

RESEARCH

Open Access



Benchmarking deep learning methods for biologically conserved single-cell integration

Chenxin Yi^{1,2}, Jinyu Cheng^{2,3}, Jiajun Chen^{1,2}, Wanquan Liu^{1*}, Junwei Liu^{2*} and Yixue Li^{2,4,5,6,7,8*}

*Correspondence:

Wanquan Liu
liuwq63@mail.sysu.edu.cn
Junwei Liu
liu_junwei@gzlab.ac.cn
Yixue Li
li_yixue@gzlab.ac.cn

Full list of author information is available at the end of the article

Abstract

Background Advancements in single-cell RNA sequencing have enabled the analysis of millions of cells, but integrating such data across samples and methods while mitigating batch effects remains challenging. Deep learning approaches address this by learning biologically conserved gene expression representations, yet systematic benchmarking of loss functions and integration performance is lacking.

Results We evaluate 16 integration methods using a unified variational autoencoder framework, incorporating batch and cell-type information. Results reveal limitations in the single-cell integration benchmarking index (scIB) for preserving intra-cell-type information. To address this, we introduce a correlation-based loss function and enhance benchmarking metrics to better capture biological conservation. Using cell annotations from lung and breast atlases, our approach improves biological signal preservation. We propose a refined integration framework, scIB-E, and metrics that provide deeper insights into the integration process and offer guidance for advanced developments in integrating increasingly complex single-cell data.

Conclusions This benchmark highlights the potential of deep learning-based approaches for single-cell data integration, emphasizing the importance of biologically informed metrics and improved benchmarking strategies.

Keywords Single-cell RNA sequencing, Data integration, Deep learning, Batch correction, Biological conservation

Background

Single-cell RNA sequencing (scRNA-seq) has revolutionized our ability to study cellular diversity by providing high-resolution insights into gene expression at the single-cell level [1]. With the advancement of scRNA-seq techniques, the volume of single-cell data across various species, tissues, and developmental stages has expanded significantly, growing from hundreds of cells to tens of millions [2]. The scRNA-seq data analysis aims to understand cellular gene expression and functional variations across different diseases or developmental stages, helping to identify mechanisms underlying cell alterations [3]. Among that, data integration is a critical process for combining data collected from



© The Author(s) 2025. **Open Access** This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

different samples and time points and is also essential for incorporating external data with a similar biological background [4]. However, due to potential data biases, high data dimension, and sparsity in scRNA-seq data, integrating large-scale single-cell data from different experiments, studies, and platforms while preserving crucial biological insights remains a significant challenge [5, 6].

Several statistical methods have been developed to address batch effects in scRNA-seq data integration. One strategy is to identify the Mutual Nearest Neighbors (MNN) of single cells across datasets, including MNN [7], Scanorama [8], and Seurat V3 [9]. Another strategy focuses on balancing cellular neighbors to prevent batch-specific clustering, exemplified by Harmony [10] and batch-balanced k-nearest neighbors (BBKNN) [11]. Additionally, Non-Negative Matrix Factorization (NMF) is employed to identify datasets-shared factors for integration, as seen in LIGER [12]. scMerge [13] and scMerge2 [14] leverage stably expressed genes or pseudo-replicates to estimate the factor of interest and mitigate unwanted batch variation. While these methods are effective in aligning and integrating scRNA-seq data, they often encounter difficulties with large-scale datasets, particularly those exhibiting high cell-type heterogeneity across datasets [15].

Deep learning-based approaches have emerged as more powerful and flexible solutions for single-cell data integration [16]. The capability of deep learning methods in learning large, high-dimensional, and complex datasets advances its ability to obtain crucial biological variation [17]. Autoencoder is a versatile framework to learn the latent data representation of high-dimensional single-cell gene expression data. Li et al. developed the DESC method, which employs an autoencoder to infer unsupervised embedding of scRNA-seq data and for batch-invariant cell clustering analysis [18]. Another notable method is single-cell Variational Inference (scVI) [19], a fully probabilistic deep learning framework that accounts for both biological and technical noise in scRNA-seq data. scVI uses a conditional variational autoencoder (cVAE) framework to treat different batches as variables while preserving true biological gene expression information. Moreover, the deep learning framework facilitates more complex model designs and enhances information regularization. The SCALEX method introduces a batch-free encoder to project the batch-invariant embeddings across datasets [20]. Additionally, in atlas-level data integration, pre-defined cell type information can also be leveraged for data integration. Single-cell ANnotation using Variational Inference (scANVI) [21] extends scVI by incorporating pre-existing cell state annotations, improving the accuracy of cell type identification in new datasets through a semi-supervised learning approach. Similarly, methods including scDREAMER [22] and scDML [23] can also utilize predefined cell clustering information for semi-supervised batch-removal data integration.

The success of deep learning methods in single-cell data batch correction and integration largely depends on the design of the loss function. Generally, the objective of these methods is to remove unwanted batch effects while preserving biological information across single-cell datasets. Both batch effects and biological signals can be partially captured by batch labels and predefined cell-type labels, respectively. To address batch correction, techniques such as adversarial learning [22, 24] and information-constraining [25] methods are employed to minimize batch-specific information across datasets. For preserving biological information, strategies like supervised domain adaptation [26] and deep metric learning [23] are utilized to ensure cell-type label information is maintained

in the integrated data. While various loss function designs are effective in different methods, a horizontal comparison of the impact of distinct loss function combinations in single-cell data integration tasks is lacking. Additionally, in the context of single-cell integration performance benchmarking, the single-cell integration benchmarking (scIB) [27] framework primarily evaluates methods in two key areas: batch correction and biological conservation, based on the batch and cell-type labels. While scIB provides a robust foundation for performance evaluation, it falls short in fully capturing unsupervised intra-cell-type variation. As deep learning models continue to evolve, there is a growing need for more refined benchmarking metrics that can accurately assess both batch effect correction and the preservation of critical biological information.

In this study, we developed 16 deep-learning single-cell integration methods across three distinct levels within a unified variational autoencoder framework. These methods were designed to comprehensively evaluate the impact of different loss function combinations on data integration, utilizing batch information, cell-type information, or both jointly. By analyzing the effects of batch correction and biological conservation across varying loss function configurations, we identified that current benchmarking metrics and batch-correction methods fail to adequately capture intra-cell-type biological conservation. This finding was validated with multi-layered annotations from the Human Lung Cell Atlas (HLCA) [28] and the Human Fetal Lung Cell Atlas [29]. To address this gap, we introduced a correlation-based loss function to better preserve biological signals and refined existing benchmarking metrics by incorporating intra-cell-type biological conservation. We further validated this result with a differential abundance analysis of the integrated single-cell data. Our findings highlight the potential of deep learning methods for single-cell data integration, with the refined framework and benchmarking metrics offering deeper insights into the integration process. These advancements are poised to drive the development of deep learning methods for integrating increasingly complex multimodal and spatiotemporal single-cell data.

Results

Deep-learning-based integration of single-cell data

Single-cell data integration is essential for atlas-level single-cell data analysis, and the advent of deep learning methods has broadened the application of data integration, enabling a deeper understanding of diverse biological processes. In this study, we present a unified benchmarking framework that evaluates different loss function designs and information regularization strategies for the data integration task (Fig. 1A, B). We also reorganize existing benchmarking metrics and expand their applications to include batch correction and biological conservation at both inter-cell-type and intra-cell-type levels (Fig. 1C). Additionally, we validate our extended single-cell integration benchmarking (scIB-E) metrics with multi-layer cell annotations and developmental single-cell atlases. Furthermore, we demonstrate the utility of our framework by designing and validating a novel loss function specifically aimed at preserving intra-cell-type biological structure, as confirmed by differential abundance testing (Fig. 1D).

In this study, we used the scVI [19] and scANVI [21] models as the foundational deep-learning framework to evaluate the effectiveness of different loss function designs and regularization modules for single-cell data integration. Both methods utilize conditional variational autoencoders to embed single-cell gene expression data, with batch labels of

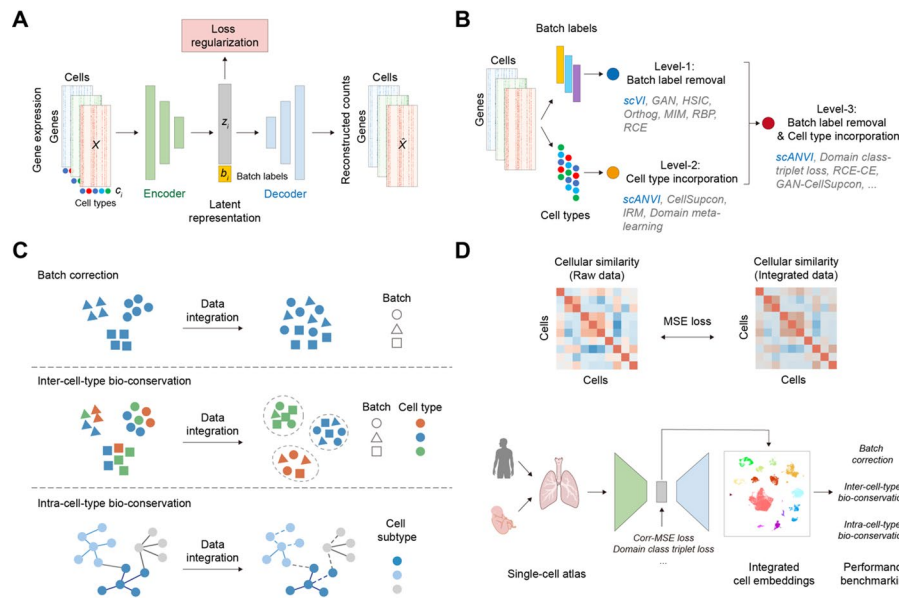


Fig. 1 Multi-level loss regularization designs for single-cell integration. **A** Diagram illustrating the unified variational autoencoder framework used in this project. **B** Schematic representation of the three-level loss designs implemented in this study. **C** Illustration of the effects of batch correction (top), inter-cell-type biological conservation (middle), and intra-cell-type biological conservation (bottom) following single-cell data integration. **D** Schematic representation of the Corr-MSE loss design (top) and the process of biologically conserved single-cell integration (bottom)

single-cell samples as conditional variables to remove batch effects. In scANVI, known cell-type labels are incorporated for semi-supervised data integration, enabling the preservation of biological information. Consequently, we applied both batch labels of single-cell samples and known cell-type labels of single cells as the proxies of both batch effects and biological information, respectively. Thereby, we developed a multi-level deep-learning strategy for single-cell data integration (Fig. 1B). Level-1 focuses on batch effect removal using batch labels, level-2 incorporates single-cell cell-type labels as biological conservation restraint, and level-3 integrates both batch and cell-type information for data integration. Overall, we designed 16 data integration methods based on the scVI framework across three levels, with the scVI method serving as the baseline for level-1, and the scANVI method as the baseline for level-2 and - 3.

Multi-level loss function designs for single-cell data integration

The level-1 methods are designed to eliminate batch effects by using the batch labels of different single-cell samples. These methods primarily focus on constraining the information shared between the learned latent embeddings of single-cell data and their batch labels. To achieve this, various loss functions are employed, including Generative Adversarial Network (GAN) [30], Hilbert-Schmidt Independence Criterion (HSIC) [25], Orthogonal Projection Loss (Orthog), Mutual Information Minimization (MIM), Reverse Backpropagation (RBP) [31] and Reverse Cross-Entropy (RCE) [32] (Methods). The level-2 methods incorporate known cell-type labels as proxy biological information to ensure that the latent embeddings from different batches remain biologically aligned. The loss functions at this level include Cell Supervised contrastive learning (CellSupcon) [33], Invariant Risk Minimization (IRM) [34], and Domain meta-learning [35] (Methods). In level-3 methods, we integrated both batch labels and cell-type labels

to simultaneously achieve batch-effect removal and biological conservation. Leveraging the flexibility of deep learning frameworks, we combine certain loss functions from level-1 and level-2 for level-3, and introduce an additional Domain Class Triplet loss [36] (Methods). The model training process follows the scVI framework, and the hyperparameters for the various methods were determined using the automated Ray Tune [37] framework (Methods). The hyperparameters used for the loss function combinations for each method are listed in Additional file 2: Tables S1 and S2.

We utilized multiple single-cell RNA-seq datasets to benchmark the performance of various loss designs, including datasets from immune cells [27], pancreas cells [38], and the Bone Marrow Mononuclear Cells (BMMC) dataset originating from the NeurIPS 2021 competition [39] (Additional file 2: Table S3). For performance evaluation, we used the single-cell integration benchmarking (scIB) [27] metrics, which provide quantitative evaluations of single-cell integration methods. The Uniform Manifold Approximation and Projection (UMAP) [40] was utilized to visualize the learned single-cell embeddings of different methods, highlighting the cell distributions across batches and cell types (Fig. 2A and Additional file 1: Figs. S1-S3). We assessed performance by comparing batch correction and biological conservation scores from scIB across multiple datasets (Fig. 2B and Additional file 2: Table S4). Compared to the scVI baseline, most level-1 methods demonstrated improved batch correction index, particularly in RBP regulation (Fig. 2C). For level-2 and level-3 methods, our results indicated that most methods achieved higher biological conservation scores compared to the scANVI baseline. Notably, the level-3 Domain Class Triplet loss, RBP-CellSupcon, and RCE-CE loss designs outperformed the baseline at both batch correction and biological conservation levels. Additionally, the level-2 IRM loss demonstrated robust biological conservation capability (Fig. 2C).

Assessing intra-cell-type biological information in data integration

Based on the performance results across various loss designs and levels of information regularization (Fig. 2), we observed that the strength of information regularization plays a pivotal role in single-cell data integration, affecting both scIB metrics for batch correction and biological conservation. To further evaluate the effects of different levels of information regularization, we applied the CellSupcon [33] loss function with varying hyperparameters (0, 10, 50, and 100), simulating the absence and increasing intensities of cell-type information regularization. Using the same immune dataset [27], we found stronger cell-type information regularization associated with more distinct separations among annotated single-cell subtypes in the resulting embeddings (Fig. 3A). Moreover, an analysis of scIB metrics across these hyperparameters revealed a gradient increase in the biological conservation score, accompanied by a slight decline in the batch correction score, ultimately resulting in an improvement in the scIB total metric with higher levels of cell-type regularization (Fig. 3B), indicating improved single-cell integration performance.

