $\mathbf{Y}$ are encoded by the leading eigenvalues and eigenvectors of their corresponding population covariance matrices $\mathbf{\Sigma}_1$ and $\mathbf{\Sigma}_2$. As a result, the null hypothesis that the signal structures underlying $\mathbf{X}$ and $\mathbf{Y}$ are identical up to a similarity transformation implies that the leading eigenvalues of $\mathbf{\Sigma}_1$ and $\mathbf{\Sigma}_2$ are identical up to a global scaling factor. As such, a test statistic $T(\mathbf{X}, \mathbf{Y})$ based on comparing the leading eigenvalues of the empirical covariance matrices of $\mathbf{X}$ and $\mathbf{Y}$ can be computed, whose theoretical null distribution as $(n_1, n_2, d) \to \infty$ is derived using random matrix theory. Thus, SMAI-test returns the $P$-value by comparing the test statistic $T(\mathbf{X}, \mathbf{Y})$ with its asymptotic null distribution (Fig. 1*A*).

For the test of partial alignability, a sample splitting procedure is adopted where the first part is used to identify subsets of the two datasets with maximal correspondence or structure-sharing (*Materials and Methods*), and the second part is used to compute the test statistic and the $P$-value concerning the alignability between such maximal correspondence subsets. As such, we avoid

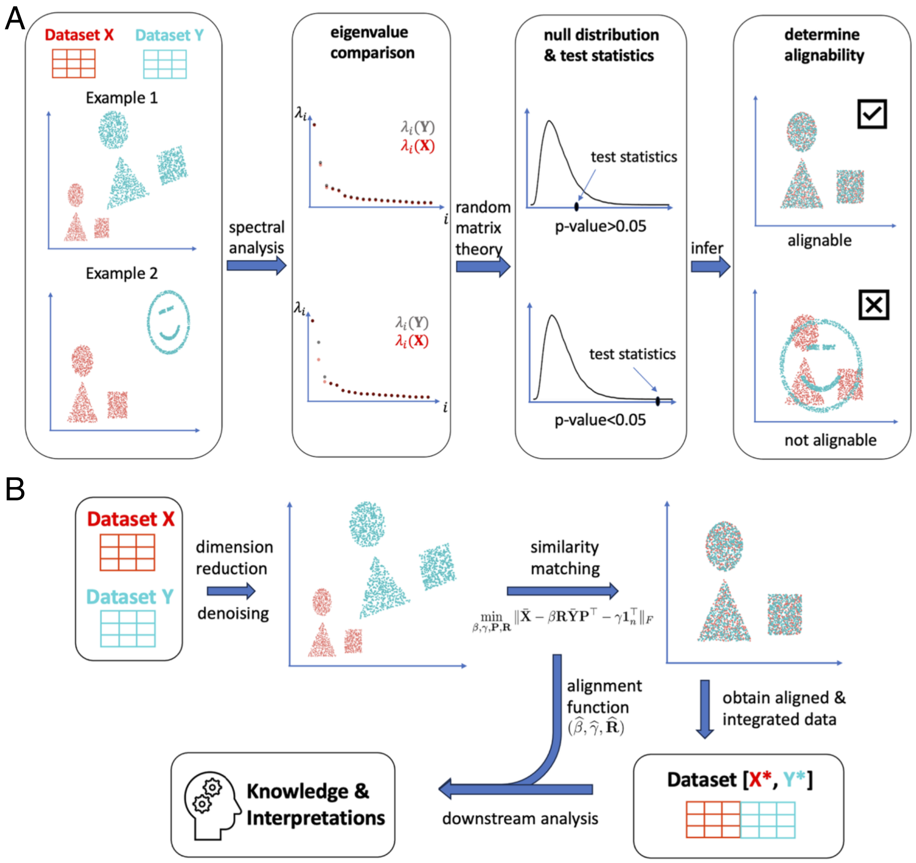

**Fig. 1.** Overview and illustration of the SMAI algorithm. (*A*) SMAI-test imposes a low-rank spiked covariance matrix model where the low-dimensional signal structures of data matrices are encoded by a few largest eigenvalues of the population covariance matrices. Under the null hypothesis that the underlying signal structures are alignable up to a similarity transformation, a test statistic based on comparing the leading eigenvalues of the empirical covariance matrices is computed, whose theoretical null distribution as $(n, d) \to \infty$ is derived using random matrix theory. The final *P*-value returned by SMAI-test is used to infer the alignability of the two datasets. (*B*) SMAI-align aims to solve the shuffled Procrustes optimization problem (**1**). To do so, SMAI-align starts with a denoising procedure, and then adopts an iterative spectral algorithm to achieve similarity matching between the two datasets using high-dimensional Procrustes analysis. The method returns an integrated dataset containing all the samples with the original features, along with a closed-form alignment function, which is interpretable and can be readily used for various downstream analyses.

selection bias due to repeated use of the samples in both selection and test steps, ensuring a valid test.

**SMAI-Align.** SMAI-align starts by filtering out the low-rank signal structures in $\mathbf{X}$ and $\mathbf{Y}$ to obtain their denoised versions $\widehat{\mathbf{X}}$ and $\widehat{\mathbf{Y}}$, and then approximately solves the following shuffled Procrustes optimization problem

$$\min_{\beta,\boldsymbol{\gamma},\mathbf{P},\mathbf{R}} \|\widehat{\mathbf{X}} - \beta\mathbf{R}\widehat{\mathbf{Y}}\mathbf{P} - \boldsymbol{\gamma}\mathbf{1}_{n_1}^\top\|_F. \qquad [1]$$

Here, $\mathbf{1}_n \in \mathbb{R}^n$ is an all-one vector, and the minimization is achieved for some global scaling factor $\beta \in \mathbb{R}$, some vector $\boldsymbol{\gamma} \in \mathbb{R}^d$ adjusting for the possible global mean shift between $\widehat{\mathbf{Y}}$ and $\widehat{\mathbf{X}}$, some extended orthogonal matrix (*Materials and Methods*) $\mathbf{P} \in \mathbb{R}^{n_2 \times n_1}$ recovering the sample correspondence between $\widehat{\mathbf{X}}$ and $\widehat{\mathbf{Y}}$, and some rotation matrix $\mathbf{R} \in \mathbb{R}^{d\times d}$ adjusting for the possible covariance shift. Compared with the traditional Procrustes analysis (31), Eq. **1** contains an additional matrix $\mathbf{P}$, allowing for a general unknown correspondence between the samples in $\mathbf{X}$ and $\mathbf{Y}$, which is the case in most of our applications. To solve for Eq. **1**, SMAI-align adopted an iterative spectral algorithm that alternatively solves for $\mathbf{P}$ and $(\beta, \boldsymbol{\gamma}, \mathbf{R})$ using high-dimensional Procrustes analysis. The final solution $(\widehat{\beta}, \widehat{\boldsymbol{\gamma}}, \widehat{\mathbf{R}})$ then gives a good similarity transformation aligning the two datasets in the original feature space. In particular, to improve robustness and reduce the effects of potential outliers in the data on the final alignment function, in each iteration we remove some leading outliers from both datasets, whose distances to the other dataset remain large. Moreover, to allow for integration of datasets containing partially shared structures (up to a user-specified threshold; *Materials and Methods*), users may also request SMAI-align to infer the final alignment function only based on the identified maximal correspondence subsets, rather than the whole datasets. This makes the alignment more robust to local structural heterogeneity. SMAI-align returns an integrated dataset containing all the samples, along with the similarity transformation, which are interpretable and readily used for various downstream analyses (Fig. 1*B*). The idea of SMAI-align is closely related to that of SMAI-test: a pass in SMAI-test essentially renders the goodness-of-fit of the model underlying SMAI-align, and therefore ensures its performance. In addition, since SMAI-align essentially learns some underlying similarity transformation, based on which all the samples are aligned, the algorithm is easily scalable to very large datasets. For example, one can first infer the alignment function by applying SMAI-align to some representative subsets of the datasets, and then use it to align all the samples (*Materials and Methods*).

To empirically evaluate the statistical validity of SMAI-test and the consistency of SMAI-align, we generate simulated data based on some signal-plus-noise matrix model with various signal structures, batch effects, and sample sizes (*SI Appendix*, Notes). Our simulation results indicate that SMAI-test has desirable type I errors across all the simulation settings, that is, achieving the nominal probability (0.05) of rejecting the null hypothesis when the datasets are truly alignable (*SI Appendix*, Table S1 and *Materials and Methods*). We also evaluate the performance of SMAI-align in recovering the true alignment function by measuring the estimation errors for each of the true parameters ($\beta^*, \boldsymbol{\gamma}^*, \mathbf{R}^*$) generating the data (*Materials and Methods*). We find that increasing the sample sizes leads to reduced estimation errors in general (*SI Appendix*, Fig. S2*B*), suggesting statistical consistency of SMAI-align.

**SMAI Robustly Determines Alignability between Diverse Single-Cell Data.** We apply SMAI-test to diverse single-cell data integration tasks and demonstrate its robust performance in determining the alignability between datasets. The detailed information about each dataset and the corresponding test results are summarized in *SI Appendix*, Table S1. In particular, our real and synthetic datasets involve diverse tissues including human livers, human pancreas, human blood (peripheral blood mononuclear cell, PBMC), human lung, human mesenteric lymph nodes (MLN), human lung-draining lymph nodes (LLN), mouse brain, mouse PBMC (either lipopolysaccharide (LPS) stimulated and control), mouse primary visual cortex (VISP), and mouse gastrulation, and contain multiple modalities measured by various sequencing technologies, such as single-cell transcriptomics (10X Genomics, Smart-seq, Smart-seq2, Drop-seq, and CEL-seq2), spatial transcriptomics (seqFISH, ISS and ExSeq), and chromatin accessibility (ATAC-seq). The 13 integration tasks cover 7 different scenarios that arise commonly in single-cell research, including 1) integration across different samples with the same cell types, 2) integration across different samples with partially overlapping cell types, 3) integration across samples with non-overlapping cell types, 4) integration across studies with different sequencing technologies, 5) integration across studies with different tissues, 6) integration across studies with different experimental conditions, and 7) integration of single-cell RNA-seq and spatial-transcriptomic data. For each task, we test for partial alignability between each pair of datasets, determining whether at least 60% of the cells are alignable in the sense of our null model (*Materials and Methods*). Among them, three out of 13 integration tasks, with zero or very low proportions ($\leq 37\%$) of cells under the overlapping cell types, are taken as negative controls (Neg1–Neg3), whose alignability is doubtful in general; the rest of the tasks, including both non-spatial integration tasks (Pos1–Pos7) and spatial integration tasks (PosS1–PosS3), are taken as positive controls, whose alignability is expected due to the association with the same tissue or largely overlapping cell types.