We then analyzed the trends across specific benchmarking indices within both scIB metric categories. For biological conservation metrics, all indices increased with stronger cell-type information regularization (Fig. 3C and Additional file 2: Table S5), suggesting that stronger cell-type regularization enhances biological conservation. In the batch correction category, we observed increases in Graph Connectivity and partial

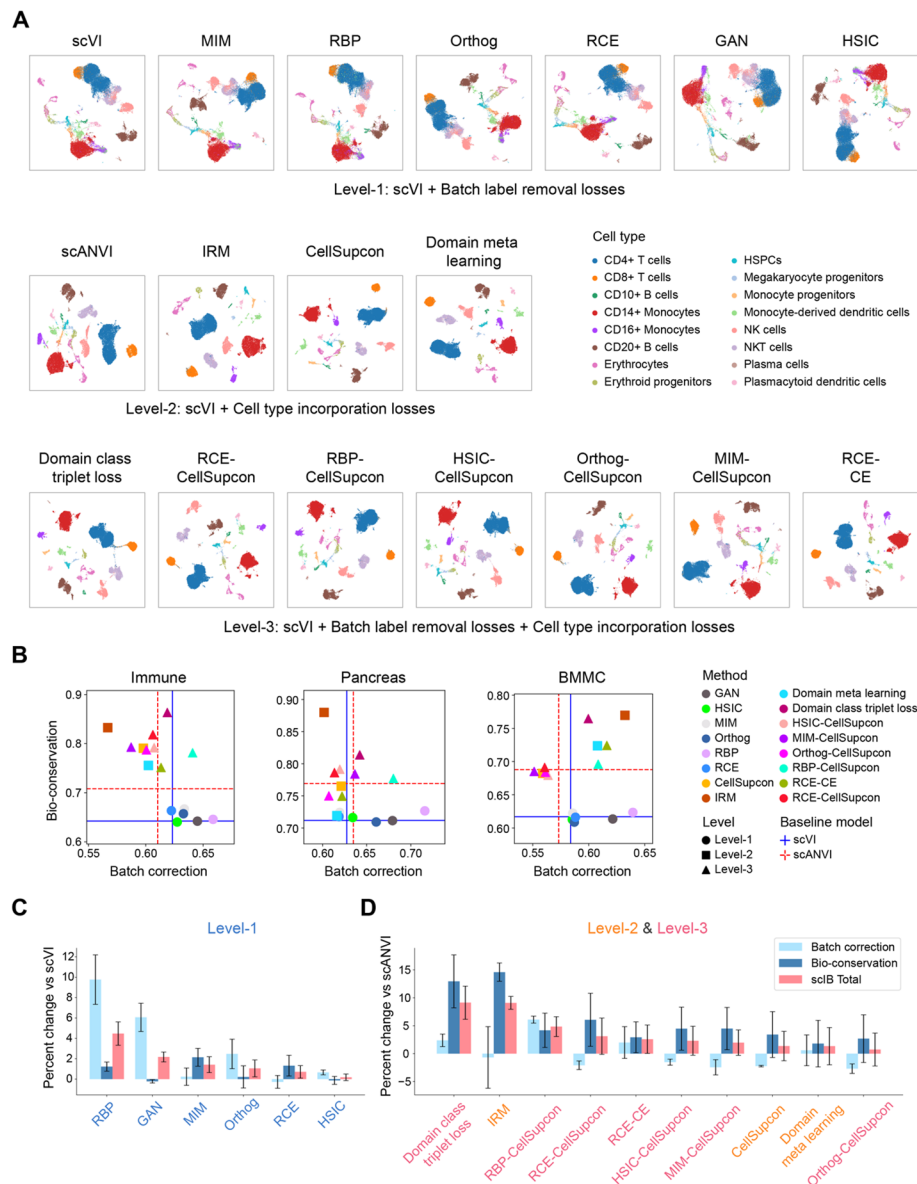


Fig. 2 Evaluation of multi-level loss functions for single-cell integration. **A** UMAP visualization of the immune dataset using three-level integration methods, with cells colored by cell type. **B** Comparison of the overall batch correction and bio-conservation scores across different methods by three datasets, with points colored by method and shaped by integration levels. Baseline performance of scVI and scANVI are marked by cross lines. **C-D**, Percentage margins of sclB metrics for level-1 (**C**) and level-2/level-3 (**D**) methods compared to scVI and scANVI baselines. Bars represent the average percentage margins for different scores, with error bars indicating the standard error of the mean. Methods are ranked in descending order based on their total sclB scores

improvements in iLISI and kBET scores. However, the Principal Component Regression (PCR) comparison index consistently decreased as cell-type information regularization intensified (Fig. 3D). Graph Connectivity, iLISI, and kBET indices all operate on the k-Nearest Neighbors (kNN) graph constructed from all single-cell data. Specifically, Graph Connectivity quantifies how well cells are connected within annotated cell types, while iLISI and kBET assess the distribution of batch labels across graph nodes (Methods). The observed increases in these indices suggest that embeddings of cells with the same cell-type labels become more similar across batches.

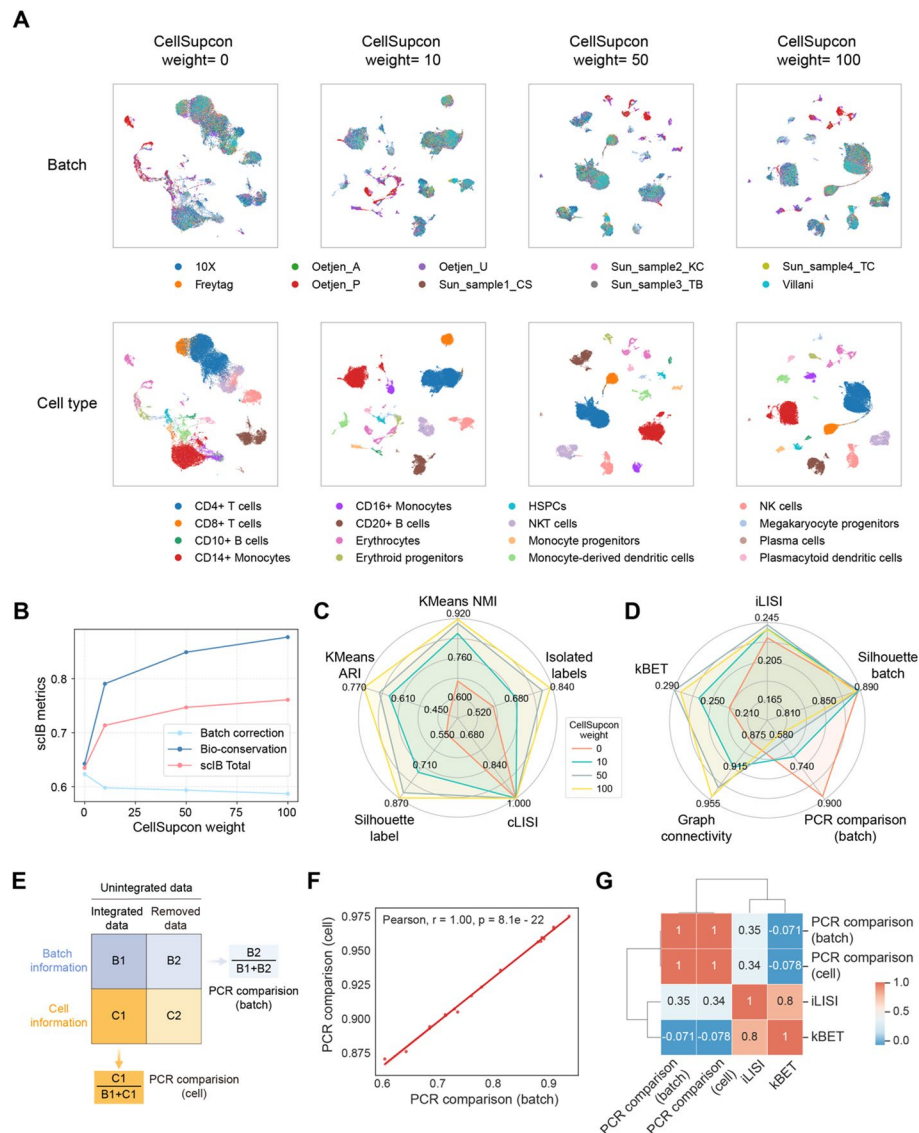


Fig. 3 Evaluation of intra-cell-type biological information across varying loss regularizations. **A** UMAP visualization of the immune dataset using methods with varying weights of the CellSupcon loss, with cells colored by batch label (top) and cell type (bottom). **B** Line plot showing scIB metrics of single-cell integration results as in (A) with different weights of the CellSupcon loss. **C–D**, Radar plots illustrating trends across detailed indices within biological conservation (C) and batch correction (D) categories of scIB metrics. Colors represent different weights of the CellSupcon loss. **E** Schematic representation of the design for PCR comparison indices (batch and-cell) to assess biological and batch information across integrated and unintegrated single-cell embeddings. **F** Pearson correlation analysis between PCR comparison-batch and PCR comparison-cell indices across single-cell integration results, as shown in Fig. 2A, for the immune dataset. The p-value is annotated. **G** Heatmap showing Pearson correlation coefficients between PCR comparison indices (PCR comparison-batch and PCR comparison-cell) and local cellular neighbor-based batch correction indices (iLISI and kBET)

The PCR index quantifies the variance in principle components of single-cell embeddings attributable to the batch variable, while the PCR comparison index evaluated the extent of batch information removed after batch correction (Methods). Since the PCR index reflects relative variance in single-cell embeddings, we hypothesized that the PCR comparison index for batch labels (PCR comparison-batch) is potentially inversely related to the preservation of biological information in integrated datasets. To test this, we designed a new PCR comparison index for cellular biological information (PCR

comparison-cell), which measures the proportion of cellular variance preserved in integrated single-cell data (Fig. 3E, Methods). Empirical correlation analyses confirmed the consistency between these PCR comparison indices, revealing that a reduction in the PCR comparison-batch with increasing regularization corresponds to a loss of underlying biological information (Fig. 3F). Meanwhile, the PCR comparison indices differed from the other batch correction indices, which primarily evaluate local cellular neighbors (Fig. 3G and Additional file 2: Table S6).

Correlation-based loss function for comprehensive biological conservation

The biological conservation metrics used in scIB are heavily dependent on pre-annotated cell-type labels. However, the PCR comparison-cell index revealed an opposite trend concerning these biological conservation metrics. This led us to hypothesize that the metrics might inadequately capture intra-cell-type biological variation, as they rely solely on the pre-annotated cell-type labels. The intra-cell-type biological variation reflects cellular diversity within annotated populations, encompassing potentially significant biological signals related to micro-populations or disease-specific subtypes.

To confirm that, we utilized a multi-layer annotated single-cell lung atlas featuring five levels of cellular annotations [28], ranging from broad cell populations to the finest resolved subsets. This framework allowed us to examine single-cell biological information at varying scales. Using the level-1 cell-type labels for model training, we assessed scIB indices across different methods with varying CellSupcon weights. To quantify the conservation of intra-cell-type biological variation, we analyzed the changes in scIB indices across cell annotation levels. Our findings revealed consistent trends in the batch correction index with varying CellSupcon weights across annotation levels (Fig. 4A). However, the biological conservation indices varied across cell levels, suggesting that current biological conservation metrics fail to fully capture cellular biological diversity (Fig. 4A and Additional file 2: Table S7).

To address this limitation, we introduced a correlation-based loss function, Correlation Mean Squared Error (Corr-MSE) Loss, designed to maintain the correlation similarity of single cells before and after data integration within each batch (Methods). Biological conservation scores that were evaluated at various annotation levels confirmed that this loss enhances the preservation of intra-cell-type biological variation (Additional file 1: Fig. S4A). Since the PCR comparison-batch index reflects the intra-cell-type biological variation, we also introduced a Jaccard index to quantify the ratio of edge connections among single cells before and after single-cell integration within a global kNN graph of each batch (Methods). We evaluated changes in the PCR comparison-batch index, Jaccard index, and other batch correction metrics, iLISI and kBET, by varying the weights of CellSupcon and Corr-MSE losses. Empirical results demonstrated that increasing the Corr-MSE weight enhances the conservation of intra-cell-type biological conservation while restraining batch correction metrics focused on local cell connections (Fig. 4B and Additional file 2: Table S7). In contrast, increasing the CellSupcon weight primarily improved the conservation of inter-cell-type information.