SMAI-test returns significant *P*-values ($<0.01$) for all the negative controls, correctly detecting their unalignability against our null model. The positive control tasks are assigned non-significant *P*-values, passing the (partial) alignability test as expected. For the seven non-spatial integration tasks (Pos1–Pos7), SMAI-test confirms the alignability between datasets with largely overlapping ($\geq 84\%$) cell types but possibly different sequencing technologies, tissues or experimental conditions. For the three spatial tasks (PosS1–PosS3), SMAI-test confirms the alignability of the paired single-cell RNA-seq and spatial-transcriptomic data from the same tissue, justifying its wide use for downstream analyses such as prediction of unmeasured spatial genes (Fig. 5).

**Necessity of Certifying Data Alignability Prior to Integration.** In the absence of a principled procedure for determining the alignability between datasets, the existing integration methods often end up forcing alignment between any datasets by significantly distorting and twisting each original dataset (Fig. 2). Specifically, for each of the negative control tasks Neg1-Neg3, we apply seven existing integration methods [Scanorama, Harmony, LIGER, fastMNN, Seurat, Pamona (32), and SCOT (33)] to obtain the integrated datasets, and then evaluate how well the relative distances between the cells within each dataset before integration are preserved in the integrated datasets. As a result, we find overall low correlations (Kendall's tau correlation on average 0.6 across the seven methods and three negative control tasks Neg1–Neg3, as compared with 0.9 achieved by SMAI-align on average across the positive control tasks Pos1–Pos7) between the relative distances of the cells within each dataset before integration, and the distances after data integration (*Materials and Methods*, Fig. 2*B* and *SI Appendix*, Fig. S3*C*). In addition, we observe many cases of false alignment of distinct cell types from different datasets, and sometimes serious distortion and creation of artificial cell clusters under the same cell type. For example, in Task Neg1, we find the false alignment between ductal cells and acinar cells by Scanorama, Harmony, and fastMNN, between alpha cells and gamma cells by Scanorama, fastMNN, and Seurat, between alpha cells and beta cells by Seurat, and between alpha cells and endothelial cells by Pamona and SCOT (Fig. 2*A* and *SI Appendix*, Fig. S3*E*); we also observe significant distortion or dissolution of the alpha cell cluster and the beta cell cluster after integration by Harmony, LIGER, and fastMNN, as compared with the original datasets (Fig. 2*A*). As such, the final integration results can be highly problematic and unfaithful to the original datasets, which may lead to erroneous conclusions from downstream analysis. SMAI-test is able to detect the lack of alignability (i.e., significant *P*-values) between these datasets, alerting users that the integrated data may not be reliable.

To evaluate the possible effects on downstream analysis, we focus on one important application following data integration, that is, the identification of DE genes for each cell type. We consider the above five integration methods (Harmony and Pamona are not included as they only produce integrated data in the low-dimensional space). For the three negative control tasks, we find that for many cell types, the set of DE genes identified based on the integrated data have low overlap with those identified based on the original datasets (Fig. 2*C* and *SI Appendix*, Figs. S3*D* and S4). These discrepancies are likely artifacts created by the respective integration methods. For instance, in Task Neg1, we find that, compared with other methods, the integrated data based on Seurat has lower power in detecting DE genes for beta cells and ductal cells, which may be a result of collapsing beta and ductal cells with other cell types during Seurat integration (Fig. 2*D*). Similarly, we observe a higher false discovery proportions (FDPs) in detecting DE genes of alpha cells based on the integrated datasets by fastMNN and LIGER than other methods, which is likely a consequence of the artificial split of the alpha cell cluster during fastMNN and LIGER integration (Fig. 2*D*).

**SMAI Enables Principled Structure-Preserving Integration of Single-Cell Data.** For the first six non-spatial positive control tasks (Tasks Pos1–Pos6) with annotated cell types, we further apply SMAI-align to obtain the integrated datasets (Fig. 3*A* and *SI Appendix*, Fig. S5) and compare the quality of alignment with the above seven existing methods. We find SMAI-align

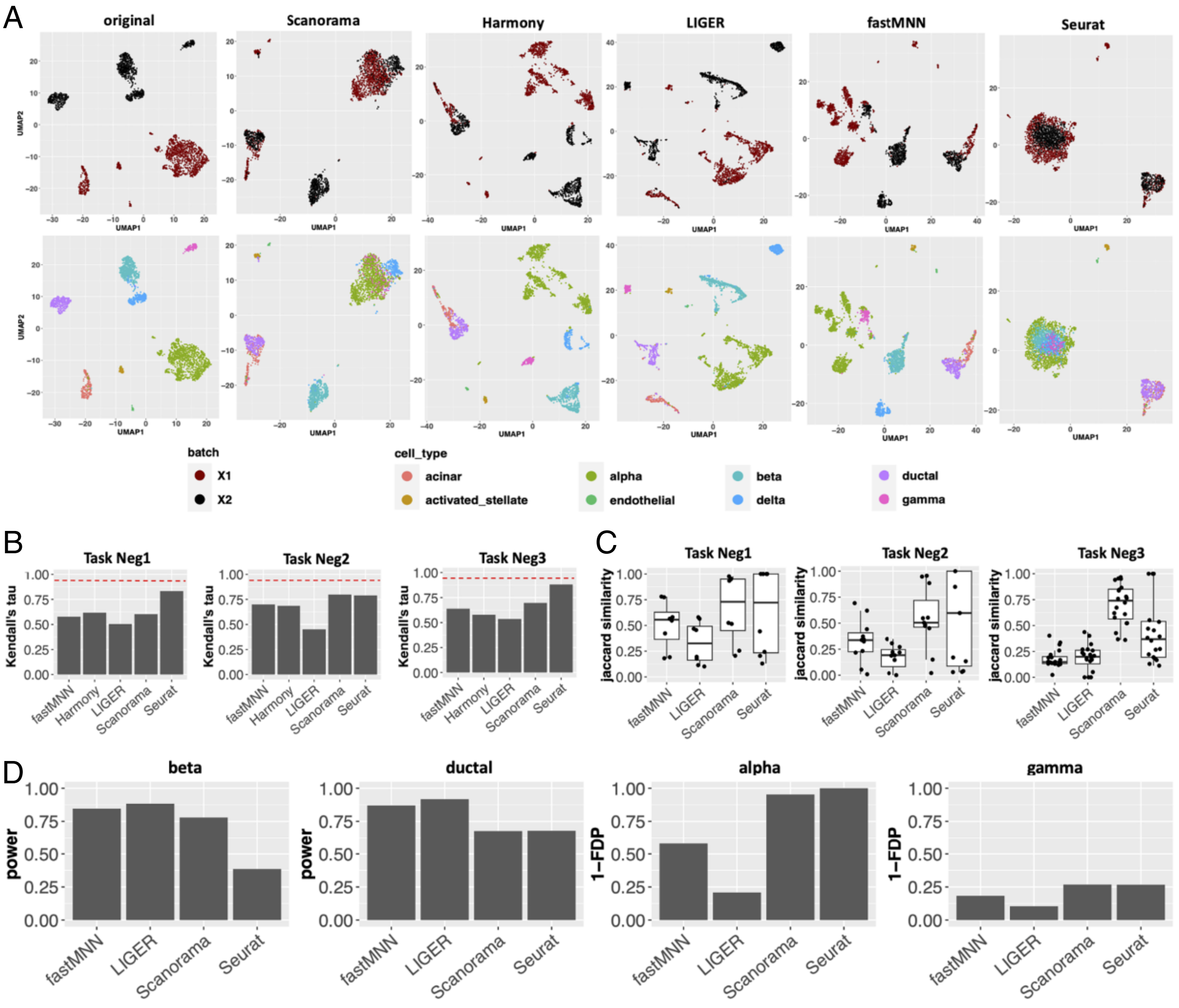

**Fig. 2.** Forcing uncertified data integration may cause false alignment, serious distortions and misleading inferences. (*A*) UMAP visualizations of the original (pooled) data under negative control task Neg1, and the integrated data as obtained by five popular methods (Scanorama, Harmony, LIGER, fastMNN, and Seurat). For each method, the *Top* figure is colored to indicate the distinct datasets being aligned, whereas the *Bottom* figure colored to indicate different cell types. See *SI Appendix*, Fig. S3 *A* and *B* for similar results about Pamona and SCOT, and the results about Neg2 and Neg3. (*B*) Under the three negative control tasks, we show barplots of Kendall's tau correlations between relative distances among the cells before integration and the distances after data integration, as achieved by each methods. The red dashed line benchmarks the average Kendall's tau correlation of 0.9 achieved by SMAI-align over the positive control tasks Pos1–Pos7. (*C*) Boxplots of Jaccard similarity between the set of differentially expressed (DE) genes associated with a distinct cell type detected based on the integrated data and the DE genes based on the original data. Each point represents a cell type. See *SI Appendix*, Fig. S3 *C* and *D* for similar results about Pamona and SCOT. (*D*) For Task Neg1, we show some representative barplots of (1−false discovery proportion) (1−FDP) and the power of detecting DE genes for some cell types based on the integrated data. Harmony is not included in (*C*) and (*D*) as its integration is only achieved in the low-dimensional space. Notably, SMAI-test correctly detects that all the datasets in Tasks Neg1–Neg3 are not alignable.

has overall better performance in preserving the within-data structures after integration (i.e., highest correlations between the relative distances of the cells before integration and the distances after integration), while achieving comparable if not better performance in removing the unwanted variations between the datasets (i.e., higher similarity in expression profiles for the same cell types across batches) (Fig. 3*A*). In particular, these features are consistent across multiple evaluation metrics, including Kendall's tau correlation and Spearman's rho correlation for structure preservation, and the D-B index (34) and the inverse Calinski–Harabasz (C-H) index (35) for batch effect removal (*SI Appendix*, Fig. S5). The advantages of these indices over other metrics such as LISI or ARI are explained in *Materials and Methods*. The desirable alignment achieved by SMAI and its advantages over the existing methods is further supported by visualizing low-dimensional embeddings of the integrated data. In Fig. 3*B* and *SI Appendix*, Fig. S6, we observe that across all six tasks, SMAI-align in general achieves good alignment of cells from different datasets under the same cell types. In contrast, distortions and misalignment of certain cell types are found for some existing methods. For instance, in Task Pos2, we observe false integration of gamma cells and beta cells by Harmony and Seurat, and significant distortion, that is, stretching and creation of multiple artificial subclusters, of the alpha cell cluster by LIGER and fastMNN (*SI Appendix*, Fig. S7*A*). As another example, for Task Pos4, we observe strong distortion and artificial split of the excitatory neurons and the inhibitory neurons, by Harmony, LIGER, and fastMNN (*SI Appendix*, Fig. S7*B*). In general, compared with the existing methods, the integrated datasets obtained by SMAI-align are overall of higher integration quality, and less susceptible to technical artifacts, structural distortions, and information loss, making them more reliable for downstream analyses.