Extended scib metrics for comprehensive single-cell integration benchmarking

The scIB metrics established a set of benchmarks for single-cell integration tasks, focusing on batch correction and biological conservation. While these metrics are effective

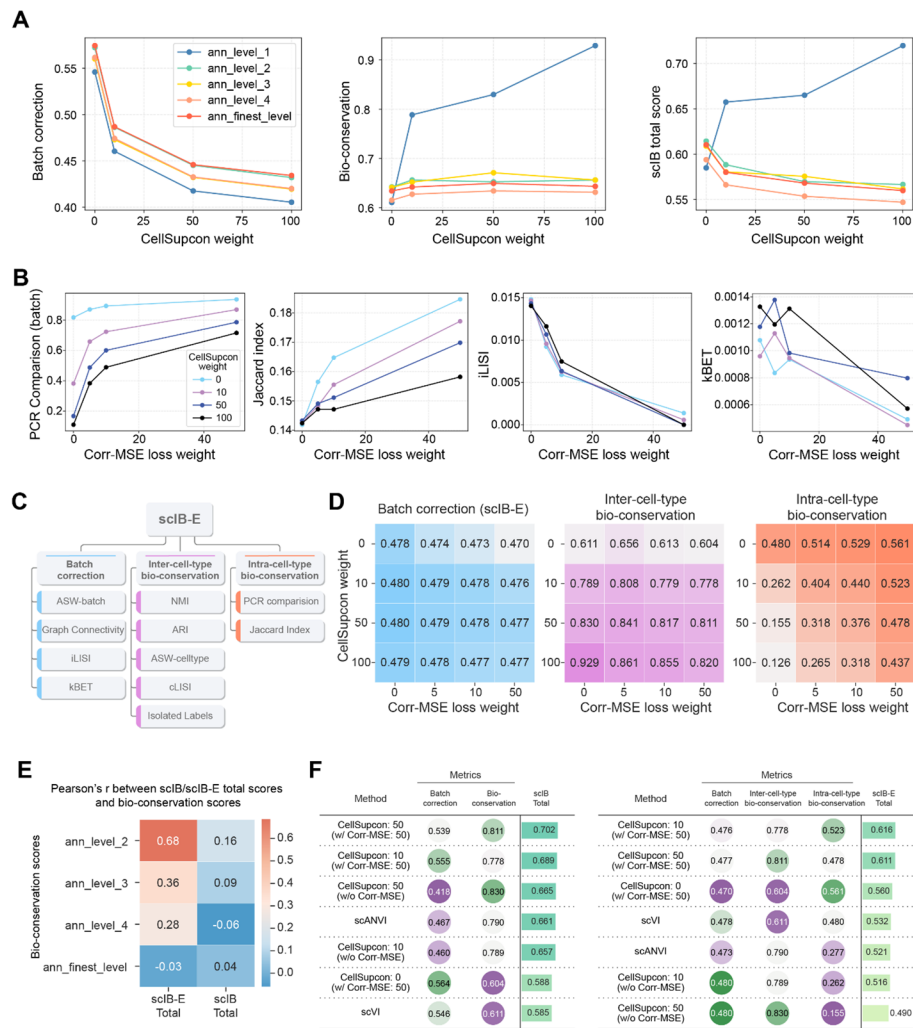


Fig. 4 Comprehensive benchmarking of single-cell integration using extended scIB metrics. **A** Line plots showing scIB metrics for the HLCA dataset integration results across varying CellSupcon loss weights: batch correction (left), biological conservation (middle), and total scIB score (right). The level-1 cell annotation was used for model training, while evaluations were performed with different levels of cell annotations. Colors indicate the levels of cell-type annotation used for evaluation. **B** Line plots depicting trends in PCR comparison-batch, Jaccard index, iLISI, and kBET indices for HLCA dataset integration results with varying CellSupcon and Corr-MSE loss weights. Colors represent different CellSupcon weights. **C** Schematic representation of the scIB-E metrics, designed for comprehensive single-cell integration benchmarking. **D** Heatmaps showing scIB-E scores for batch correction (left), inter-cell-type biological conservation (middle), and intra-cell-type biological conservation (right) across different configurations of Corr-MSE and CellSupcon weights in the HLCA dataset. **E** Heatmap showing Pearson correlation coefficients of scIB-E total scores (left) and scIB total scores (right) of HLCA dataset integration results using level-1 cell annotation, across paired bio-conservation scores estimated by different level cell annotations. **F** Tables summarizing scIB (left) and scIB-E (right) metrics for HLCA dataset integration results with varying CellSupcon and Corr-MSE loss weight configurations, alongside baseline performances of scVI and scANVI. A score of 1 represents optimal performance

for method benchmarking, they have limitations in assessing cell-label-free intra-cell-type biological conservation. To address this, we developed an extended version, scIB-E, which encompasses three categories: batch correction, inter-cell-type biological conservation, and intra-cell-type biological conservation (Fig. 4C, Methods). Specifically, PCR comparison-batch and Jaccard indices were introduced to evaluate intra-cell-type conservation. To assess the effectiveness of the new metrics across different methods, we compared scores across these categories using various Corr-MSE and CellSupcon weight

configurations applied to the single-cell lung atlas dataset (Fig. 4D). As previously noted, Corr-MSE primarily enhances intra-cell-type biological information conservation, while CellSupcon predominantly influences inter-cell-type bio-conservation (Additional file 2: Table S7). Additionally, we analyzed the total scores of scIB and scIB-E metrics to estimate multi-layer cell biological conservation estimation, the scIB-E total score exhibited higher Pearson correlation coefficients, indicating its superior ability to capture comprehensive biological variation in single-cell data (Fig. 4E and Additional file 1: Fig. S4B). Furthermore, compared to the original scIB metrics, scIB-E provided more stable batch correction scores and a clearer distinction between inter- and intra-cell-type bio-conservation scores (Fig. 4F). These findings suggest that a balanced combination of Corr-MSE and CellSupcon weights optimizes single-cell integration analysis, highlighting the advantages of scIB-E in advancing biological conservation assessment.

We then employed the scIB-E metrics to re-evaluate our previous benchmarking analysis involving different loss functions (Fig. 2) and to assess the effects of Corr-MSE loss on single-cell integration. The updated scIB-E metrics revealed that the Domain Class Triplet loss consistently outperformed other designs, both with and without Corr-MSE regulation, while the Level-2 IRM loss demonstrated substantial performance advantages under less stringent regulation (Fig. 5A and Additional file 1: Figs. S5-S7). By comparing index changes across all methods with and without Corr-MSE regulation, we confirmed that this regulation enhances intra-cell-type biological conservation by $6.61 \pm 5.49\%$, while slightly reducing the batch correction index by $-4.02 \pm 1.84\%$ and having no significant effect on inter-cell-type biological conservation (Fig. 5B). Overall, Corr-MSE regulation improves the scIB-E total index by $1.43 \pm 1.71\%$, suggesting that this approach enhances single-cell integration by better preserving comprehensive biological variation.

Enhanced single-cell integration methods for preserving biological information

To further evaluate single-cell integration methods by optimizing on preserving biological variation, we analyzed a multi-layer annotated single-cell dataset from the Human Fetal Lung Cell Atlas [29]. Our analysis compared scVI, scANVI, and two methods based on the Domain Class Triplet Loss, which was identified as the top-performing method in our previous analysis. We evaluated a baseline version against an enhanced Domain Class Triplet with Corr-MSE Loss (DCT-Corr) method, where the Corr-MSE loss was applied to specifically assess its impact on intra-cell-type variation regulation. For model training, broad cell type annotations were utilized as known cell labels, and UMAP projections were employed to visualize the developmental lung cell representations learned by different methods (Fig. 6A and Additional file 1: Fig. S8A). Performance was assessed using scIB-E metrics across three sub-categories. The DCT-Corr method outperformed baseline scANVI in preserving both inter and intra-cell-type biological structures. In contrast, scVI, which lacks cell label constraints, achieved the highest intra-cell-type bio-conservation score. Notably, integrating Corr-MSE Loss further enhanced intra-cell-type biological variation preservation, resulting in optimized single-cell integration performance (Fig. 6B).

For a detailed examination of biological variation preservation, we subsetting learned embeddings of fibroblast and distal epithelial cells across methods (Fig. 6A). To assess the biological variation within these embeddings, we focused on high-resolution cell labels and developmental stages of selected major cell types (Fig. 6C and Additional file

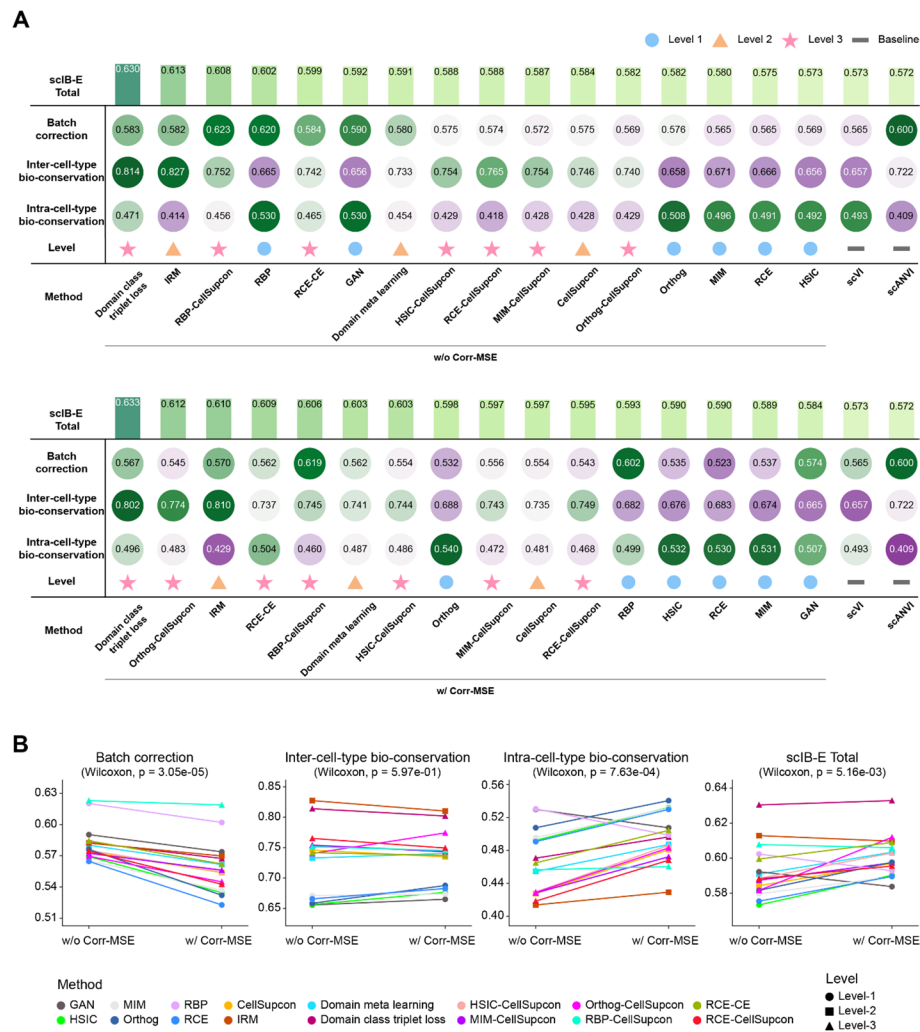


Fig. 5 Assessing multi-level loss functions for single-cell integration using scIB-E metrics. **A** Tables summarizing the average scIB-E scores across three datasets for multi-level loss functions, as described in Fig. 2. Results are shown without Corr-MSE loss (top) and with Corr-MSE loss (bottom). Methods are ranked in descending order by total score, with a score of 1 indicating optimal performance. **B** Paired line plots illustrating average scIB-E metrics across three datasets for batch correction, inter-cell-type biological conservation, intra-cell-type biological conservation, and total score, evaluated with and without Corr-MSE loss across multi-level loss functions. Colors represent distinct methods, and shapes denote different levels. Statistical significance was assessed using a paired Wilcoxon test, with p-values labeled

1: Fig. S8B). Clustering performance was evaluated using ARI and NMI indices, while developmental trajectory conservation was assessed with Moran's I and Geary's C indices (Additional file 2: Table S10 and Methods). Methods like scANVI and the baseline Domain Class Triplet loss relied on broad cell labels as biological constraints, limiting their ability to conserve intra-cell-type variation (Fig. 6D). In contrast, scVI achieved higher scores in developmental trajectory conservation (Fig. 6D). Furthermore, our DCT-Corr method demonstrated superior performance in local cell representation learning (Fig. 6D), as reflected in its high scores for broad cell type-level biological conservation (Fig. 6B). These findings highlight the versatility of the deep learning framework for single-cell integration. By leveraging different loss designs, this framework effectively balances comprehensive biological variation conservation and the removal of unwanted signals, achieving optimal performance for single-cell integration tasks.

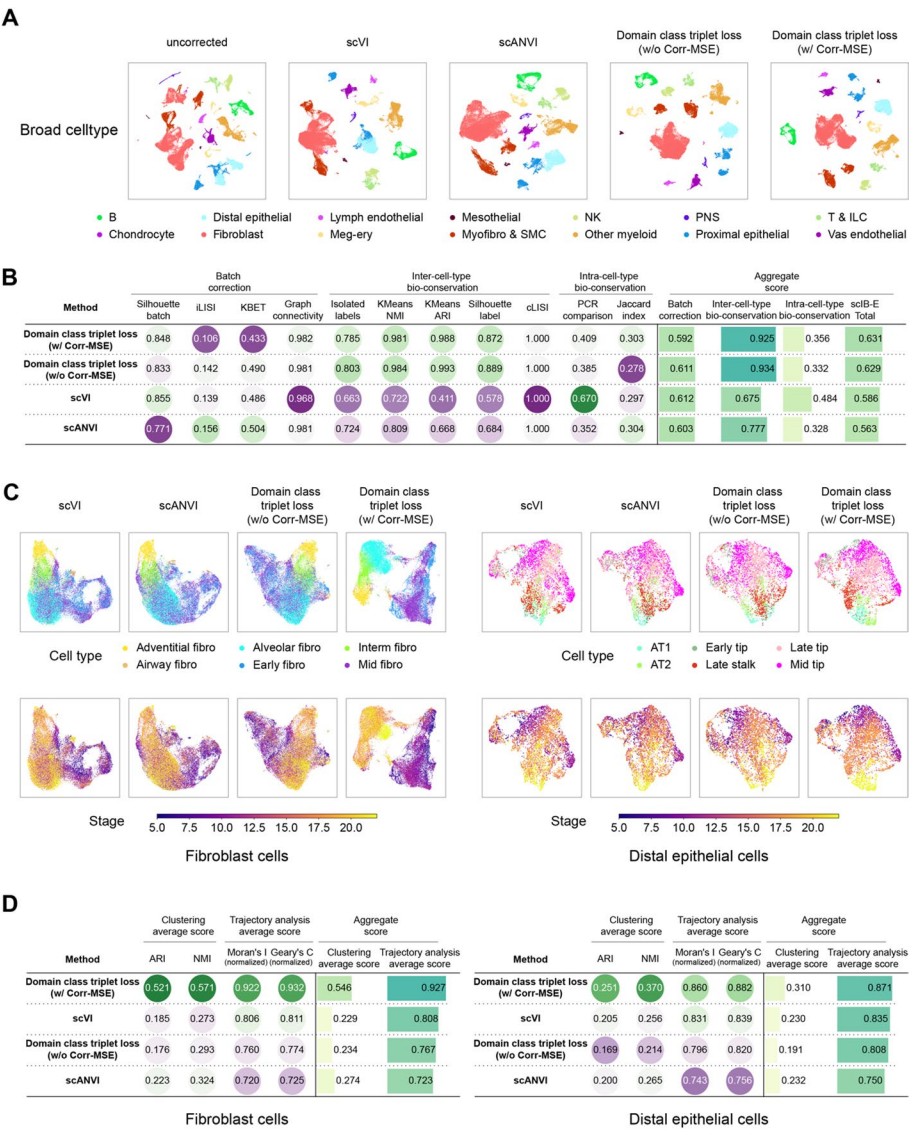


Fig. 6 Enhanced methods for biological preserving single-cell integration. **A** UMAP visualization of broad cell-type annotations for the Human Fetal Lung Cell Atlas, comparing the integration results of scVI, scANVI, the baseline Domain Class Triplet loss method, and the enhanced Domain Class Triplet with Corr-MSE loss (DCT-Corr) method. **B** Table summarizing the scIB-E metrics for integration results as shown in **(A)**. **C** UMAP visualization of learned single-cell embeddings for fibroblast (left) and distal epithelial cells (right) from the Human Fetal Lung Cell Atlas across different integration methods as in **(A)**, cells are colored by high-resolution cell labels (top) and developmental stages (bottom). **D** Tables summarizing clustering and developmental trajectory conservation performance of learned single-cell embeddings for fibroblast (left) and distal epithelial cells (right) as in **(C)**. Methods are ranked in descending order based on aggregated scores