Preserving and characterizing rare cell types is a desirable property of data integration methods. In this respect, SMAI is by

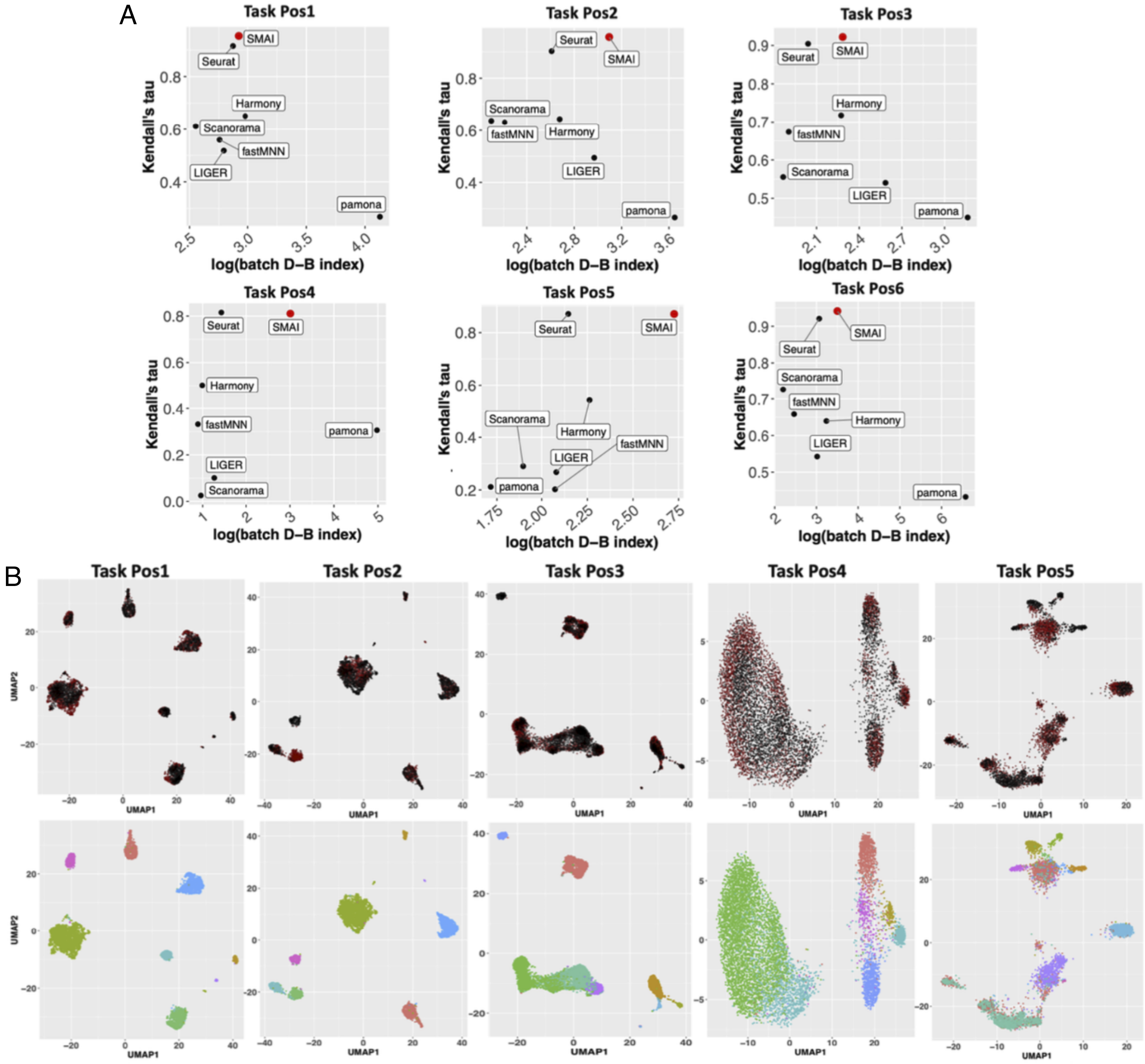

Fig. 3. Performance of SMAI-align on the six positive control integration tasks. (*A*) Compared with the six existing algorithms (black), SMAI-align (red) has an overall best performance in preserving the within-data structures after integration while achieving a competitive performance in removing the unwanted variations. The former is characterized by the highest Kendall's tau correlations between the relative distances of the cells within a dataset before integration and the distances after integration (*y*-axis), whereas the latter is reflected by higher values of the batch-associated Davies–Bouldin (D-B) index (*x*-axis shown in log-scale). See *SI Appendix*, Fig. S5 for more comparisons about additional datasets and metrics. (*B*) UMAP visualizations of the integrated data as obtained by SMAI-align. For each integration task, the *Top* figure is colored to indicate the two datasets being aligned, whereas the *Bottom* figure is colored to indicate different cell types. See also *SI Appendix*, Fig. S7 for UMAP visualizations associated with other integration methods.

design less likely to merge rare cell clusters with other cell types during alignment. To evaluate SMAI's performance in preserving and aligning rare cell types compared with the existing methods, we focus on two positive tasks with more cell types (Pos5 and Pos6), and assess for each rare cell type (containing less than 5% of the total cells) the performance of different methods in both structure preservation and batch-effect removal. Our results confirm the advantages of using SMAI in dealing with rare cell types (*SI Appendix*, Fig. S8).

**SMAI Improves Reliability and Power of Differential Expression Analysis.** A common and important downstream analysis following data integration is to identify the marker genes associated with individual cell types based on the integrated data (7, 8, 11). To demonstrate the advantage of SMAI-align in improving the reliability of downstream differential expression analysis, we focus on the first six positive control tasks and evaluate how many DE genes for each cell type are preserved after integration, and how many new DE genes are introduced after integration. Specifically, for each integrated dataset produced by fastMNN, LIGER, Scanorama, Seurat, or SMAI, we identify the DE genes for each cell type based on the Benjamini–Hochberg adjusted *P*-values, and compare their agreement with the DE genes identified from the individual datasets before