Enhanced biological discovery in a multi-condition breast cancer atlas

To rigorously test our framework’s capacity for biological discovery, we applied it to a complex human breast cell atlas (HBCA) that maps cellular ecosystems across age, parity, and high-risk BRCA1/2 germline mutations [41]. We integrated this dataset using our top-performing loss function, DCT-Corr, alongside two baseline methods, scVI and scANVI. All models were trained using processing date as the batch covariate, and broad cell-type labels were provided to scANVI and DCT-Corr. Our method demonstrated superior performance, achieving the highest overall scIB-E and scIB score (Fig. 7A, Additional file 1: Fig. S9A, Additional file 2: Tables S11-S12), which reflects a more

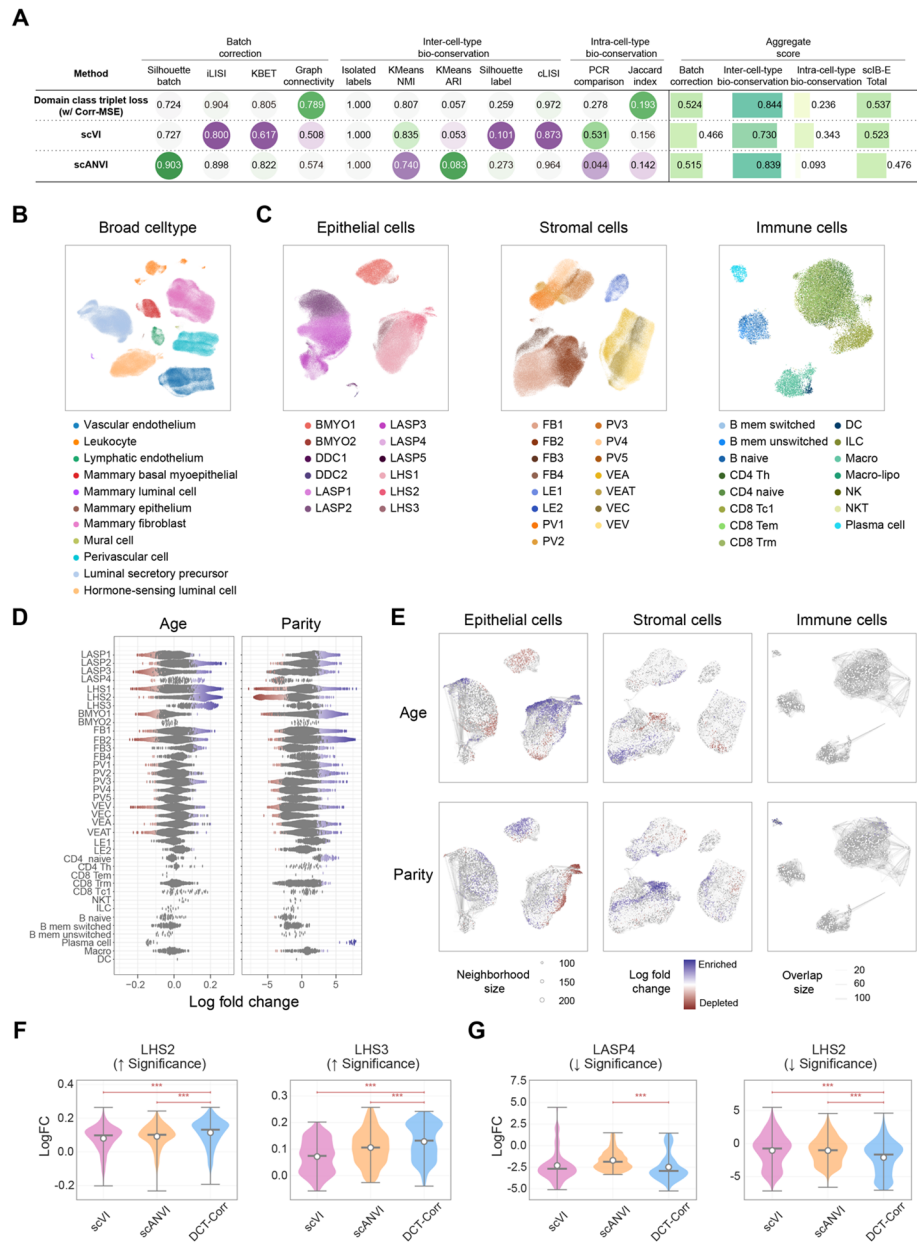


Fig. 7 Enhanced biological discovery in the Human Breast Cell Atlas (HBCA) with DCT-Corr integration. **A** Table summarizing the scIB-E metrics for the HBCA integration results, comparing the performance of the DCT-Corr method against scVI and scanVI baselines. **B** UMAP visualization of the integrated HBCA dataset using the DCT-Corr method, with cells colored by broad cell-type annotations. **C** UMAP projections showing detailed subcluster annotations within the epithelial, stromal, and immune compartments. **D** Beeswarm plot illustrating the log fold change of cell abundance across all subclusters for the age and parity comparisons from the differential abundance analysis. **E** UMAP visualizations of the three main cellular compartments, with neighborhoods colored by their log fold change to show enrichment or depletion of cell populations in the age and parity analyses. **F–G** Comparison of log fold changes for key epithelial subpopulations, showing results from the age analysis for LHS2 and LHS3 (F) and the parity analysis for LASP4 and LHS2 (G), evaluated across embeddings from scVI, scanVI, and DCT-Corr. Upward (↑) and downward (↓) arrows indicate a significant increase or decrease in cell abundance, respectively. Statistical significance is denoted by asterisks (* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$)

effective balance between removing technical artifacts and preserving biological variation (Fig. 7B-C, Additional file 1: Fig. S10A-B, Fig. S11A-B).

We then performed a differential abundance (DA) analysis to compare how well each integration captured subtle cell population changes across different conditions, mirroring the analysis in Reed et al. [41]. The DCT-Corr integration demonstrated exceptional sensitivity in resolving cellular state shifts within epithelial cells, which are central to breast biology. For instance, in the context of aging, DCT-Corr accurately captured significant changes in the hormone-sensing luminal populations LHS2 and LHS3 (Fig. 7D-F, Additional file 2: Table S13), consistent with the original study's conclusions. In the analysis of parity, while all methods identified the enrichment of BMYO1 and depletion of LSP4 cells, DCT-Corr uniquely revealed a more pronounced dysregulation in the LHS2 subpopulation (Fig. 7D-E, G, Additional file 2: Table S14). This finding suggests a previously underappreciated role for LHS2 cells in the extensive tissue remodeling associated with parity, offering a more refined biological insight.

Beyond resolving dynamics within a single lineage, DCT-Corr provided a more holistic view of the complex interplay between different cellular compartments. In high-risk BRCA1 carriers, the original study described an early immune-escape mechanism, where pro-inflammatory CD8 T cells are enriched while immune-suppressive programs are concurrently upregulated in epithelial cells. The DCT-Corr model successfully recapitulated this multi-compartment signal by not only identifying enrichment of CD8 Tc1 and CD8 T_{reg} immune cells but also powerfully co-highlighting significant responses within the BMYO2 and LHS1 epithelial populations (Additional file 1: Fig. S9B-D, Additional file 2: Table S15). Similarly, in the analysis of changes driven by BRCA2 mutations, the model effectively captured the key cross-compartment signals: a reduction of the epithelial LHS2 subpopulation and a concurrent enrichment of the stromal VEAT population (Additional file 1: Fig. S9B-C, E, Additional file 2: Table S16). Collectively, these results demonstrate that the improved integration by DCT-Corr translates directly to enhanced biological discovery in complex, multi-condition single-cell datasets.

Discussion

In this study, we introduce a multi-level benchmarking framework for single-cell data integration. By leveraging a unified deep learning architecture, we systematically evaluated the impact of diverse loss functions and regularization strategies on integration outcomes. Our work extends foundational benchmarking studies [15, 27, 42] by providing a modular analysis of the components driving model performance. The results demonstrate that targeted loss functions are critical for achieving an optimal balance between batch effect removal and biological signal conservation. Furthermore, our framework comprehensively assesses information preservation by disentangling technical artifacts from crucial inter- and intra-cell-type biological variance. Informed by these findings, we propose the scIB-E evaluation framework and a Corr-MSE loss function, both designed to more accurately assess and perform integration on complex single-cell datasets.

A key insight from our work is the importance of assessing intra-cell-type biological conservation, an area where existing methods have limitations. Specifically, common batch correction evaluations reliant on metrics like iLISI and kBET primarily focus on local cell neighbors, lacking the ability to preserve structural biological variation within

batches. To address this gap, we incorporated PCR comparison and Jaccard indices as complementary approaches for evaluating the conservation of biological information. Additionally, by incorporating the Corr-MSE loss to maintain correlation similarities within each batch, our analyses demonstrated improved conservation of biological variation. Using the multi-layer annotated lung atlases, we assessed various methods for intra-cell-type biological conservation, leveraging high-resolution single-cell labels and biological statuses not used in model training as proxies for intrinsic biological variation. The distinction between inter- and intra-cell-type bio-conservation scores offered by scIB-E demonstrated its superior capacity for comprehensively evaluating single-cell data integration.

A central challenge in single-cell integration is balancing the removal of technical variation with the preservation of biological heterogeneity. Our findings indicate that the performance of integration methods is critically dependent on the degree of information regularization applied. Stronger regularization, such as the CellSupcon [33] loss, improves biological conservation at inter-cell-type levels but can compromise intra-cell-type information and lead to over-correction [43]. This trade-off underscores the need for carefully calibrated approaches. We demonstrate that combining Corr-MSE with CellSupcon [33] achieves a more nuanced integration, mitigating batch effects while retaining subtle biological variations. Furthermore, the Domain Class Triplet loss [36] emerged as an effective strategy for robust biological conservation in complex integration tasks. Although we provide optimized hyperparameters and a weighted scIB-E metric (Methods), we recognize that fixed weights may be insufficient for all integration scenarios. Ultimately, optimal integration is not achieved by maximizing a single metric but by attaining a favorable balance across multiple factors, including technical artifacts and known biological patterns, as proposed by the scIB-E framework.

It is crucial to acknowledge the potential risk of information leakage when employing supervised or semi-supervised integration strategies that utilize cell-type labels during model training. As noted, evaluating performance with metrics that also rely on these same labels can lead to artificially inflated scores. This circularity may produce an embedding that is well-structured according to the provided annotations but lacks generalizability, potentially obscuring novel or unannotated cell states and failing to generalize to datasets with different compositions. Similarly, while the Corr-MSE loss improves the preservation of biological variation, the intra-batch biological signal can still be obscured by technical noise or other unwanted sources of variation. These limitations highlight the need for methods that can better disentangle true biological signals from batch-specific artifacts. Potential solutions include reference-based mapping to integrate a new query dataset into a common atlas-level context [44, 45], as well as emerging approaches like contrastiveVI [46] and DA-seq [47] that aim to isolate condition-specific biological signals.

Conclusions

In this study, we develop and validate a comprehensive benchmarking framework, scIB-E, to enable a more rigorous evaluation of single-cell data integration methods, with a particular focus on preserving biological information. We introduce the Corr-MSE loss function as a novel tool to specifically maintain the intra-cell-type biological variation often lost during integration. Our findings demonstrate that this framework, combined

with tailored loss functions, provides a more nuanced approach to balancing batch effect removal with the conservation of complex biological signals, thereby enhancing the ability to uncover subtle cellular processes in large-scale datasets. Future work could expand these frameworks to other single-cell modalities or experimental factors to further enhance the extraction of true biological insights. Collectively, this work underscores the value of flexible deep learning frameworks and well-designed loss functions in pushing the boundaries of single-cell data analysis, contributing to the creation of high-quality single-cell atlases that can facilitate deeper biological insights.

Methods

A unified benchmarking framework for single-cell integration

To systematically evaluate various loss functions, we established a unified benchmarking framework wherein all 16 methods were implemented upon a common deep generative model architecture. The framework is built on a conditional variational autoencoder (cVAE) that learns latent representations by conditioning the generative process on variables such as batch labels. The decoder models the count-based nature of scRNA-seq data using a zero-inflated negative binomial (ZINB) distribution.

scVI The scVI (Single-cell Variational Inference) model [19] served as the foundational baseline for our Level-1 methods. As a probabilistic framework, it generates low-dimensional embeddings suitable for batch correction and other downstream analyses. We selected scVI to benchmark strategies focused exclusively on batch effect removal.

scANVI The scANVI (Single-cell ANnotation using Variational Inference) model [21], a semi-supervised extension of scVI, was employed as the baseline for Level-2 and Level-3 methods. scANVI incorporates pre-existing cell-type annotations into the generative model to guide the latent space, thereby enhancing the preservation of biological signals during integration. It was selected to evaluate methods that leverage cell-type labels for biological conservation.

Multi-Level loss designs for single-cell data integration

To systematically dissect the impact of information regularization on integration performance, we designed a three-level evaluation framework. Level-1 methods focus on batch effect removal using batch labels; Level-2 methods incorporate cell-type labels for biological conservation; and Level-3 methods utilize both for joint optimization. A total of 16 integration methods were evaluated across these levels.

Level-1: batch effect removal

Methods at this level aim to eliminate batch effects by minimizing the dependence between the latent embeddings and batch labels. The evaluated loss functions include:

GAN Generative Adversarial Network (GAN) [30] is an adversarial framework that involves a generator and a discriminator engaged in a min-max optimization. The generator creates synthetic samples, while the discriminator distinguishes between real and synthetic ones, with the goal of improving the generator's ability to produce realistic samples. We adopt the GAN loss design from scGAN [24], tailored for batch effect removal

in scRNA-seq data. This loss uses a generator to produce cell embeddings and a discriminator to predict batch labels, generating batch-independent embeddings through adversarial training. The optimization process is described by Eq. (1):

$$\min_{\varphi} \max_{\omega} \mathbb{E}_{q_{\varphi}(z_{ij}|x_{ij})} [\log p(b_j|z_{ij})] \quad (1)$$

where x_{ij} represents the scRNA-seq profile of cell i from subject j , b_j is the corresponding batch label, z_{ij} is the latent embedding generated by the encoder parameterized by φ , and the discriminator, parameterized by ω , predicts b_j from z_{ij} .

HSIC Hilbert-Schmidt Independence Criterion (HSIC) [25] is a non-parametric statistical test that measures the dependence between two random variables using kernel methods. HSIC computes the Hilbert-Schmidt norm of the cross-covariance operator in reproducing kernel Hilbert spaces (RKHS) to quantify the dependence between variables. We minimize the HSIC loss to ensure that cell representations are independent of batch information. The HSIC measure is defined by Eq. (2):

$$\text{HSIC}_n(P) = \frac{1}{n^2} \sum_{i,j}^n k(u_i, u_j) l(v_i, v_j) + \frac{1}{n^4} \sum_{i,j,k,l}^n k(u_i, u_j) l(v_k, v_l) - \frac{2}{n^3} \sum_{i,j,k}^n k(u_i, u_j) l(v_i, v_k) \quad (2)$$

where P represents the joint probability distribution of the random variables, the random variable u corresponds to the cell embeddings, and v corresponds to the batch labels, (u_i, v_i) are independent and identically distributed (iid) copies of the random variables u and v , k and l are kernel functions.

Orthog Orthogonal Projection Loss (Orthog) is a statistical approach designed to reduce the correlation between two sets of embeddings by applying orthogonal constraints. This loss is minimized to enforce orthogonality between cell embeddings and batch embeddings, effectively disentangling biological signals from technical variations. The loss is formulated as the sum of the squared elements of the covariance matrix:

$$\mathcal{L}_{\text{Orthog}} = \| \text{Cov}(E_1, E_2) \|_F^2 \quad (3)$$

where E_1 represents the matrix of cell embeddings and E_2 represents the one-hot encoded matrix of batch embeddings. $\text{Cov}(E_1, E_2)$ denotes the covariance matrix between them.

MIM Mutual Information Minimization (MIM) is an information-theoretic method that reduces the mutual information (MI) between two variables. MI quantifies the amount of information obtained about one variable through another. To minimize the MI between cell representations x and batch information y , we use the definition of MI in Eq. (4):

$$I(x; y) = \mathbb{E}_{p(x,y)} \left[\log \frac{p(x, y)}{p(x) p(y)} \right] \quad (4)$$

We used the Contrastive Log-ratio Upper Bound (CLUB) [48] estimator to approximate the upper bound of MI by treating it as a divergence between joint and product distributions. To minimize the MI between cell representations and batch information, we apply the sampled vCLUB (vCLUB-S) estimator, which employs a negative sampling strategy to reduce computational complexity. It samples a single negative pair $(x_i, y_{k'_i})$ for each

positive pair (x_i, y_i) , where k'_i is uniformly chosen from the set $\{1, 2, \dots, N\}$, excluding i . The MI is then estimated in Eq. (5):

$$\hat{I}_{\text{vCLUB-S}} = \frac{1}{N} \sum_{i=1}^N [\log_{q_\theta}(y_i | x_i) - \log_{q_\theta}(y_{k'_i} | x_i)] \quad (5)$$

where N is the number of samples, and θ represents the parameters of the variational approximation.

RBP Reverse Backpropagation (RBP) [31] is a domain adaptation method that leverages a Gradient Reversal Layer (GRL) to learn domain-invariant representations. During forward propagation, the GRL functions as an identity transformation. During backpropagation, it reverses the gradient by multiplying it with $-\lambda$, where λ is a fixed meta-parameter:

$$R_\lambda(x) = x, \quad \frac{dR_\lambda}{dx} = -\lambda \mathbf{I} \quad (6)$$

where $\mathbf{I}$ is the identity matrix. This loss helps eliminate batch-specific signals from the learned representations.

RCE Reverse Cross-Entropy (RCE) [32] encourages a uniform probability distribution across incorrect classes, introducing ambiguity to enhance the model's robustness against label noise. We use this loss to distribute labels evenly across batches, reducing batch-specific variations. The RCE loss is defined:

$$\mathcal{L}_{CE}^R(x, y) = -R_y^\top \log F(x) \quad (7)$$

where x is the input feature, y is the target label. $F(x)$ represents the model output as a probability value or confidence score. R_y is the reverse label vector, where the y -th element is zero, and all others are $\frac{1}{L-1}$, with L as the number of labels.

Level-2: biological information preservation

These methods leverage known cell-type labels to ensure the biological alignment of embeddings across batches. The benchmarked loss functions are:

CellSupcon Cell Supervised contrastive learning (CellSupcon) applies the principles of supervised contrastive learning [33], leveraging label information to optimize contrastive learning. In this work, cell-type labels are used as class labels, where samples from the same cell type are treated as positives, and those from different cell types are treated as negatives. The CellSupcon loss function incorporates multiple positives and negatives for each anchor sample, eliminating the need for hard negative mining. The loss is defined as:

$$\mathcal{L}_{\text{out}}^{\text{sup}} = \sum_{i \in I} \frac{-1}{|P(i)|} \sum_{p \in P(i)} \log \frac{\exp(z_i \cdot z_p / \tau)}{\sum_{a \in A(i)} \exp(z_i \cdot z_a / \tau)} \quad (8)$$

where $P(i)$ represents the set of indices for positive samples, and $A(i)$ comprises all other indices except i . z_i is the feature vector of the anchor sample, z_p is the feature vector of a positive sample from the same class, and z_a is the feature vector of any other sample in the mini-batch. The temperature parameter τ is used for normalization.

IRM Invariant Risk Minimization (IRM) [34] is a method that enhances model generalization by learning features that remain invariant across different environments. It defines a data representation $\Phi : \mathcal{X} \rightarrow \mathcal{H}$ that enables an optimal classifier $w : \mathcal{H} \rightarrow \mathcal{Y}$ to perform consistently across all environments. The goal is to minimize the risk $R^e(\bar{w} \circ \Phi)$ for each environment e to ensure stable correlations with the target variable, regardless of environmental variations. IRM is formulated as:

$$\min_{\Phi : \mathcal{X} \rightarrow \mathcal{Y}} \sum_{e \in \mathcal{E}_{tr}} R^e(\Phi) + \lambda \cdot \|\nabla_{w|w=1.0} R^e(w \cdot \Phi)\|^2 \quad (9)$$

where $\mathcal{E}_{tr}$ is the set of training environments, Φ is the entire invariant predictor, $w = 1.0$ is a fixed “dummy” classifier, and the gradient norm penalty evaluates the classifier’s optimality in each environment. The regularization parameter $\lambda \in [0, \infty)$ balances predictive power (an empirical risk minimization term) and the invariance of the predictor $1 \cdot \Phi(x)$. We apply IRM by treating batch information as environmental variables and learning invariant features across batches to improve model generalization.

Domain meta-learning Domain meta-learning [35] is a method designed to improve model adaptability and generalization of models across different domains. It consists of two phases: meta-train and meta-test. In the meta-train phase, the model is trained on labeled data from source domains to learn lower-dimensional representations that effectively predict class labels. This is achieved using a feature extractor defined by $\bar{y} = \sigma(T_\varphi(F_\theta(x_i)))$, where $F_\theta(x)$ is the feature extractor for $x_i \in D_{tr_i}$, T_φ is a task-specific module, and σ is the softmax activation. The task-specific loss is defined in Eq. (10):

$$l_{task}(y, \bar{y}; \theta, \varphi) = - \sum_c 1[y = c] \log \bar{y}_c \quad (10)$$

The meta-train phase loss and parameter updates are defined in Eqs. (11 and 12):

$$l_{meta-train} = l_{task}(y, \bar{y}; \theta, \varphi) \quad (11)$$

$$(\theta', \varphi') = (\theta, \varphi) - \alpha_0 \nabla [l_{meta-train}] \quad (12)$$

In the meta-test phase, the model learns domain-invariant representations to generalize across unseen environments. This involves aligning the geometric configuration of class centroids between domains. The centroid for class c in the domain d is defined as in Eq. (13):

$$\mathcal{C}_{l,d}^c = \frac{1}{N_c} \sum_{(x_i, y_i=c)} F_\theta(x_i) \quad (13)$$

The alignment loss that measures the difference in pairwise distances between class centroids across meta-train and meta-test domains is defined in Eq. (14):

$$l_{alignment}(\mathcal{C}; \theta', \varphi') = \sum_{c \in \mathcal{C}} \sum_{[d_{tr}, d_{ts}] \in D_{tr} \times D_{ts}} d(\mathcal{C}_{d_{tr}}^c, \mathcal{C}_{d_{ts}}^c) \quad (14)$$

where $d_{tr} \in D_{tr}$ are the meta-train domains and $d_{ts} \in D_{ts}$ are the meta-test domains. The loss and the parameter update in the meta-test phase are defined in Eqs. (15 and 16):

$$l_{meta-test} = l_{task}(d_{ts}; \theta', \varphi') + \beta l_{alignment} \quad (15)$$

$$(\theta, \varphi) = (\theta, \varphi) - \alpha_1 \nabla [l_{meta-train} + l_{meta-test}] \quad (16)$$

where l_{task} is derived from both labeled and unlabeled samples in d_{ts} , and β is the regularization coefficient. We use domain meta learning in batch correction to improve the model's ability to generalize across batches, ensuring consistent features are learned across different batches.

Level-3: joint optimization

Methods at this level jointly optimize for batch-effect removal and biological conservation by integrating both batch and cell-type information. This was achieved by combining loss functions from the previous levels or by using a dedicated loss function:

Domain Class Triplet loss Domain class triplet loss [36] is an extension of traditional triplet loss designed to integrate domain information for improved generalization across domains. This method modifies the triplet configuration by selecting the positive sample from the same class but a different domain, and the negative sample from the same domain but a different class. The loss function is defined in Eq. (17):

$$Loss_{DCT} = [D(a_{i_c, i_d}, p_{i_c, j_d}) - D(a_{i_c, i_d}, n_{j_c, i_d}) + \tau]_+ \quad (17)$$

where D denotes the Euclidean distance, τ is a margin that specifies the separation between the positive and negative pairs. α, p, n represent the anchor, positive, and negative samples respectively, with c indicating class labels and d indicating domain labels. We apply this loss to integrate cell type and batch information, optimizing biological signal retention while eliminating batch effects.

scIB-E: an extended framework for evaluating intra-cell-type conservation

For a more holistic evaluation, we developed scIB-E, an extension of the established scIB framework. The scIB-E framework improves upon the original by reorganizing established metrics and, critically, by introducing a novel third category for assessing intra-cell-type biological conservation. The complete framework organizes metrics into the following three distinct categories.

Batch correction metrics

This category is composed of established metrics from the scIB framework that directly quantify the effective removal of batch effects and the mixing of cells from different batches. It includes the following metrics:

Average Silhouette Width (ASW) Batch Average Silhouette Width (ASW) Batch score quantifies the effectiveness of batch correction by measuring the extent of batch mixing within each cell type. Let batch labels be $y_{batch}(i) \in \mathcal{B}$. For each cell type $c \in \mathcal{C}$, a silhouette score $s_{batch}(i)$ is computed for each cell i with respect to its batch label, as defined in Eq. (18):

$$s_{batch}(i) = \frac{b_{batch}(i) - a_{batch}(i)}{\max(a_{batch}(i), b_{batch}(i))} \in [-1, 1] \quad (18)$$

where $a_{\text{batch}}(i)$ is the mean distance from cell i to all other cells in the same batch (within cell type c), and $b_{\text{batch}}(i)$ is the mean distance from cell i to all cells in the nearest neighboring batch (within cell type c).

Following common practice, symmetrize and invert so that higher means better mixing:

$$\text{ASW}_{\text{batch}}(c) = \frac{1}{|C_c|} \sum_{i \in C_c} (1 - |s_{\text{batch}}(i)|) \quad (19)$$

The overall batch ASW score is the unweighted mean across cell types,

$$\text{ASW}_{\text{batch}} = \frac{1}{|C|} \sum_{c \in C} \text{ASW}_{\text{batch}}(c) \quad (20)$$

where larger values indicate more effective batch integration within each cell type.