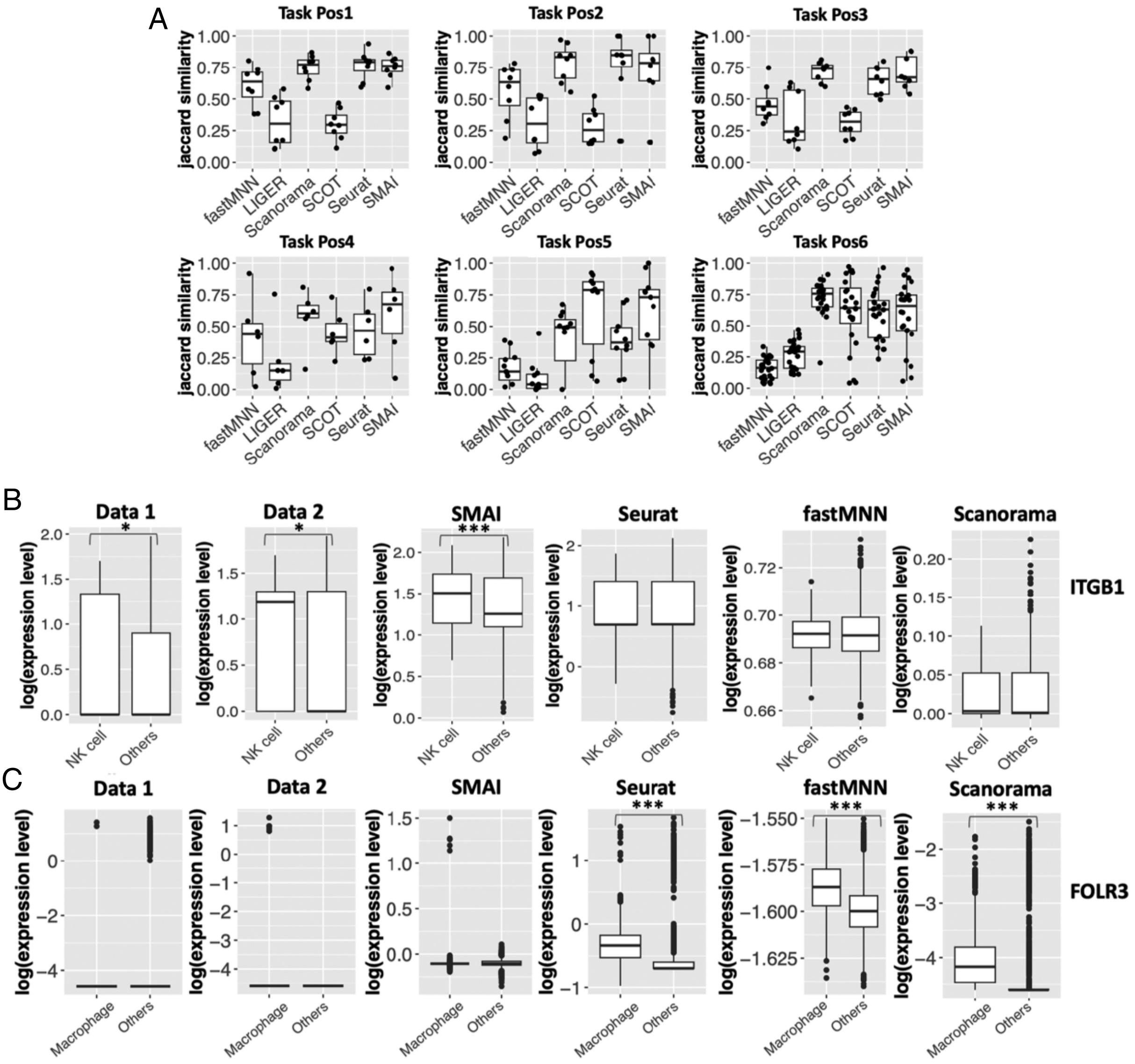

**Fig. 4.** SMAI improves reliability and power of DE analysis. (*A*) Boxplots of Jaccard similarity between the DE genes for each cell type identified based on the integrated data, obtained by one of the six integration methods, and the genes identified based on the individual datasets before integration. Each point represents a distinct cell type. The results indicate that SMAI-align oftentimes leads to more consistent and more reliable characterization of DE genes, as compared with other methods. (*B*) Boxplots of log-expression levels of ITGB1 as grouped by cell types in the two datasets about human PBMCs (Task Pos3: Data 1 contains 219 natural killer (NK) cells and 3,143 other cells, and Data 2 contains 194 NK cells and 3,028 other cells), and in the integrated datasets (413 NK cells and 6,171 other cells) as produced by SMAI-align, Seurat, fastMNN, and Scanorama. The DE pattern of ITGB1 is only preserved by SMAI after integration. (*C*) Boxplots of log-expression levels of FOLR3 as grouped by cell types in the two datasets about human lung tissues (Task Pos5: Data 1 contains 68 macrophages and 2,285 other cells, and Data 2 contains 911 macrophages and 1,000 other cells), and in the integrated datasets (979 macrophages and 3,285 other cells) as produced by SMAI-align, Seurat, fastMNN, and Scanorama. Artificial DE patterns are created by existing integration methods. The stars above the boxplots indicate statistical significance of DE test. Specifically, * means adjusted *P*-value $< 0.05$; ** means adjusted *P*-value $< 0.01$; *** means adjusted *P*-value $< 0.001$. Harmony and LIGER are not included in (*B*) and (*C*) as they do not produce gene-specific integrated data.

integration using the Jaccard similarity index, which accounts for both power and false positive rate in signal detection (*SI Appendix*, Notes). As a result, we find that compared with other methods, SMAI-align oftentimes leads to more consistent and more reliable characterization of DE genes based on the integrated data (Fig. 4*A*).

Biological insights can be obtained from the improved DE analysis with SMAI-align. For instance, under Task Pos3 concerning human PBMCs, an important protein coding gene ITGB1 (CD29), involved in cell adhesion and recognition (36), has been found DE in natural killer (NK) cells compared with other cell types in the SMAI-integrated data (adjusted *P*-value $< 10^{-11}$), but not in the Seurat-, fastMNN-, or Scanorama-integrated data. In both original datasets before integration, we also find statistical evidence supporting ITGB1 as a DE gene for NK cells (Fig. 4*B*). The functional relevance of ITGB1 to NK cells has been reported previously (37). In this case, the biological signal is blurred and compromised during data alignment by existing methods. See *SI Appendix*, Fig. S9 for similar examples. On the other hand, we also observe that SMAI-align is less likely to introduce artificial signals or false discoveries as compared with the existing methods. For example, under Task Pos5 concerning human lung tissues, we find almost no expression of the gene FOLR3 in macrophages in both datasets before integration. However, after integration, this gene is detected as DE gene based on the Seurat-, fastMNN-, and Scanorama-integrated datasets, but not based on the SMAI-integrated dataset (Fig. 4*C*). Similar examples are shown in *SI Appendix*, Fig. S10.

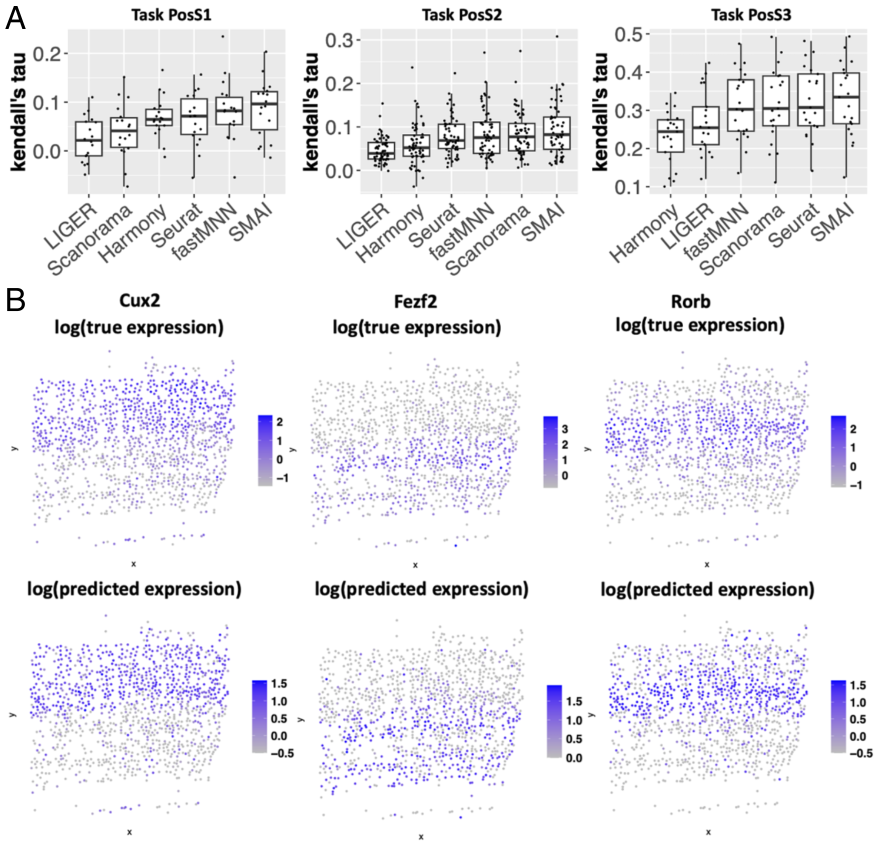

**Fig. 5.** SMAI improves prediction of single-cell spatial transcriptomic data. (*A*) Boxplots of Kendall's tau correlation between the actual expression levels of the spatial genes and their predicted values based on the two-step procedure (alignment followed by *k* nearest neighbor regression) where the data alignment is achieved by LIGER, Scanorama, Harmony, Seurat, fastMNN, or SMAI-align. Each point represents a distinct spatial gene. The methods are ordered according to their median predictive performance, showing the overall best performance of SMAI. (*B*) Examples of true expression levels of some spatial genes from Task PosS3, presented according to the cells' spatial layout, and their predicted values based on SMAI-align. The colors are in log scale.

**SMAI Improves Integration of scRNA-seq with Spatial Transcriptomics.** Another important application of data integration techniques is the imputation of the spatial expression levels of unmeasured transcripts in single-cell spatial transcriptomic data. Spatial transcriptomics technologies extend high-throughput characterization of gene expression to the spatial dimension and have been used to characterize the spatial distribution of cell types and transcripts across multiple tissues and organisms (38–40). A major trade-off across all spatial transcriptomics technologies is between the number of genes profiled and the spatial resolution such that most spatial transcriptomics technologies with single-cell resolution are limited to the measurement of a few hundred genes rather than the whole transcriptome (41). Given the resource-intensive nature of single-cell spatial transcriptomics data acquisition, computational methods for upscaling the number of genes and/or predicting the expression of additional genes of interest have been developed, which oftentimes make use of some paired single-cell RNA-seq data. Among the existing prediction methods, an important class of methods (8, 42–44) are based on first aligning the spatial and RNA-seq datasets and then predicting the expression of new spatial genes by aggregating the nearest neighboring cells in the RNA-seq data. Applications of these methods have been found, for example, in the characterization of spatial differences in the aging of mouse neural and glial cell populations (44), and recovery of immune signatures in primary tumor samples (45). As a key step within these prediction methods, we show that data alignment achieved by SMAI-align may lead to improved performance in predicting unmeasured spatial genes. To ensure fairness, we compare various prediction workflows which only differ in the data integration step (*Materials and Methods*). For the three spatial positive control tasks (PosS1–PosS3), we withhold each gene from the spatial transcriptomic data, and compare its actual expression levels with the predicted values based on the aforementioned two-step procedure where the data alignment is achieved by one of the six methods (LIGER, Scanorama, Harmony, Seurat, fastMNN, and SMAI-align). Due to the intrinsic difficulty of predicting some spatial genes that are nearly independent of any other genes, we only focus on predicting the first half of spatial genes that have overall higher correlations with other genes. Our analysis of the three pairs of datasets yields the best predictive performance of the SMAI-based prediction method (Fig. 5).

**SMAI's Interpretability Reveals Insights into the Sources of Batch Effects.** Unlike the existing methods, SMAI-align not only returns the aligned datasets, but also outputs explicitly the underlying alignment function achieving such alignment, which enables further inspection and a deeper understanding of possible sources of batch effects. Specifically, recall that the final alignment function obtained by SMAI-align consists of a scaling factor ($\widehat{\beta}$), a global mean-shift vector ($\widehat{\boldsymbol{\gamma}}$), and a rotation matrix ($\widehat{\mathbf{R}}$); each of them may contain important information as to the nature of the corrected batch effects. For example, applying SMAI-align to the human pancreas data (Task Pos1) leads to an alignment function in terms of ($\widehat{\beta}, \widehat{\boldsymbol{\gamma}}, \widehat{\mathbf{R}}$), linking the CEL-Seq2 dataset to