Graph Connectivity (GC) Graph Connectivity (GC) is a batch correction metric that measures the connectivity within label-specific subgraphs in a k -nearest neighbors (kNN) graph. The GC score ranges from 0 to 1, with higher values indicating better batch correction and well-connected cells of the same label, while lower scores suggest poor correction and fragmented subgraphs. The GC score is calculated in Eq. (21):

$$\text{GC} = \frac{1}{|C|} \sum_{c \in C} \frac{|\text{LCC}(G(N_c; E_c))|}{|N_c|} \quad (21)$$

where C denotes the set of cell identity labels, $|\text{LCC}(G(N_c; E_c))|$ is the number of nodes in the largest connected component for each label c , and $|N_c|$ represents the total number of nodes with that label.

Local Inverse Simpson's Index (iLISI) Local Inverse Simpson's Index (iLISI) metric quantifies the degree of batch mixing by measuring the diversity of batch labels within local neighborhoods. Construct a k -nearest-neighbor graph on X . For cell i , let $p_{\mathcal{A}}(i)$ be the proportion of neighbors with batch label $\mathcal{A} \in \mathcal{B}$ (excluding i). Define the local diversity as in Eq. (22):

$$\text{LISI}_{\mathcal{B}}(i) = \left(\sum_{\mathcal{A} \in \mathcal{B}} p_{\mathcal{A}}(i)^2 \right)^{-1} \in [1, |\mathcal{B}|] \quad (22)$$

High $\text{LISI}_{\mathcal{B}}$ indicates good batch mixing. Rescale per cell to $\frac{\text{LISI}_{\mathcal{B}}(i) - 1}{B - 1}$ with $B = |\mathcal{B}|$, and aggregate by the median as in Eq. (23):

$$\text{iLISI} = \text{median}_i \left(\frac{\text{LISI}_{\mathcal{B}}(i) - 1}{B - 1} \right) \quad (23)$$

k-Nearest Neighbor Batch Effect Test (kBET) kBET (k-nearest neighbor Batch Effect Test) [49] evaluates batch effect correction by comparing local and global batch compositions within a cell's k -nearest neighbors. The kBET score ranges from 0 to 1, where a higher score means the local batch composition closely matches the global batch composition, indicating better batch effect correction. The kBET score is calculated in Eq. (24):

$$\text{kBET} = \frac{\sum_{i=1}^N \mathbb{I}(p_i \geq \alpha)}{N} \quad (24)$$

where N is the number of neighborhoods, $\mathbb{I}(p_i \geq \alpha)$ is the indicator function for the i -th neighborhood subset, equal to 1 if the p-value p_i is greater than or equal to the significance level α and 0 otherwise.

Inter-cell-type Bio-conservation metrics

This category also utilizes established label-based metrics from the scIB framework to assess the preservation of biological identity based on discrete cell-type labels. It includes:

Normalized Mutual Information (NMI) Normalized Mutual Information (NMI) is a statistical metric used to evaluate the similarity between two clustering results. It quantifies the extent of shared information between the two clusterings by normalizing the mutual information (MI) against the geometric mean of their entropies. The NMI score ranges from 0 to 1, where a higher value indicates better agreement between the clusterings, reflecting more accurate correspondence in assignments. The NMI is calculated in Eq. (25):

$$\text{NMI} = \frac{MI(P, T)}{\sqrt{H(P) H(T)}} \quad (25)$$

where P and T are the two sets of clustering results, $MI(P, T)$ is the mutual information between them, and $H(P)$ and $H(T)$ represent the entropies, measuring the randomness within each set of clusters.

Adjusted Rand Index (ARI) Adjusted Rand Index (ARI) is a metric that measures the similarity between two clustering results. It considers both correct and incorrect assignments, adjusting for chance agreement. The ARI score ranges from 0 to 1, with higher values indicating greater agreement with the ground truth. ARI is calculated in Eq. (26):

$$\text{ARI} = \frac{TP + TN - E}{TP + TN + FP + FN - E} \quad (26)$$

where TP and TN are the true positive and true negative pairs, FP and FN are the false positive and false negative pairs, and E is the expected number of random agreements calculated in Eq. (27):

$$E = \frac{(TP + FP) \cdot (TP + FN) + (FN + TN) \cdot (FP + TN)}{TP + TN + FP + FN} \quad (27)$$

Average Silhouette Width (ASW) Cell Type Average Silhouette Width (ASW) Cell Type score assesses the preservation of biological structure post-integration by quantifying the separation between distinct cell types in the embedding space. Let the integrated embedding be $X = \{x_i\}_{i=1}^n \subset \mathbb{R}^d$ with a distance metric $d(\cdot, \cdot)$. Denote cell-type labels by $y_{\text{cell}}(i) \in \mathcal{C}$. Define the silhouette for cell i under the cell-type partition as in Eq. (28):

$$s_{\text{cell}}(i) = \frac{b_{\text{cell}}(i) - a_{\text{cell}}(i)}{\max(a_{\text{cell}}(i), b_{\text{cell}}(i))} \in [-1, 1] \quad (28)$$

Where $a_{\text{cell}}(i)$ is the mean distance from x_i to cells sharing $y_{\text{cell}}(i)$, and $b_{\text{cell}}(i)$ is the minimum, over all $c' \neq y_{\text{cell}}(i)$, of the mean distance from x_i to cells with c' .

Let $\overline{s_{\text{cell}}} = \frac{1}{n} \sum_{i=1}^n s_{\text{cell}}(i)$. We report the cell-type ASW scaled to $[0, 1]$:

$$\text{ASW}_{\text{cell}} = \frac{\overline{s_{\text{cell}}} + 1}{2} \quad (29)$$

where larger values indicate better preservation of cell-type structure after integration.

Local Inverse Simpson's Index (cLISI) Local Inverse Simpson's Index (cLISI) metric assesses the preservation of biological structure by measuring the purity of cell-type labels within local neighborhoods. Using the same graph as iLISI, let $p_{\mathcal{A}}(i)$ be the proportion of neighbors with batch label $\mathcal{A} \in \mathcal{C}$. Define

$$\text{LISI}_{\mathcal{C}}(i) = \left(\sum_{\mathcal{A} \in \mathcal{C}} p_{\mathcal{A}}(i)^2 \right)^{-1} \in [1, |\mathcal{C}|] \quad (30)$$

Here, lower diversity signifies well-separated cell types, so we invert and rescale so that higher is better separation as in Eq. (31):

$$\text{cLISI} = \text{median}_i \left(\frac{L - \text{LISI}_{\mathcal{C}}(i)}{L - 1} \right) \quad (31)$$

Isolated label score Isolated label score is a metric designed to evaluate the effectiveness of data integration in managing cell identity labels that are present in only a few batches. It uses the average silhouette width (ASW) to assess the separation of isolated labels. The score ranges from 0 to 1, with higher values indicating better separation. The isolated label score is defined in Eq. (32):

$$\text{Isolated label score} = \frac{1}{N} \sum_{i=1}^N \left(\frac{1}{n_i} \sum_{j=1}^{n_i} s(j) \right) \quad (32)$$

where N is the number of isolated labels, n_i is the number of samples in the i -th isolated label, and $s(j)$ is the silhouette score for each sample j .

Intra-cell-type Bio-conservation metrics

This third category is the primary extension introduced in the scIB-E framework, designed to address a key limitation of previous benchmarks by evaluating the preservation of biological variation that is not dependent on cell-type labels. Specifically, we introduced or repurposed the following metrics for this category:

PCR comparison Principal component regression (PCR) comparison [49] measures the difference in explained variance before and after data integration. The total variance explained by the variable is calculated by summing the variance contributions from its influence across all principal components defined in Eq. (33):

$$\text{Var}(C|V) = \sum_{i=1}^G \text{Var}(C|\text{PC}_i) \times R^2(\text{PC}_i|V) \quad (33)$$

where G is the total number of principal components, C denotes the centered data matrix of the dataset under evaluation at a given stage, and V denotes the explanatory variable whose effect is being quantified. $\text{Var}(C|V)$ denotes the total variance in C explained by variable V , $\text{Var}(C|PC_i)$ denotes the variance of C captured by the i -th principal component; and $R^2(PC_i|V)$ is the squared correlation coefficient indicating how much of the variance of the i -th component is explained by the variable V .

The PCR comparison method provides two metrics: PCR comparison-batch and PCR comparison-cell. Both metrics are scaled to a range of 0 to 1, with higher values indicating better performance. In our study, PCR comparison-cell was used to validate the logic of our framework, while PCR comparison-batch was the sole PCR-based metric ultimately incorporated into the scIB-E framework for final performance scoring.

PCR comparison-batch is the PCR-based metric selected for inclusion in the final scIB-E framework. While part of the original scIB benchmark, we have repositioned it to the intra-cell-type conservation category based on its role in our analysis. It serves as a critical proxy to assess over-correction; based on our findings, strong regularization can entangle biological signals with batch effects, and an excessively low score on this metric indicates a corresponding loss of subtle biological signals. The PCR comparison-batch is calculated in Eq. (34):

$$\text{PCR comparison - batch} = \frac{\text{Var}(C_{pre}|B) - \text{Var}(C_{post}|B)}{\text{Var}(C_{pre}|B)} \quad (34)$$

where B denotes the batch variable, C_{pre} and C_{post} refer to the original (pre-correction) and integrated (post-correction) data. $\text{Var}(C_{pre}|B)$ represents the variance in the principal components explained by the batch variable before correction, and $\text{Var}(C_{post}|B)$ represents the variance in the principal components after correction.

PCR comparison-cell directly quantifies the proportion of cellular variance preserved after integration. Although not incorporated into the final aggregated scIB-E score, it was instrumental in our analysis (e.g., Fig. 3F) for validating that a decrease in the PCR comparison-batch score corresponds to a tangible loss of biological information. The PCR comparison-cell information retention is calculated in Eq. (35):

$$\text{PCR comparison - cell} = \frac{\text{Var}(C_{pre}|PC_{post}) - \text{Var}(C_{post}|B)}{\text{Var}(C_{pre}|PC_{post})} \quad (35)$$

where PC_{post} denotes the set of principal components of the post-correction data. $\text{Var}(C_{pre}|PC_{post})$ represents the variance in the principal components before correction, explained by the principal components after batch correction.

Jaccard index Jaccard index is a global metric used to evaluate the preservation of cell information during batch correction. It measures the similarity between neighborhood structures by calculating the overlap of k -nearest neighbor graphs between the cell embeddings before and after batch correction. Jaccard index ranges from 0 to 1, with higher values indicating better preservation of neighborhood structures through the batch correction process. The Jaccard index is calculated in Eq. (36):

$$\text{Jaccard Index} = \frac{1}{N} \sum_{i=1}^N \frac{|E_{PCA_i} \cap E_{\text{embedding}_i}|}{|E_{PCA_i} \cup E_{\text{embedding}_i}|} \quad (36)$$

where N is the number of batches, E_{PCA_i} and $E_{\text{embedding}_i}$ represent the sets of edges connecting the k -nearest neighbors in cell embeddings before and after batch correction, respectively.

Corr-MSE: A correlation-based loss for preserving intra-batch structure

To preserve continuous intra-cell-type biological variation independent of discrete cell-type labels, we introduce the Correlation Mean Squared Error (Corr-MSE) loss. This loss function is designed to maintain the global cell correlation structure within each batch during integration.

For each batch, the Pearson correlation coefficient matrices are computed for both the original PCA embeddings and the batch-corrected embeddings. The loss is computed as the mean squared error (MSE) between these two correlation matrices as described in Eq. (37):

$$\mathcal{L}_{\text{Corr-MSE}} = \sum_{b=1}^B \text{MSE} \left(\text{Corr} \left(\mathbf{Z}_{\text{PCA}}^{(b)} \right), \text{Corr} \left(\mathbf{Z}_{\text{corrected}}^{(b)} \right) \right) \quad (37)$$

where B is the number of batches, $\mathbf{Z}_{\text{PCA}}^{(b)}$ represents the original PCA embeddings for batch b , and $\mathbf{Z}_{\text{corrected}}^{(b)}$ refers to the batch-corrected embeddings for batch b .

Benchmarking implementation and evaluation

Benchmark datasets and preprocessing The performance of the benchmarked integration methods was assessed across several diverse, publicly available single-cell RNA-sequencing datasets to ensure the robustness and generalizability of our findings. A consistent preprocessing step was applied to all datasets, where the feature space for analysis was limited to the 4,000 most highly variable genes.

Immune The immune dataset consists of human immune cells collected from bone marrow and peripheral blood mononuclear cells (PBMCs), sequenced using multiple platforms. It includes 33,506 cells from 10 donors and 16 cell types, collected from five studies. The immune dataset was derived from the study [27].

Pancreas The pancreas dataset consists of 16,382 cells collected from 9 batches, annotated and classified into 14 distinct cell types. It was obtained from the study [38].

BMMC The Bone Marrow Mononuclear Cells (BMMC) dataset was created for the 2021 NeurIPS Multimodal Single-Cell Data Integration competition [39]. It contains 90,261 cells and 13,953 genes, originating from bone marrow mononuclear cells collected from multiple donors and sequenced across 12 batches [50]. The dataset includes 45 cell types and is generated using various technologies, such as RNA and protein measurements (CITE-seq) and RNA with chromatin accessibility (10x Multiome). For this analysis, the dataset focuses exclusively on transcriptomic data.

HLCA The Human Fetal Lung Cell Atlas (HLCA) is an integrated reference atlas of the human respiratory system, including lung parenchyma, respiratory airways, and the nose. We used the core dataset from the HLCA, comprising 584,944 lung cells derived from 166 samples across 107 individuals. The HLCA provides detailed annotations at multiple

cell-type levels, along with comprehensive metadata. The dataset was sourced from the study [28].

Human fetal lung cell atlas The Human fetal lung cell atlas is a dataset based on a multiomic analysis of human fetal lungs, covering 5 to 22 post-conception weeks. This dataset includes 29 batches from 12 donors and spans 8 developmental stages, encompassing 14 broad cell types and 144 newly classified cell types. The dataset was sourced from the study [29].