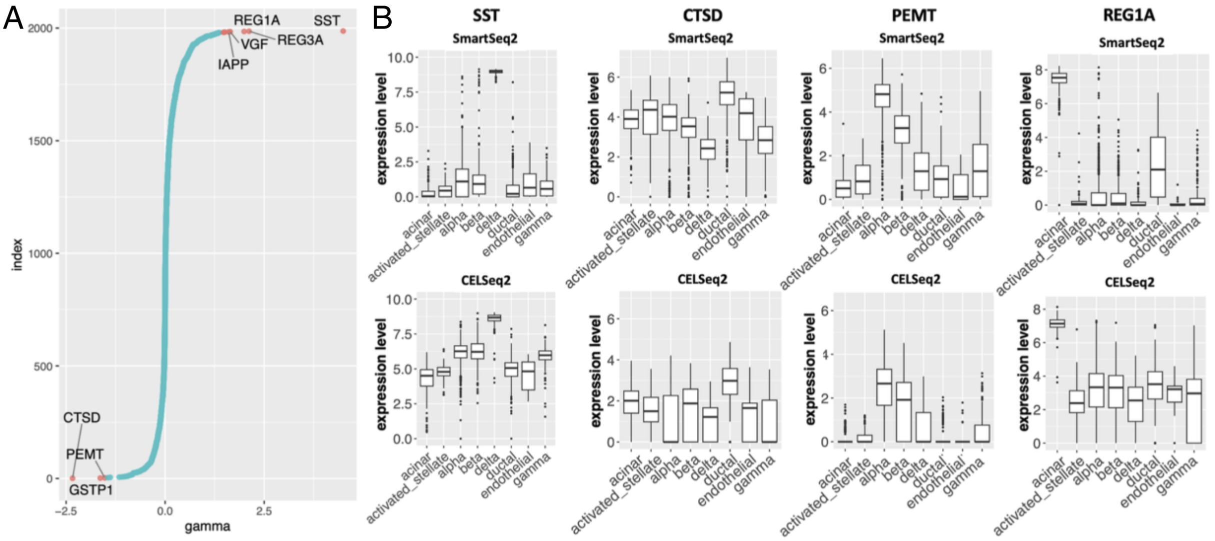

**Fig. 6.** SMAI's interpretability brings insights into the batch effects. (*A*) Visualization of the estimated mean-shift vector $\widehat{\gamma}$ from integrating the pancreatic data (Task Pos1) by SMAI-align, whose components are ordered from the smallest (*Bottom*) to the largest (*Top*). In $\widehat{\gamma}$ only a sparse set of genes such as SST, CTSD, PEMT, and REG1A are substantially affected by the batch effects. (*B*) Boxplots of expression levels of some genes highlighted in panel (*A*), grouped by different cell types. *Top*: Smart-Seq2 dataset with $n = 2{,}364$. *Bottom*: CEL-Seq2 dataset with $n = 2{,}244$. Note that SST, CTSD, PEMT, and REG1A are all DE genes. The batch effects on these genes are relatively uniform across cell types, and therefore SMAI-align does not affect DE results after integration.

the Smart-Seq2 dataset. The scaling factor $\widehat{\beta} = 1.01$ suggests little scaling difference between the two datasets. However, the obtained global mean-shift vector $\widehat{\boldsymbol{\gamma}}$ highlighted a sparse set of genes such as SST, CTSD, PEMT, and REG1A, affected by the batch effects (Fig. 6*A*). In particular, for both datasets, we observe similar patterns in the relative abundances of the transcript across different cell types (Fig. 6*B*), suggesting the batch effects on these genes are relatively uniform across cell types. Moreover, while SST, CTSD, PEMT, and REG1A are all DE genes associated with some cell types, SMAI-align does not affect DE patterns after integration due to its ability to identify and remove such global discrepancy. Similar observations can be made on other integration tasks such as human PBMCs (Task Pos3, *SI Appendix*, Fig. S12). As for the rotation matrix $\widehat{\mathbf{R}}$, it captures and removes the batch effects altering the gene correlation structures in each dataset (*SI Appendix*, Fig. S13). The obtained SMAI-align parameters $(\widehat{\beta}, \widehat{\boldsymbol{\gamma}}, \widehat{\mathbf{R}})$ can also be converted into distance metrics to quantify and compare the geometrically constitutive features of the batch effects (*SI Appendix*, Fig. S14*A*) and their overall magnitudes (*SI Appendix*, Fig. S14*B*).

## Discussion

We develop a spectral manifold alignment and inference algorithm that can determine general alignability between datasets and achieve robust and reliable alignment of single-cell data. The method explores the best similarity transformation that minimizes the discrepancy between the datasets, and thus is interpretable and able to reveal the source of batch effects. Despite SMAI being limited to modeling batch effects as linear transformations, our analyses of the ten positive control integration tasks suggest its practical suitability and competence in removing batch effects between datasets across diverse scenarios, oftentimes achieving superior performance compared with the existing nonlinear data integration methods such as Seurat or Harmony. In terms of computational time, standard implementation of SMAI requires a similar running time as existing methods such as Seurat (*SI Appendix*, Fig. S15*A*). The hyperparameters of SMAI, such as the number of informative eigenvalues for each data matrix, have been carefully determined (*SI Appendix*, Notes), and shown to work robustly in our benchmark datasets.

SMAI has a few limitations that deserve further development. First, although our restriction to the similarity class already yields promising performance over diverse applications, extending to and allowing for more flexible nonlinear transformations may lead to further improvement, especially in tracking and addressing local discrepancies associated with particular cell types. We describe two possible nonlinear extensions of SMAI in *SI Appendix*, Notes. Second, the current method only makes use of the overlapping features in both datasets. However, in many other applications, such as integrative analysis of single-cell DNA copy number variation data and single-cell RNA-seq data (46, 47), the features are related but not shared in general. To achieve integration across different modalities, one could first embed multimodal data into a common lower-dimensional embedding space using, for example, MOFA+(48), and then apply SMAI to the embedded data with common features. Third, although the current framework mainly concerns testing and aligning two datasets, direct extension to multiple datasets is available, which is achieved by applying SMAI in a sequential manner, based on some pre-specified order for integrating the datasets. Moreover, we point out that given the nature of our alignability test and the spectral alignment algorithm, extension to the simultaneous (non-sequential) testing and alignment of multiple datasets may be achieved by replacing the Procrustes analysis objective in Eq. **1** with a generalized Procrustes analysis objective involving all the available datasets and multiple alignment functions (31, 49). Expanding SMAI's scope of applications along these lines are interesting direction for future work.

## Materials and Methods

For clarity, here we introduce SMAI-test and SMAI-align along with their theoretical properties, focusing on the global alignment of two datasets with the same sample size. Extensions to aligning datasets with unequal sample sizes and SMAI for partial alignment, can be achieved by slightly modifying these basic algorithms, whose details are provided in *SI Appendix*, Notes.

**SMAI-Test Algorithm.** The basic SMAI-test algorithm requires as input the data matrices $\mathbf{X} \in \mathbb{R}^{d\times n}$ and $\mathbf{Y} \in \mathbb{R}^{d\times n}$, each containing $n$ cells and $d$ features (genes), a pre-determined parameter $r_{\max}$ corresponding to the number of leading eigenvalues to be used in the test, and a global rescaling factor $b > 0$. In principle, the number $r_{\max}$ should reflect the dimension of the underlying true signal structure of interest, whereas the global scaling factor $b > 0$ adjusts for the potential scaling difference between the two matrices. Estimators of $r_{\max}$ and $b$ are discussed in *SI Appendix*, Notes. The formal procedure of the test is summarized as follows:

1. Compute and normalize eigenvalues: Define the centered matrices $\bar{\mathbf{X}} = \mathbf{X}(\mathbf{I} - \mathbf{1}\mathbf{1}^\top)$ and $\bar{\mathbf{Y}} = \mathbf{Y}(\mathbf{I} - \mathbf{1}\mathbf{1}^\top)$. Let $\lambda_1^{(1)} \geq \lambda_2^{(1)} \geq \ldots \geq \lambda_n^{(1)}$ and $\lambda_1^{(2)} \geq \lambda_2^{(2)} \geq \ldots \geq \lambda_n^{(2)}$ be the ordered eigenvalues of matrices $d^{-1}\bar{\mathbf{X}}^\top\bar{\mathbf{X}}$ and $d^{-1}\bar{\mathbf{Y}}^\top\bar{\mathbf{Y}}$, respectively. Update $\lambda_i^{(1)} \leftarrow b\lambda_i^{(1)}$ for all $i = 1, 2, \ldots, n$.
2. Construct test statistic: i) define $\alpha_i^{(\ell)} = \big(\frac{1-n/d}{\lambda_i^{(\ell)}} + \frac{1}{d}\sum_{j=r_{\max}+1}^{d} \frac{1}{\lambda_j^{(\ell)}-\lambda_i^{(\ell)}}\big)^{-2}$, and $\phi_i^{(\ell)} = \frac{1}{\sqrt{\alpha_i^{(\ell)}}}\big(-\frac{1-n/d}{(\lambda_i^{(\ell)})^2} + \frac{1}{d}\sum_{j=r_{\max}+1}^{d} \frac{1}{(\lambda_j^{(\ell)}-\lambda_i^{(\ell)})^2}\big)^{-1}$ for $\ell = 1, 2$ and $i = 1, 2, \ldots, r_{\max}$.

   ii) define the test statistic $T_n = \sum_{i=1}^{r_{\max}} \frac{d(\lambda_i^{(1)}-\lambda_i^{(2)})^2}{2\alpha_i^{(1)}\phi_i^{(1)}+2\alpha_i^{(2)}\phi_i^{(2)}}$.
3. Obtain $P$-value: the $P$-value is defined as $1 - F^{-1}(T_n; r_{\max})$, where $F(\cdot; s)$ is the cumulative distribution function of the $\chi^2$ random variable with degree of freedom $s$.