HBCA The Human Breast Cell Atlas (HBCA) is a comprehensive single-cell RNA-sequencing dataset of the adult human breast. It comprises over 800,000 cells collected from 55 donors who had undergone reduction mammoplasties or risk-reduction mastectomies, with samples processed across 45 distinct batches. The dataset profiles the three main cellular compartments: epithelial, immune, and stromal, and provides detailed annotations at multiple resolutions, including 11 broad cell types and 41 distinct cell sub-clusters. The atlas was created to study how cellular composition shifts in response to breast cancer risk factors, including age, parity, and germline BRCA1/2 mutations. The dataset was sourced from the study [41].

Model implementation and hyperparameter optimization

All models were implemented using PyTorch (v1.13.0+cu117) and trained on a single NVIDIA A100 GPU. Baseline hyperparameters for the scVI and scANVI models were retained from their original implementations (Additional file 2: Table S1). For training, the models were trained for a maximum of 400 epochs, with a batch size of 128. Early stopping was enabled to prevent overfitting.

The various loss functions were integrated into the standard variational autoencoder (VAE) models. The final objective function for each method is a weighted aggregation of the VAE loss (L_{VAE}) and the additional regularization loss terms (L_{reg}). The general formulation is as follows:

$$L_{total} = L_{VAE} + \sum_{i=1}^N \lambda_i L_{reg,i} \quad (38)$$

where N is the number of regularization terms and λ_i is the corresponding weight that balances the contribution to the total loss. To systematically determine the optimal weights for each loss function combination, we employed the Ray Tune [37] framework for hyperparameter optimization. The optimization process was conducted on the immune dataset, which was selected as our primary benchmark dataset, with the objective of maximizing the overall scIB total score.

Our parameter selection procedure began by defining a logarithmic search space for each hyperparameter, performing a grid search over the discrete set of values: {0.001, 0.01, 0.1, 1, 10, 100}. For methods in Level-1 and Level-2 involving a single regularization term, we identified the weight that yielded the highest scIB total score. For the CellSupcon loss, while higher weights generally increased the scIB score, our analysis in Fig. 3 revealed that a weight of 10 provided an optimal balance, effectively enhancing biological conservation without significant overcorrection, and was therefore selected for subsequent analyses. For Level-3 methods that combine two regularization loss terms, we simplified the search by fixing the CellSupcon weight at 10 and performing a grid search

for the weight of the second loss term. For the RCE-CE method, a two-dimensional grid search was executed. The hyperparameters determined through this optimization process on the immune dataset were then fixed and applied uniformly across all other datasets for the final benchmarking evaluation to ensure a standardized and reproducible comparison. A detailed list of the final hyperparameters used for each of the 16 methods is provided in Additional file 2: Table S2.

Performance score aggregation

The scIB-E framework quantifies overall single-cell integration performance using the scIB-E Total score. This score is a weighted average of the three primary performance categories: batch correction (S_{batch}), inter-cell-type biological conservation (S_{inter}), and intra-cell-type biological conservation (S_{intra}). The aggregation is defined by the formula:

$$scIB - E_{Total} = w_{batch} \cdot S_{batch} + w_{inter} \cdot S_{inter} + w_{intra} \cdot S_{intra} \quad (39)$$

Each category score (S) is the unweighted arithmetic mean of its constituent metrics (Fig. 4C). The empirically determined weights are set to $w_{batch}=0.2$, $w_{inter}=0.4$, and $w_{intra}=0.4$.

These weights were derived from a data-driven optimization procedure designed to ensure that the scIB-E Total score serves as a robust proxy for the preservation of complex biological signals. This procedure utilized the multi-layer annotated Human Lung Cell Atlas (HLCA) dataset. Our objective was to identify a weighting scheme where the scIB-E Total score, calculated using broad level-1 cell annotations for training, maximally correlates with a quantitative benchmark for biological preservation. To establish this benchmark, we first calculated biological conservation scores for each of the finer-grained annotation levels (2, 3, 4, and 'finest'). The arithmetic mean of these scores served as our final reference, representing a measure of deeper biological truth not exposed to the model.

The optimization involved a grid search over all valid weight combinations, constrained by $w_{batch} + w_{inter} + w_{intra} = 1$ and $0.2 \leq w_{batch} \leq 0.5$. For each candidate weight combination, we computed the Pearson correlation between the resulting scIB-E Total scores and our quantitative benchmark for biological preservation, assessed across all integration results. The analysis revealed that the weight combination of 0.2 for batch correction, 0.4 for inter-cell-type conservation, and 0.4 for intra-cell-type conservation maximized this correlation. This empirical weighting ensures that the scIB-E Total score is a reliable indicator of an integration method's ability to balance effective batch removal with the preservation of multi-resolution biological variation.

Downstream task evaluation

Moran's I Moran's I is a metric that evaluates the preservation of developmental trajectories in batch-corrected embeddings. It quantifies the global autocorrelation by measuring the correlation between developmental information and graph structures derived

from batch-corrected embeddings. Moran's I ranges from -1 to 1 , with higher values indicating better trajectory preservation. The formula for Moran's I is defined in Eq. (40):

$$\text{Moran's } I = \frac{N}{W} \frac{\sum_{i=1}^N \sum_{j=1}^N w_{ij} (x_i - \bar{x})(x_j - \bar{x})}{\sum_{i=1}^N (x_i - \bar{x})^2} \quad (40)$$

where N denotes the number of cells. x_i is the developmental information of cell i . $\bar{x}$ is the average developmental information. w_{ij} defines neighborhood connectivity, which is 1 if cells are neighbors and 0 otherwise. W is the sum of all connections. For analysis, it is normalized to $[0, 1]$, where values closer to 1 indicate stronger preservation.

Geary's C Geary's C evaluates local differences in developmental information among neighboring cells in batch-corrected embeddings. Its values range from 0 to 2, with lower values indicating better local similarity. The formula for Geary's C is defined in Eq. (41):

$$\text{Geary's } C = \frac{(N-1)}{2W} \frac{\sum_{i=1}^N \sum_{j=1}^N w_{ij} (x_i - x_j)^2}{\sum_{i=1}^N (x_i - \bar{x})^2} \quad (41)$$

where w_{ij} is the weight between cells i and j , $W = \sum_{i,j} w_{ij}$ is the sum of all weights. For analysis, Geary's C is normalized to $[0, 1]$, where higher values indicate better preservation of the developmental trajectory.

Differential Abundance (DA) Analysis We benchmarked the performance of the DCT-Corr method against scVI and scANVI on a differential abundance (DA) analysis task using the Human Breast Cell Atlas (HBCA) [41]. To simulate a realistic workflow, models were trained on 11 broad cell-type labels, while subsequent DA testing was conducted on 41 granular subclusters. We replicated the analytical strategy from the source publication by adopting the Milo [51] methodology. Tests were performed independently within each of the three major cellular compartments (epithelial, stromal, and immune). For each compartment, cell neighborhoods were defined on a k -nearest neighbor (kNN) graph of the integrated embeddings ($k = 50$, $d = 20$, $\text{prop} = 0.3$). We then performed four key comparisons using generalized linear models (GLMs) with blocking covariates: (1) the effect of age within the average-risk (AR) cohort ($n = 22$); (2) the effect of parity within the AR cohort; (3) high-risk BRCA1 mutation carriers (HR-BR1, $n = 11$) versus the AR cohort; and (4) high-risk BRCA2 mutation carriers (HR-BR2, $n = 11$) versus the AR cohort. To validate biological discovery, the magnitude of abundance shifts for each subcluster was quantified via the average absolute log-fold change and benchmarked against the findings reported in the source publication [41].

Supplementary Information

The online version contains supplementary material available at <https://doi.org/10.1186/s13059-025-03869-z>.

Additional file 1. SupplementaryFigs S1-S11 FigS1: Evaluation of multi-level loss functions for single-cell integration on immune dataset. FigS2: Evaluation of multi-level loss functions for single-cell integration on pancreas dataset. FigS3: Evaluation of multi-level loss functions for single-cell integration on BMMC dataset. FigS4: Extended scIB metrics for intra-cell-type biological conservation evaluation. FigS5: Assessing multi-level loss functions for single-cell integration using scIB-E metrics on immune dataset. FigS6: Assessing multi-level loss functions for single-cell integration using scIB-E metrics on pancreas dataset. FigS7: Assessing multi-level loss functions for single-cell integration using scIB-E metrics on BMMC dataset. FigS8: Assessing single-cell integration methods for comprehensive biological conservation. Fig S9: Performance evaluation and differential abundance analysis on the HBCA dataset for BRCA

carriers. Fig S10: Visualization of the HBCA dataset integration and differential abundance analysis using scVI. Fig S11: Visualization of the HBCA dataset integration and differential abundance analysis using scANVI

Additional file 2. Supplementary Tables S1–S16 Table S1: Baseline model hyperparameters and training settings for scVI and scANVI. Table S2: Hyperparameters for 16 deep learning single-cell integration methods across three levels. Table S3: Description of the datasets used for benchmarking. Table S4: Detailed scIB evaluation results across three datasets for multi-level loss functions in single-cell integration. Table S5: Detailed scIB evaluation results for the immune dataset under varying loss regularizations. Table S6: Detailed evaluation of metrics for multi-level loss functions in single-cell integration: local cell-neighbors-based batch correction metrics and PCR comparison scores. Table S7: Detailed evaluation of metrics for multi-level loss functions, including local cell-neighbors-based batch correction metrics and PCR comparison scores. Table S8: Detailed scIB-E evaluation results across three datasets for multi-level loss functions in single-cell integration. Table S9: Detailed scIB-E metrics for enhanced methods in biologically preserving single-cell integration. Table S10: Downstream analysis results for enhanced methods in biologically preserving single-cell integration. Table S11: Detailed scIB-E metrics for the Human Breast Cell Atlas (HBCA) integration. Table S12: Detailed scIB metrics for the Human Breast Cell Atlas (HBCA) integration. Table S13: Differential abundance analysis results for the age comparison in the Human Breast Cell Atlas (HBCA) using the DCT-Corr integration. Table S14: Differential abundance analysis results for the parity comparison in the Human Breast Cell Atlas (HBCA) using the DCT-Corr integration. Table S15: Differential abundance analysis results for the BRCA1 comparison in the Human Breast Cell Atlas (HBCA) using the DCT-Corr integration. Table S16: Differential abundance analysis results for the BRCA2 comparison in the Human Breast Cell Atlas (HBCA) using the DCT-Corr integration.

Additional file 3. Peer review history

Acknowledgements

We thank technical support from the Data Science Platform of Guangzhou National Laboratory and the Bio-medical Big Data Operating System (Bio-OS).

Peer review information

Barbara Cheifet and Claudia Feng were the primary editors of this article and managed its editorial process and peer review in collaboration with the rest of the editorial team. The peer-review history is available in Additional File 3.

Authors' contributions

Y.L., J.L. and W.L. conceived the project. J.L. and C.Y. designed the framework and loss designs. J.Cheng. helped data analysis. J.Chen. contributed to manuscript revision. J.L. and C.Y. wrote the manuscript with contribution from all authors. Y.L. supervised the entire project. All authors read and approved the final manuscript.

Funding

This study was supported by the National Key R&D Program (2023YFF1204701), the Startup Program of Guangzhou National Laboratory (YW-YFYJ0101), the Major Project of Guangzhou National Laboratory (GZNL2025C01013), the National Natural Science Foundation of China (No. 12371485 and No.82400622), the Guangdong Basic and Applied Basic Research Foundation (2023B1515130008), and the Department of Science and Technology of Guangdong Province (2021CX02G450).

Data availability

The immune dataset is publicly available at <https://doi.org/10.6084/m9.figshare.12420968> [52]. The pancreas dataset is available at <https://doi.org/10.6084/m9.figshare.23694912.v1> [53]. The BMBC dataset can be obtained from Gene Expression Omnibus (GEO) database using accession number GSE194122 [54]. The HLCA core dataset can be downloaded via cellxgene (<https://cellxgene.cziscience.com/collections/6f6d381a-7701-4781-935c-db10d30de293>) [55]. The Human fetal lung cell atlas can be accessed from <https://fetal-lung.cellgeni.sanger.ac.uk/scRNA.html> [56]. The HBCA dataset is available on cellxgene (<https://cellxgene.cziscience.com/collections/48259aa8-f168-4bf5-b797-af8e88da6637>) [57]. The code in this study is available on GitHub (<https://github.com/Chenxin-Yi/scIB-E>) [58] and in the Zenodo repository (<https://doi.org/10.5281/zenodo.17481059>) [59] under the GNU Affero General Public License v3.0.

Declarations

Ethics approval and consent to participate

Not applicable.

Competing interests

The authors declare no competing interests.