**Theoretical Guarantee of SMAI-Test.** We first introduce our assumption on the centered matrices $\bar{\mathbf{X}}$ and $\bar{\mathbf{Y}}$, as well as the formal null hypothesis, based on which our testing procedure is developed. Suppose $\bar{\mathbf{X}} = \mathbf{Z}_1\mathbf{\Sigma}_1^{1/2}$ and $\bar{\mathbf{Y}} = \mathbf{Z}_2\mathbf{\Sigma}_2^{1/2}$, where $\mathbf{\Sigma}_1, \mathbf{\Sigma}_2 \in \mathbb{R}^{n\times n}$ are positive definite matrices, and $\mathbf{Z}_1, \mathbf{Z}_2$ are independent copies of random matrix $\mathbf{Z} = (z_{ij}) \in \mathbb{R}^{d\times n}$ with entries $z_{ij} = d^{-1/2}q_{ij}$ where the double array $\{q_{ij} : i = 1, 2, \ldots, d, j = 1, 2, \ldots, n\}$ consists of independent and identically distributed random variables whose first four moments match those of a standard normal random variable. We assume each of $\mathbf{\Sigma}_1$ and $\mathbf{\Sigma}_2$ has $r$ spiked/outlier eigenvalues, and the remaining bulk eigenvalues have some limiting spectral distribution. In other words, for each $\mathbf{\Sigma}_\ell, \ell = 1, 2$, there are exactly $r$ eigenvalues $\theta_1^{(\ell)} \geq \theta_2^{(\ell)} \geq \ldots \geq \theta_r^{(\ell)}$ larger than a certain threshold, characterizing the dominant global signal structures in the data, whereas for the rest of the eigenvalues $\theta_{r+1}^{(\ell)} \geq \theta_{r+2}^{(\ell)} \geq \ldots \geq \theta_n^{(\ell)}$ characterizing the remaining signal structures of much smaller magnitude, their empirical distribution (i.e., histogram) has the same deterministic limit as $n \to \infty$. The precise statements of these conditions are given as assumptions (A1)–(A3) in *SI Appendix*, Notes. Moreover, to account for the high dimensionality of the datasets, we assume that the number of genes is comparable to the number of cells, in the sense that $n/d \to c \in (0, \infty)$ as $n \to \infty$. The above model is commonly referred to as the high-dimensional generalized spiked population model (50–53), which is widely used for modeling high-dimensional noisy datasets with certain low-dimensional signal structures. In particular, the key assumption about the spiked eigenvalue structure can be empirically verified in our real single-cell datasets (*SI Appendix*, Fig. S1 and refs. 54 and 55). Essentially, such a model ensures the existence and statistical regularity of some underlying low-dimensional signal structure. For instance, it implies that when the signal strength of the low-dimensional structure is strong enough, that is, when top eigenvalues $\{\theta_i^{(\ell)}\}_{1\leq i\leq r}$ are large enough, the underlying low-dimensional signal structure would be roughly captured by the leading $r$ eigenvalues and eigenvectors of $d^{-1}\bar{\mathbf{X}}^\top\bar{\mathbf{X}}$ and $d^{-1}\bar{\mathbf{Y}}^\top\bar{\mathbf{Y}}$. Unlike many earlier works (56–59), where the bulk eigenvalues $\{\theta_i^{(\ell)}\}_{i>r}$ are assumed to be identical or all ones, the current framework allows for more flexibility as to the possible heterogeneity in the signal and/or noise structures.

Suppose the eigendecompositions of $\mathbf{\Sigma}_1$ and $\mathbf{\Sigma}_2$ are expressed as $\mathbf{\Sigma}_\ell = \mathbf{U}_\ell\mathbf{\Theta}_\ell\mathbf{U}_\ell^\top, \ell = 1, 2,$ where $\mathbf{\Theta} = \mathrm{diag}(\theta_1^{(\ell)}, \theta_2^{(\ell)}, \ldots, \theta_n^{(\ell)}\}$ is a diagonal matrix containing the ordered eigenvalues, and the columns of $\mathbf{U}_\ell = [\mathbf{u}_1^{(\ell)}, \mathbf{u}_2^{(\ell)}, \ldots, \mathbf{u}_n^{(\ell)}]$ are the corresponding eigenvectors. By definition, the (centered) low-dimensional structure associated with $\bar{\mathbf{X}}$ or $\bar{\mathbf{Y}}$ can be represented by the leading $r$ eigenvectors of $\mathbf{\Sigma}_\ell$ weighted by the square root of their corresponding eigenvalues, that is, $\mathbf{L}_\ell = \left[\sqrt{\theta_1^{(\ell)}}\cdot\mathbf{u}_1^{(\ell)} \quad \sqrt{\theta_2^{(\ell)}}\cdot\mathbf{u}_2^{(\ell)} \quad \ldots \quad \sqrt{\theta_r^{(\ell)}}\cdot\mathbf{u}_r^{(\ell)}\right] \in \mathbb{R}^{n\times r}$, where $\ell = 1, 2$. With these, the null hypothesis about the general alignability between $\mathbf{X}$ and $\mathbf{Y}$ can be formulated as the alignability between their low-dimensional structures $\mathbf{L}_1$ and $\mathbf{L}_2$ up to a possible rescaling and a rotation (note that translation is not needed as $\mathbf{L}_\ell$'s are already centered). Formally, the null hypothesis $H_0$ under the above statistical model states that "there exists a rotation matrix $\mathbf{R} \in \mathbb{R}^{n\times n}$ and a scalar $\beta > 0$ satisfying $\mathbf{L}_1 = \beta\mathbf{R}\mathbf{L}_2$." To develop a statistical test against such a null hypothesis, we notice that under $H_0$, it necessarily follows that $\theta_i^{(1)} = \beta\theta_i^{(2)}$ for all $1 \leq i \leq r$. As a result, it suffices to develop a statistical test that can evaluate if the spiked eigenvalues of $\mathbf{\Sigma}_1$ and $\mathbf{\Sigma}_2$ are identical up to a global scaling factor. This leads us to the proposed test. The next theorem, proved in *SI Appendix*, Notes, theoretically justifies the proposed test and ensures its statistical validity in terms of type I errors.

**Theorem 1.** *Suppose $\bar{\mathbf{X}}$ and $\bar{\mathbf{Y}}$ are independent and satisfy the above high-dimensional generalized spiked population model. Let $p(\bar{\mathbf{X}}, \bar{\mathbf{Y}}, r_{\max}, \beta)$ be the P-value returned by SMAI-test with $(r_{\max}, b) = (r, 1/\beta)$. Under the null hypothesis $H_0$, for any $0 < \alpha < 1/2$, it holds that $\lim_{(d,n)\to\infty} P_{H_0}(p(\bar{\mathbf{X}}, \bar{\mathbf{Y}}, r_{\max}, b) < \alpha) < \alpha$.*

**SMAI-Align Algorithm.** The basic SMAI-align algorithm requires as input the normalized data matrices $\mathbf{X}$ and $\mathbf{Y}$, the eigenvalue threshold $r_{\max}$, the maximum number of iterations $T$, and the outlier control parameter $k$. The algorithm starts with a denoising procedure during which the centered, best rank $r_{\max}$ approximations ($\bar{\mathbf{X}}$ and $\bar{\mathbf{Y}}$) of the data matrices are obtained. Multiple ways are available to denoise a high-dimensional data matrix with low-rank signal structures (60–62). To ensure both computational efficiency and theoretical guarantee, we adopt the hard-thresholding denoiser (*SI Appendix*, Notes). The second step is a robust iterative manifold matching and correspondence algorithm, motivated by the shuffled Procrustes optimization problem [**1**]. Spectral methods that generalize the classical ideas of shape matching and correspondence analysis in computer vision (63, 64) and the theory of Procrustes analysis (31, 49), are used to account for high dimensions. Specifically, the algorithm alternatively searches for the best basis transformation over the sample space and the feature space. In each iteration, the following ordinary Procrustes analyses are considered in order:

$$\min_{\mathbf{P}\in\mathbb{O}(n),\alpha\in\mathbb{R}} \|\bar{\mathbf{X}} - \alpha\mathbf{R}\bar{\mathbf{Y}}\mathbf{P}\|_F^2, \quad \text{with } \mathbf{R} \text{ given,} \qquad [2]$$

$$\min_{\mathbf{R}\in\mathbb{O}(d),\beta\in\mathbb{R},\boldsymbol{\gamma}\in\mathbb{R}^d} \|\bar{\mathbf{X}} - \beta\mathbf{R}\bar{\mathbf{Y}}\mathbf{P} - \boldsymbol{\gamma}\mathbf{1}_n^\top\|_F^2, \quad \text{with } \mathbf{P} \text{ given.} \qquad [3]$$

The first optimization looks for an orthogonal matrix $\mathbf{P} \in \mathbb{O}(n)$ and a scaling factor $\alpha \in \mathbb{R}$ so that the data matrix $\bar{\mathbf{X}}$ is close to the transformed data matrix $\alpha\mathbf{R}\bar{\mathbf{Y}}$, subject to recombination of its samples. In the second optimization, a similarity transformation $(\beta, \boldsymbol{\gamma}, \mathbf{R})$ is obtained to minimize the discrepancy between the data matrix $\bar{\mathbf{X}}$ and the sample-matched data matrix $\bar{\mathbf{Y}}\mathbf{P}$. The matrix $\mathbf{P}$ relaxes the original permutation class, allowing for more flexible sample matching between the two data matrices; $\mathbf{R}$ represents the rotation needed to align the features between the two matrices. To improve robustness against potential outliers in the data, in each iteration we remove the top $k$ outliers from both datasets, whose distances to the other dataset are the largest after alignment. $T$ and $k$ are tunable parameters, about which we find $3 \leq T \leq 5$ and $5 \leq k \leq 20$ work robustly for our benchmark datasets. In the last step, we determine the direction of alignment, that is, whether aligning $\mathbf{Y}$ to $\mathbf{X}$, or aligning $\mathbf{X}$ to $\mathbf{Y}$. In general, such directionality is less important as our similarity transformation is invertible and symmetric with respect to both datasets. As a default setting, our algorithm automatically determines the directionality by pursuing whichever direction that leads to a smaller between-data distance. Users can also specify the preferred directionality, which can be useful in some applications such as in the prediction of unmeasured spatial genes using single-cell RNA-seq data. Additional technical details of SMAI-align are provided in *SI Appendix*, Notes.

**Evaluation of Within-Data Structure Preservation and Batch Effect Removal.** For the negative control tasks Neg1–Neg3, and the positive control tasks Pos1–Pos7, we compare the overall correlation between the pairwise distances of cells before and after alignment. Specifically, suppose $\mathbf{X} \in \mathbb{R}^{d\times n_1}$ and $\mathbf{Y} \in \mathbb{R}^{d\times n_2}$ are the original normalized datasets, and $\mathbf{X}^* \in \mathbb{R}^{r\times n_1}$ and $\mathbf{Y}^* \in \mathbb{R}^{r\times n_2}$ are the aligned datasets obtained by one of the integration methods, where $r$ could be different from $d$. We first obtain $\mathcal{D}(\mathbf{X}) \in \mathbb{R}^{n_1\times n_1}$, whose entries contain pairwise Euclidean distances between the columns of $\mathbf{X}$. Similarly, we obtain $\mathcal{D}(\mathbf{Y})$, $\mathcal{D}(\mathbf{X}^*)$, and $\mathcal{D}(\mathbf{Y}^*)$. Let $[\mathcal{D}(\mathbf{X})]_{i.}$ be the $i$-th row of $\mathcal{D}(\mathbf{X})$. We calculate Kendall's tau correlations and Spearman's rho correlations between $[\mathcal{D}(\mathbf{X})]_{i.}$ and $[\mathcal{D}(\mathbf{X}^*)]_{i.}$, for all $i$'s, and then use the average correlation $\bar{r}_X$ across all $i$'s to quantify how well the structure within $\mathbf{X}$ is preserved after integration, in $\mathbf{X}^*$. Similarly, we also calculate the average correlation $\bar{r}_Y$ between the rows of $\mathcal{D}(\mathbf{Y})$ and the rows of $\mathcal{D}(\mathbf{Y}^*)$, to quantify the preservation of structure within $\mathbf{Y}$ after integration. Then we take the mean of $\bar{r}_X$ and $\bar{r}_Y$ as the final metric for the within-data structure preservation of the method, reported as the $y$-axis of the scatter plots in Figs. 2*B* and 3*A* and *SI Appendix*, Fig. S5.

Once the integrated data $(\mathbf{X}^*, \mathbf{Y}^*)$ are obtained, we calculate the D-B index and the inverse C-H index, to quantify how well the two integrated datasets are mixed. For a given integrated dataset consisting of $K$ batches $C_1, ..., C_K$, we define the D-B index $= \frac{1}{K}\sum_{k=1}^{K} \max_{j\neq k} \frac{S_k+S_j}{M_{ij}}$, where $S_i = [\frac{1}{|C_k|}\sum_{i\in C_k} \|X_i - A_k\|_2^2]^{-1/2}$, $M_{kj} = \|A_k - A_j\|_2$, with $A_k$ being the centroid of batch $k$ of size $|C_k|$, and $X_i$ being the $i$-th data point. We define the inverse C-H index $= \left[\frac{\sum_{k=1}^{K}\sum_{i\in C_k}\|X_i-A_k\|_2^2}{N-K}\right] \Big/ \left[\frac{\sum_{k=1}^{K}|C_k|\cdot\|A_k-A\|_2^2}{K-1}\right]$, where $A$ is the global centroid, and $N$ is the total sample size. These metrics essentially quantify the ratio between within-batch variations and between-batch variations. As a result, a method achieving better alignment quality will have a higher D-B index, and a higher inverse C-H index. These metrics are shown as the x-axis of the scatter plots in Fig. 3*A* and *SI Appendix*, Fig. S5, where we evaluate, for a given pair of datasets (i.e., fixing the within-batch variations), which alignment method leads to smaller between-batch variations. Compared with other metrics such as LISI and ARI, the D-B and inverse C-H indices are calculated from the pairwise distances, and do not rely on pre-specified nearest neighbor graphs (LISI), or the predicted cluster labels (ARI), which may be sensitive to specific methods.

**Differential Expression Analysis.** For each of the positive control tasks Pos1–Pos6, we first obtain an integrated dataset containing all the samples, by using one of the alignment methods. Then, for each cell type $k$, we identify the set of marker genes $S_k$ based on the aforementioned procedure. We use a threshold of 0.01 on the BH-adjusted $P$-values to select the DE genes. In the meantime, for each cell type $k$, we also obtain the benchmark set $S_k^b$ of marker genes, which contains all the DE genes identified based on the individual datasets before alignment. Finally, we compute the Jaccard similarity index between the set $S_k$ and the set $S_k^b$. The results are reported in Fig. 3*C*.

**Prediction of Spatial Genes.** For the three spatial positive control tasks PosS1–PosS3, we withhold each gene from the spatial transcriptomic data, and predict its values based on the following procedure. In particular, we denote $\mathbf{X} \in \mathbb{R}^{d\times n_1}$ as the spatial transcriptomic data, and $\mathbf{Y} \in \mathbb{R}^{d\times n_2}$ as the paired RNA-seq data with the same features. We also denote $\mathbf{X}_{-i.} \in \mathbb{R}^{(d-1)\times n_1}$ as the submatrix of $\mathbf{X}$ after removing the $i$-th row, and denote $\mathbf{X}_{i.} \in \mathbb{R}^d$ as the $i$-th row of $\mathbf{X}$. For each $1 \leq i \leq d$, we

1. Apply one of the alignment methods (LIGER, Scanorama, Seurat, fastMNN, or SMAI) to $\mathbf{X}_{-i.}$ and $\mathbf{Y}_{-i.}$, and obtain the aligned datasets $\mathbf{X}^*_{-i.}$ and $\mathbf{Y}^*_{-i.}$;
2. Fit a $k$-nearest neighbor regression (with $k = 5$ in all our analysis) between the predictor matrix $\mathbf{Y}^*_{-i.} \in \mathbb{R}^{(d-1)\times n_2}$ and the outcome vector $\mathbf{Y}_{i.} \in \mathbb{R}^{n_2}$;
3. Predict the outcome vector $\widehat{\mathbf{X}}_{i.}$ associated with the predictor matrix $\mathbf{X}^*_{-i.}$ based on the above regression model.

To evaluate the prediction accuracy, we calculate Kendall's $\tau$ correlation between $\mathbf{X}_{i.}$ and $\widehat{\mathbf{X}}_{i.}$.

**Cell Type Specific Alignment using SMAI.** To characterize possible intra-cell type variability, one can apply SMAI to each cell type separately, to learn the cell-type specific alignment. In this way, each cell type will have its own alignment function, characterizing the intra-cell type variability across different studies or conditions. In particular, the availability of a closed-form expression of the obtained alignment function allows for quantitatively comparing the similarity between the obtained alignment functions for different cell types. We show a scatter plot, where each point represents a cell-type specific alignment function, whose coordinates demonstrate the magnitude of the associated rotation and translation, obtained by converting the obtained rotation and translation parameters into normalized metrics between 0 and 1. Focusing on two positive integration tasks (Tasks Pos4 and Pos5), we applied SMAI to each cell type to learn the cell-type specific alignment functions. Interestingly, for each of the tasks, we found a remarkable similarity in the obtained cell-type specific alignment functions (*SI Appendix*, Fig. S14*C*), which supports the default design of SMAI using a common global alignment across cell types.

**Data Preprocessing.** The raw counts data listed in *SI Appendix*, Table S1 were filtered, normalized, and scaled by following the standard procedure (R functions CreateSeuratObject, NormalizeData and ScaleData under default settings) as incorporated in the R package Seurat. For datasets with more than 2,000 genes, we also applied the R function FindVariableFeatures in Seurat to identify the top 2,000 most variable genes for subsequent analysis. For the human pancreatic data associated with Task Pos1, we remove the cell types containing less than 20 cells. For the cross-tissue integration tasks (Neg3 and Pos6), a subset of 4,000 cells were randomly sampled from each of the tissues for our analysis.

**Data, Materials, and Software Availability.** The human pancreatic data can be accessed in the R package SeuratData (https://github.com/satijalab/seurat-data) under the dataset name panc8. The PBMC data can be accessed in the R package SeuratData (https://github.com/satijalab/seurat-data) under the dataset name pbmcsca. The mouse brain chromatin accessibility data were downloaded from Figshare (https://figshare.com/ndownloader/files/25721789), containing a dataset from Fang et al. (65) (single-nucleus ATAC-seq protocol), and a 10X Genomics dataset for fresh adult mouse brain cortex (sample retrieved from https://support.10xgenomics.com/single-cell-atac/datasets/1.2.0/atac_v1_adult_brain_fresh_5k). Both ATAC-seq datasets have been preprocessed by Luecken et al. (6) to characterize gene activities. The human lung data were downloaded as Anndata objects (samples 1, A3, B3 and B4) on Figshare (https://figshare.com/ndownloader/files/24539942). The human liver, MLN, and LLN immune cell data were downloaded from https://www.tissueimmunecellatlas.org. The mouse PBMC datasets (samples "Control 1h" and "LPS 1h") were downloaded from Gene Expression Omnibus (GSE178431) https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE178429. The mouse gastrulation seqFISH data were downloaded from https://content.cruk.cam.ac.uk/jmlab/SpatialMouseAtlas2020/, and the RNA-seq (10X Chromium) data can be accessed as "Sample 21" in MouseGastrulationData within the R package MouseGastrulationData. For the mouseVISP data, the ISS spatial transcriptomic data can be downloaded from https://github.com/spacetx-spacejam/data, the ExSeq spatial transcriptomic data can be downloaded from https://github.com/spacetx-spacejam/data, and the Smart-seq data can be downloaded from https://portal.brain-map.org/atlases-and-data/rnaseq/mouse-v1-and-alm-smart-seq. The R and Python packages of SMAI, and the R codes for reproducing our simulations and data analyses, are available at our GitHub repository https://github.com/rongstat/SMAI.

**ACKNOWLEDGMENTS.** We would like to thank the editors and two referees for their helpful comments and suggestions, which significantly improved the quality and presentation of the paper.