Author details

¹School of Intelligent Systems Engineering, Sun Yat-sen University, Shenzhen 518107, China

²Guangzhou National Laboratory, Guangzhou 510005, China

³Multimedia Laboratory (MMLab), The Chinese University of Hong Kong, Hongkong 999077, China

⁴GMU-GIBH Joint School of Life Sciences, The Guangdong-Hong Kong-Macao Joint Laboratory for Cell Fate Regulation and Diseases, Guangzhou Medical University, Guangzhou 511436, China

⁵Key Laboratory of Systems Health Science of Zhejiang Province, School of Life Science, Hangzhou Institute for Advanced Study, University of Chinese Academy of Sciences, Hangzhou 310024, China

⁶School of Life Sciences and Biotechnology, Shanghai Jiao Tong University, Shanghai 200240, China

⁷Shanghai Institute of Nutrition and Health, Chinese Academy of Sciences, Shanghai 200030, China

⁸Bioland Laboratory, Guangzhou 510005, China

Received: 30 October 2025 / Accepted: 11 November 2025

Published online: 20 November 2025

References

1. Gulati GS, D'Silva JP, Liu Y, Wang L, Newman AM. Profiling cell identity and tissue architecture with single-cell and spatial transcriptomics. *Nat Rev Mol Cell Biol.* 2024;(1):21. <https://doi.org/10.1038/s41580-024-00768-2>.
2. Regev A, Teichmann SA, Lander ES, Amit I, Benoist C, Birney E, et al. The human cell atlas. *Elife.* 2017. <https://doi.org/10.7554/eLife.27041>.
3. Heumos L, Schaar AC, Lance C, Litnetskaya A, Drost F, Zappia L, et al. Best practices for single-cell analysis across modalities. *Nat Rev Genet.* 2023. <https://doi.org/10.1038/s41576-023-00586-w>.
4. Adey AC. Integration of single-cell genomics datasets. *Cell.* 2019;177:1677–9.
5. Stuart T, Butler A, Hoffman P, Hafemeister C, Papalexi E, Mauck WM III, et al. Comprehensive integration of single-cell data. *Cell.* 2019;177:1888–1902.e1821.
6. Butler A, Hoffman P, Smibert P, Papalexi E, Satija R. Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nat Biotechnol.* 2018;36:411–20.
7. Haghverdi L, Lun ATL, Morgan MD, Marioni JC. Batch effects in single-cell RNA-sequencing data are corrected by matching mutual nearest neighbors. *Nat Biotechnol.* 2018;36:421–7.
8. Hie B, Bryson B, Berger B. Efficient integration of heterogeneous single-cell transcriptomes using scanorama. *Nat Biotechnol.* 2019;37:685–91.
9. Mereu E, Lafzi A, Moutinho C, Ziegenhain C, McCarthy DJ, Álvarez-Varela A, Batlle E, Sagar, Grün D, Lau JK, et al. Benchmarking single-cell RNA-sequencing protocols for cell atlas projects. *Nat Biotechnol.* 2020;38:747–55.
10. Korsunsky I, Millard N, Fan J, Slowikowski K, Zhang F, Wei K, et al. Fast, sensitive and accurate integration of single-cell data with Harmony. *Nat Methods.* 2019;16:1289–96.
11. Polański K, Young MD, Miao Z, Meyer KB, Teichmann SA, Park J-E. BBKNN: fast batch alignment of single cell transcriptomes. *Bioinformatics.* 2020;36:964–5.
12. Welch JD, Kozareva V, Ferreira A, Vanderburg C, Martin C, Macosko EZ. Single-cell multi-omic integration compares and contrasts features of brain cell identity. *Cell.* 2019;177:1873–1887. e1817.
13. Lin Y, Ghazanfar S, Wang KY, Gagnon-Bartsch JA, Lo KK, Su X, et al. ScMerge leverages factor analysis, stable expression, and pseudoreplication to merge multiple single-cell RNA-seq datasets. *Proc Natl Acad Sci USA.* 2019;116:9775–84.
14. Lin Y, Cao Y, Willie E, Patrick E, Yang JYH. Atlas-scale single-cell multi-sample multi-condition data integration using scMerge2. *Nat Commun.* 2023;14:4272.
15. Tran HTN, Ang KS, Chevrier M, Zhang X, Lee NYS, Goh M, Chen J. A benchmark of batch-effect correction methods for single-cell RNA sequencing data. *Genome Biol.* 2020;21:12.
16. Ma Q, Xu D. Deep learning shapes single-cell data analysis. *Nat Rev Mol Cell Biol.* 2022;23:303–4.
17. Eraslan G, Avsec Ž, Gagneur J, Theis FJ. Deep learning: new computational modelling techniques for genomics. *Nat Rev Genet.* 2019;20:389–403.
18. Li X, Wang K, Lyu Y, Pan H, Zhang J, Stambolian D, Susztak K, Reilly MP, Hu G, Li M. Deep learning enables accurate clustering with batch effect removal in single-cell RNA-seq analysis. *Nat Commun.* 2020;11:2338.
19. Lopez R, Regier J, Cole MB, Jordan MI, Yosef N. Deep generative modeling for single-cell transcriptomics. *Nat Methods.* 2018;15:1053–8.
20. Xiong L, Tian K, Li Y, Ning W, Gao X, Zhang QC. Online single-cell data integration through projecting heterogeneous datasets into a common cell-embedding space. *Nat Commun.* 2022;13:6118.
21. Xu C, Lopez R, Mehlman E, Regier J, Jordan MI, Yosef N. Probabilistic harmonization and annotation of single-cell transcriptomics data with deep generative models. *Mol Syst Biol.* 2021;17:e9620.
22. Shree A, Pavan MK, Zafar H. ScDREAMER for atlas-level integration of single-cell datasets using deep generative model paired with adversarial classifier. *Nat Commun.* 2023;14:7781.
23. Yu X, Xu X, Zhang J, Li X. Batch alignment of single-cell transcriptomics data using deep metric learning. *Nat Commun.* 2023;14:960.
24. Bahrami M, Maitra M, Nagy C, Turecki G, Rabiee HR, Li Y. Deep feature extraction of single-cell transcriptomes by generative adversarial network. *Bioinformatics.* 2021;37:1345–51.
25. Lopez R, Regier J, Jordan MI, Yosef N. Information constraints on auto-encoding variational Bayes. *Adv Neural Inf Process Syst.* 2018;31:6114–25.
26. Sun Y, Qiu P. Domain adaptation for supervised integration of scRNA-seq data. *Commun Biol.* 2023;6:274.
27. Luecken MD, Buttner M, Chaichoompu K, Danese A, Interlandi M, Mueller MF, et al. Benchmarking atlas-level data integration in single-cell genomics. *Nat Methods.* 2022;19:41–50.
28. Sikkema L, Ramirez-Suastegui C, Strobl DC, Gillett TE, Zappia L, Madisoos E, et al. An integrated cell atlas of the lung in health and disease. *Nat Med.* 2023;29:1563–77.
29. He P, Lim K, Sun D, Pett JP, Jeng Q, Polanski K, Dong Z, Bolt L, Richardson L, Mamanova L, et al. A human fetal lung cell atlas uncovers proximal-distal gradients of differentiation and key regulators of epithelial fates. *Cell.* 2022;185:4841–e48604825.
30. Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, Courville A, Bengio Y. Generative adversarial Nets. *Adv Neural Inf Process Syst.* 2014;27:2672–80.
31. Ganin Y, Lempitsky V. Unsupervised domain adaptation by backpropagation. In *International conference on machine learning*. PMLR; 2015: 1180–1189.
32. Pang T, Du C, Dong Y, Zhu J. Towards robust detection of adversarial examples. *Advances in neural information processing systems.* 2018;31:4579–89.
33. Khosla P, Teterwak P, Wang C, Sarna A, Tian Y, Isola P, Maschinot A, Liu C, Krishnan D. Supervised contrastive learning. *Adv Neural Inf Process Syst.* 2020;33:18661–73.
34. Arjovsky M, Bottou L, Gulrajani I, Lopez-Paz D. Invariant risk minimization. *ArXiv Preprint arXiv:190702893* 2019. <https://arxiv.org/abs/1907.02893>

35. Sharifi-Noghabi H, Asghari H, Mehrasa N, Ester M. Domain generalization via semi-supervised meta learning. ArXiv Preprint arXiv:2009.12658 2020. <https://arxiv.org/abs/2009.12658>
36. Guo K, Lovell BC. Domain-aware triplet loss in domain generalization. *Comput Vis Image Underst.* 2024;243:103979.
37. Liaw R, Liang E, Nishihara R, Moritz P, Gonzalez JE, Stoica I. Tune: A research platform for distributed model selection and training. ArXiv Preprint arXiv:1807.05118 2018. <https://arxiv.org/abs/1807.05118>
38. De Donno C, Hediye-Zadeh S, Moinfar AA, Wagenstetter M, Zappia L, Lotfollahi M, Theis FJ. Population-level integration of single-cell datasets enables multi-scale analysis across samples. *Nat Methods.* 2023;20:1683–92.
39. Lance C, Luecken MD, Burkhardt DB, Cannoodt R, Rautenstrauch P, Laddach A, Ubingazhibov A, Cao Z-J, Deng K, Khan S et al. Multimodal single cell data integration challenge: Results and lessons learned. In *Proceedings of the NeurIPS 2021 Competitions and Demonstrations Track* (Douwé K, Marco C, Barbara C eds.), vol. 176. pp. 162–176. *Proceedings of Machine Learning Research: PMLR*; 2022:162–176.
40. Becht E, McInnes L, Healy J, Dutertre CA, Kwok IWH, Ng LG, et al. Dimensionality reduction for visualizing single-cell data using UMAP. *Nat Biotechnol.* 2018. <https://doi.org/10.1038/nbt.4314>.
41. Reed AD, Pensa S, Steif A, Stenning J, Kunz DJ, Porter LJ, Hua K, He P, Twigger AJ, Siu AJQ, et al. A single-cell atlas enables mapping of homeostatic cellular shifts in the adult human breast. *Nat Genet.* 2024;56:652–62.
42. Xiao C, Chen Y, Meng Q, Wei L, Zhang X. Benchmarking multi-omics integration algorithms across single-cell RNA and ATAC data. *Brief Bioinform.* 2024. <https://doi.org/10.1093/bib/bbae095>.
43. Andreatta M, Hérault L, Gueguen P, Gfeller D, Berenstein AJ, Carmona SJ. Semi-supervised integration of single-cell transcriptomics data. *Nat Commun.* 2024;15:872.
44. Heimberg G, Kuo T, DePianto DJ, Salem O, Heigl T, Diamant N, Scalia G, Biancalani T, Turley SJ, Rock JR, et al. A cell atlas foundation model for scalable search of similar human cells. *Nature.* 2025;638:1085–94.
45. Dominguez Conde C, Xu C, Jarvis LB, Rainbow DB, Wells SB, Gomes T, et al. Cross-tissue immune cell analysis reveals tissue-specific features in humans. *Science.* 2022;376:eabl5197.
46. Weinberger E, Lin C, Lee SI. Isolating salient variations of interest in single-cell data with contrastivevi. *Nat Methods.* 2023;20:1336–45.
47. Zhao J, Jaffe A, Li H, Lindenbaum O, Sefik E, Jackson R, et al. Detection of differentially abundant cell subpopulations in scRNA-seq data. *Proc Natl Acad Sci U S A.* 2021. <https://doi.org/10.1073/pnas.2100293118>.
48. Cheng P, Hao W, Dai S, Liu J, Gan Z, Carin L. Club: A contrastive log-ratio upper bound of mutual information. In *International conference on machine learning*. PMLR. 2020;119:1779–88.
49. Buttner M, Miao Z, Wolf FA, Teichmann SA, Theis FJ. A test metric for assessing single-cell RNA-seq batch correction. *Nat Methods.* 2019;16:43–9.
50. Luecken MD, Burkhardt DB, Cannoodt R, Lance C, Agrawal A, Aliee H, Chen AT, Deconinck L, Detweiler AM, Granados AA. A sandbox for prediction and integration of DNA, RNA, and proteins in single cells. In *Thirty-fifth conference on neural information processing systems datasets and benchmarks track (Round 2)*. La Jolla, CA:2021.
51. Dann E, Henderson NC, Teichmann SA, Morgan MD, Marioni JC. Differential abundance testing on single-cell data using k-nearest neighbor graphs. *Nat Biotechnol.* 2022;40:245–53.
52. Luecken MD, Buttner M, Chaichoompu K, Danese A, Interlandi M, Mueller MF, Strobl DC, Zappia L, Dugas M, Theis MC-T. FJ: Benchmarking atlas-level data integration in single-cell genomics - integration task datasets. *Datasets*. Figshare. <https://doi.org/10.6084/m9.figshare.12420968> (2020).
53. De Donno C, Hediye-Zadeh S, Moinfar AA, Wagenstetter M, Zappia L, Lotfollahi M, Theis FJ. Population-level integration of single-cell datasets enables multi-scale analysis across samples. *Datasets*. Figshare. <https://doi.org/10.6084/m9.figshare.23694912.v1> (2023).
54. Burkhardt D, Luecken M, Lance C, Cannoodt R, Pisco A, Krishnaswamy S, Theis F, Bloom J. A sandbox for prediction and integration of DNA, RNA, and proteins in single cells. *Datasets*. *Gene Expression Omnibus*. (2022). Accessed on 14 Mar 2024. <https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE194122>
55. Sikkema L, Ramirez-Suastegui C, Strobl DC, Gillett TE, Zappia L, Madisson E, Markov NS, Zaragosi LE, Ji Y, Ansari M. The integrated Human Lung Cell Atlas. *Datasets*. *CZ CELLxGENE Discover*. (2023). Accessed on 14 Mar 2024. <https://cellxgene.cziscience.com/collections/6f6d381a-7701-4781-935c-db10d30de293>
56. He P, Lim K, Sun D, Pett JP, Jeng Q, Polanski K, Dong Z, Bolt L, Richardson L, Mamanova L et al. A human fetal lung cell atlas uncovers proximal-distal gradients of differentiation and key regulators of epithelial fates. *Datasets*. *Sanger Institute*. (2022). Accessed on 29 Aug 2024. <https://fetal-lung.cellgeni.sanger.ac.uk/scRNA.html>
57. Reed AD, Pensa S, Steif A, Stenning J, Kunz DJ, Porter LJ, Hua K, He P, Twigger A-J, Siu AJQ et al. Human breast cell atlas. *Datasets*. *CZ CELLxGENE Discover*. (2024). Accessed on 25 Aug 2025. <https://cellxgene.cziscience.com/collections/48259aa8-f168-4bf5-b797-af8e88da6637>
58. Yi C, Cheng J, Chen J, Liu W, Liu J, Li Y. Benchmarking deep learning methods for biologically conserved single-cell integration. *Github*. (2025). Accessed on 17 Nov 2025. <https://github.com/Chenxin-Yi/scIB-E>
59. Yi C, Cheng J, Chen J, Liu W, Liu J, Li Y. Benchmarking deep learning methods for biologically conserved single-cell integration. *Zenodo*. (2025). Accessed on 17 Nov 2025. <https://doi.org/10.5281/zenodo.17481059>

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.