Author affiliations: [a]Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, MA 02115; [b]Department of Biomedical Data Science, Stanford University, Stanford, CA 94305; and [c]Department of Statistics, Stanford University, Stanford, CA 94305

1. C. Trapnell, Defining cell types and states with single-cell genomics. *Genom. Res.* **25**, 1491–1498 (2015).
2. A. Tanay, A. Regev, Scaling single-cell genomics from phenomenology to mechanism. *Nature* **541**, 331–338 (2017).
3. R. Elmentaite, C. Domínguez Conde, L. Yang, S. A. Teichmann, Single-cell atlases: Shared and tissue-specific cell types across human organs. *Nat. Rev. Genet.* **23**, 395–410 (2022).
4. D. Lähnemann *et al*., Eleven grand challenges in single-cell data science. *Genom. Biol.* **21**, 1–35 (2020).
5. H. T. N. Tran *et al*., A benchmark of batch-effect correction methods for single-cell RNA sequencing data. *Genom. Biol.* **21**, 1–32 (2020).
6. M. D. Luecken *et al*., Benchmarking atlas-level data integration in single-cell genomics. *Nat. Methods* **19**, 41–50 (2022).
7. T. Stuart *et al*., Comprehensive integration of single-cell data. *Cell* **177**, 1888–1902 (2019).
8. J. D. Welch *et al*., Single-cell multi-omic integration compares and contrasts features of brain cell identity. *Cell* **177**, 1873–1887 (2019).
9. I. Korsunsky *et al*., Fast, sensitive and accurate integration of single-cell data with harmony. *Nat. Methods* **16**, 1289–1296 (2019).
10. L. Haghverdi, A. T. Lun, M. D. Morgan, J. C. Marioni, Batch effects in single-cell RNA-sequencing data are corrected by matching mutual nearest neighbors. *Nat. Biotechnol.* **36**, 421–427 (2018).
11. B. Hie, B. Bryson, B. Berger, Efficient integration of heterogeneous single-cell transcriptomes using scanorama. *Nat. Biotechnol.* **37**, 685–691 (2019).
12. E. Mereu *et al*., Benchmarking single-cell RNA-sequencing protocols for cell atlas projects. *Nat. Biotechnol.* **38**, 747–755 (2020).
13. T. S. Consortium *et al*., The tabula sapiens: A multiple-organ, single-cell transcriptomic atlas of humans. *Science* **376**, eabl4896 (2022).
14. J. S. Ranek, N. Stanley, J. E. Purvis, Integrating temporal single-cell gene expression modalities for trajectory inference and disease prediction. *Genom. Biol.* **23**, 1–32 (2022).
15. R. Sugihara, Y. Kato, T. Mori, Y. Kawahara, Alignment of single-cell trajectory trees with capital. *Nat. Commun.* **13**, 5972 (2022).
16. E. Stephenson *et al*., Single-cell multi-omics analysis of the immune response in Covid-19. *Nat. Med.* **27**, 904–916 (2021).
17. D. J. Ahern *et al*., A blood atlas of Covid-19 defines hallmarks of disease severity and specificity. *Cell* **185**, 916–938 (2022).
18. T. Chari, J. Banerjee, L. Pachter, The specious art of single-cell genomics. *PLoS Comput. Biol.* **19**, e1011288 (2021).
19. S. M. Cooley, T. Hamilton, S. D. Aragones, J. C. J. Ray, E. J. Deeds, A novel metric reveals previously unrecognized distortion in dimensionality reduction of scRNA-seq data. bioRxiv [Preprint] (2019). https://doi.org/10.1101/689851 (Accessed 11 February 2024).
20. M. Büttner, Z. Miao, F. A. Wolf, S. A. Teichmann, F. J. Theis, A test metric for assessing single-cell RNA-seq batch correction. *Nat. Methods* **16**, 43–49 (2019).
21. S. E. Reese *et al*., A new statistic for identifying batch effects in high-throughput genomic data that uses guided principal component analysis. *Bioinformatics* **29**, 2877–2883 (2013).
22. G. Nyamundanda, P. Poudel, Y. Patil, A. Sadanandam, A novel statistical method to diagnose, quantify and correct batch effects in genomic studies. *Sci. Rep.* **7**, 10849 (2017).
23. F. Batool, C. Hennig, Clustering with the average silhouette width. *Comput. Stat. Data Anal.* **158**, 107190 (2021).
24. Z. Wu, H. Wu, Accounting for cell type hierarchy in evaluating single cell RNA-seq clustering. *Genom. Biol.* **21**, 1–14 (2020).
25. M. E. Ritchie *et al*., limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.* **43**, e47 (2015).
26. G. Finak *et al*., Mast: A flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data. *Genom. Biol.* **16**, 1–13 (2015).
27. X. Bai, L. Ma, L. Wan, Statistical test of structured continuous trees based on discordance matrix. *Bioinformatics* **35**, 4962–4970 (2019).
28. L. Aparicio, M. Bordyuh, A. J. Blumberg, R. Rabadan, A random matrix theory approach to denoise single-cell data. *Patterns* **1**, 100035 (2020).
29. M. Nitzan, M. P. Brenner, Revealing lineage-related signals in single-cell gene expression using random matrix theory. *Proc. Natl. Acad. Sci. U.S.A.* **118**, e1913931118 (2021).
30. S. Leviyang, A random matrix approach to single cell RNA-seq analysis. bioRxiv [Preprint] (2023). https://doi.org/10.1101/2023.06.28.546922 (Accessed 11 February 2024).
31. C. Goodall, Procrustes methods in the statistical analysis of shape. *J. R. Stat. Soc.: Ser. B (Methodological)* **53**, 285–321 (1991).
32. K. Cao, Y. Hong, L. Wan, Manifold alignment for heterogeneous single-cell multi-omics data integration using Pamona. *Bioinformatics* **38**, 211–219 (2022).
33. P. Demetci, R. Santorella, B. Sandstede, W. S. Noble, R. Singh, SCOT: Single-cell multi-omics alignment with optimal transport. *J. Comput. Biol.* **29**, 3–18 (2022).
34. D. L. Davies, D. W. Bouldin, A cluster separation measure. *IEEE Trans. Pattern Anal. Mach. Intell.* **PAMI-1**, 224–227 (1979).
35. T. Caliński, J. Harabasz, A dendrite method for cluster analysis. *Commun. Stat.-Theory Methods* **3**, 1–27 (1974).
36. R. O. Hynes, Integrins: versatility, modulation, and signaling in cell adhesion. *Cell* **69**, 11–25 (1992).
37. N. A. Bezman *et al*., Molecular definition of the identity and activation of natural killer cells. *Nat. Immunol.* **13**, 1000–1009 (2012).
38. M. Asp *et al*., A spatiotemporal organ-wide gene expression and cell atlas of the developing human heart. *Cell* **179**, 1647–1660 (2019).
39. L. Moses, L. Pachter, Museum of spatial transcriptomics. *Nat. Methods* **19**, 534–546 (2022).
40. R. Wei *et al*., Spatial charting of single-cell transcriptomes in tissues. *Nat. Biotechnol.* **40**, 1190–1199 (2022).
41. B. Li *et al*., Benchmarking spatial and single-cell transcriptomics integration methods for transcript distribution prediction and cell type deconvolution. *Nat. Methods* **19**, 662–670 (2022).
42. T. Abdelaal, S. Mourragui, A. Mahfouz, M. J. T. Reinders, SpaGE: Spatial gene enhancement using scRNA-seq. *Nucleic Acids Res.* **48**, e107 (2020).
43. C. Shengquan, Z. Boheng, C. Xiaoyang, Z. Xuegong, J. Rui, stPlus: A reference-based method for the accurate enhancement of spatial transcriptomics. *Bioinformatics* **37**, i299–i307 (2021).
44. W. E. Allen, T. R. Blosser, Z. A. Sullivan, C. Dulac, X. Zhuang, Molecular and spatial signatures of mouse brain aging at single-cell resolution. *Cell* **186**, 194–208.e18 (2023).
45. M. R. Vahid *et al*., High-resolution alignment of single-cell and spatial transcriptomes with cytospace. *Nat. Biotechnol.* **41**, 1543–1548 (2023).
46. V. Zachariadis, H. Cheng, N. Andrews, M. Enge, A highly scalable method for joint whole-genome sequencing and gene-expression profiling of single cells. *Mol. Cell* **80**, 541–553 (2020).
47. L. Yu *et al*., A single-cell multi-omics method enables simultaneous dissection of phenotype and genotype heterogeneity from frozen tumors. *Sci. Adv.* **9**, eabp8901 (2023).
48. R. Argelaguet *et al*., Mofa+: A statistical framework for comprehensive integration of multi-modal single-cell data. *Genom. Biol.* **21**, 1–17 (2020).
49. I. L. Dryden, K. V. Mardia, *Statistical Shape Analysis: With Applications in R* (John Wiley & Sons, 2016), vol. 995.
50. Z. Bai, J. Yao, On sample eigenvalues in a generalized spiked population model. *J. Multivar. Anal.* **106**, 167–177 (2012).
51. Z. Li, F. Han, J. Yao, Asymptotic joint distribution of extreme eigenvalues and trace of large sample covariance matrix in a generalized spiked population model. *Ann. Stat.* **48**, 3138–3160 (2020).
52. T. T. Cai, X. Han, G. Pan, Limiting laws for divergent spiked eigenvalues and largest nonspiked eigenvalue of sample covariance matrices. *Ann. Stat.* **48**, 1255–1280 (2020).
53. Z. Zhang, S. Zheng, G. Pan, P. S. Zhong, Asymptotic independence of spiked eigenvalues and linear spectral statistics for large sample covariance matrices. *Ann. Stat.* **50**, 2205–2230 (2022).
54. B. Landa, T. T. Zhang, Y. Kluger, Biwhitening reveals the rank of a count matrix. *SIAM J. Math. Data Sci.* **4**, 1420–1446 (2022).
55. B. Landa, Y. Kluger, The Dyson equalizer: Adaptive noise stabilization for low-rank signal detection and recovery. arXiv [Preprint] (2023). https://doi.org/10.48550/arXiv.2306.11263 (Accessed 11 February 2024).
56. I. M. Johnstone, On the distribution of the largest eigenvalue in principal components analysis. *Ann. Stat.* **29**, 295–327 (2001).
57. J. Baik, J. W. Silverstein, Eigenvalues of large sample covariance matrices of spiked population models. *J. Multivar. Anal.* **97**, 1382–1408 (2006).
58. D. Paul, Asymptotics of sample eigenstructure for a large dimensional spiked covariance model. *Stat. Sin.*, 1617–1642 (2007).
59. Z. Bao, X. Ding, J. Wang, K. Wang, Statistical inference for principal components of spiked covariance matrices. *Ann. Stat.* **50**, 1144–1169 (2022).
60. R. R. Nadakuditi, OptShrink: An algorithm for improved low-rank signal matrix denoising by optimal, data-driven singular value shrinkage. *IEEE Trans. Inf. Theory* **60**, 3002–3018 (2014).
61. M. Gavish, D. L. Donoho, Optimal shrinkage of singular values. *IEEE Trans. Inf. Theory* **63**, 2137–2152 (2017).
62. W. E. Leeb, Matrix denoising for weighted loss functions and heterogeneous signals. *SIAM J. Math. Data Sci.* **3**, 987–1012 (2021).
63. L. S. Shapiro, J. M. Brady, Feature-based correspondence: An eigenvector approach. *Image Vision Comput.* **10**, 283–288 (1992).
64. S. Belongie, J. Malik, J. Puzicha, Shape matching and object recognition using shape contexts. *IEEE Trans. Pattern Anal. Mach. Intell.* **24**, 509–522 (2002).
65. R. Fang *et al*., Comprehensive analysis of single cell ATAC-seq data with SnapATAC. *Nat. Commun.* **12**, 1337 (2021).