

A Benchmarking Study of Batch Integration Methods in Single-Cell Genomics

Keyan Su, Xueyi Liu, Zichen Zhuang

Abstract

The proliferation of single-cell RNA sequencing (scRNA-seq) has generated numerous computational tools for integrating datasets and removing technical batch effects. However, selecting the optimal tool for a given study remains challenging due to the context-dependent performance of these methods. This project addresses this gap by conducting a systematic benchmark of four leading batch integration methods—Harmony, scVI, BBKNN, and Scanorama—across two distinct, biologically relevant datasets. Our study evaluates their efficacy in removing batch effects while conserving biological variance. The results provide nuanced, data-driven guidance for researchers on method selection based on specific dataset characteristics, such as technical complexity and tissue diversity.

1 Introduction

Single-cell RNA sequencing (scRNA-seq) has revolutionized biology by enabling the profiling of gene expression at the resolution of individual cells. A major obstacle in combining multiple scRNA-seq datasets is the batch effect—non-biological technical variations introduced by differences in experimental conditions, instruments, or donors. These effects can obscure true biological signals and lead to erroneous conclusions in downstream analyses. While many batch integration algorithms have been developed, their performance is highly variable and dependent on the specific context of the data, such as the severity of the batch effect, the biological complexity, and the sample type. There is a pressing need for clear, context-specific guidelines to help researchers choose the most appropriate method.

The input to our benchmarking framework is a collection of raw or normalized scRNA-seq count matrices from multiple batches, along with corresponding batch labels and, where available, cell-type annotations. The output is a comprehensive evaluation report detailing each method’s performance in removing batch artifacts and preserving biological structures, culminating in practical recommendations.

2 Related Work

Recent years have seen rapid development in scRNA-seq batch integration methods, which can be broadly categorized into several groups. Correction-based methods, such as ComBat, use linear models to adjust for batch effects but may not handle complex non-linearities. Nearest-neighbor graph-based methods, like BBKNN and Scanorama, focus on constructing corrected cell-cell similarity graphs across batches. Deep learning-based generative models, such as scVI, learn a non-linear, low-dimensional representation of the data that explicitly models and disentangles batch effects. Iterative clustering-based methods, exemplified by Harmony, perform iterative clustering and correction to align datasets.

The Open Problems in Single-Cell Analysis consortium has established standardized benchmarks for these tools, providing a foundational framework for comparison. However, many benchmarks rely on a fixed set of simulated or similar datasets. Our work builds upon this foundation by intentionally selecting and evaluating methods on two publicly available, real-world datasets that present distinct and realistic biological integration challenges. This approach allows us to move beyond general performance rankings to deliver context-specific insights that are directly applicable to experimental researchers.

3 Dataset

We curated two publicly available datasets from the Single Cell Portal to represent distinct and common integration challenges in biological research.

3.1 GSE132044: A Multi-Technological Benchmark Dataset

This dataset is from a systematic study comparing seven different scRNA-seq protocols. We utilized the human peripheral blood mononuclear cell (PBMC) and human/mouse mixed cell line samples.

- **Samples & Splits:** The data comprises multiple technical replicates for each protocol (e.g., 10x Chromium v2/v3, Drop-seq, Smart-seq2). We treated each protocol-library combination as a distinct batch. No predefined train/validation/test split was used; the entire dataset was processed for unsupervised integration evaluation.
- **Preprocessing:** Data was downloaded in the form of raw count matrices. Standard preprocessing was applied using the Scanpy toolkit: cells with an extreme number of genes or high mitochondrial gene percentage were filtered out. Genes expressed in a minimal number of cells were removed. Counts were normalized per cell, log1p-transformed, and the top 2,000 highly variable genes were selected for downstream analysis.
- **Key Features:** This dataset presents the challenge of integrating data generated by fundamentally different technologies and protocols, which induce strong, non-linear batch effects. It tests an algorithm’s ability to separate technical noise from true biology, especially in the mixed-species sample where the “biological signal” (species identity) must be preserved.

3.2 GSE201333 (Tabula Sapiens): A Multi-Tissue Human Cell Atlas

We selected a subset of this comprehensive atlas focusing on immune-related tissues from multiple human donors.

- **Samples & Splits:** Data from different donors and different tissues were considered separate batches. The goal was to integrate across donors within the same tissue and across different tissues within the immune system.
- **Preprocessing:** A similar preprocessing pipeline as for GSE132044 was applied. Additional care was taken to harmonize cell-type annotations across tissues using the provided ontology.
- **Key Features:** This dataset introduces the challenge of integrating across biological batches (donors) and tissue microenvironments. The effects are often more subtle and confounded with real biological variation than in technical batches. A successful method must remove donor-specific bias while preserving delicate tissue-specific and cell-type-specific expression patterns.

4 Methods

We benchmarked four representative and widely used batch integration methods, each employing a distinct algorithmic strategy.

4.1 Harmony

Harmony is an iterative clustering-based integration algorithm. It first embeds cells into a PCA space, then performs soft clustering. Within each cluster, it computes a correction factor to centroid-align cells from different batches. This process repeats until convergence. We chose Harmony for its proven effectiveness, speed, and particular strength in integrating datasets with strong batch effects while preserving fine-grained population structure.

4.2 scVI (Single-Cell Variational Inference)

scVI is a deep generative model based on variational autoencoders (VAEs). It models the observed scRNA-seq count data as generated from a low-dimensional latent random variable, which is conditioned on both batch information and latent biological state. By learning to reconstruct the data, it infers a batch-corrected latent representation. We included scVI for its principled probabilistic framework, its ability to model uncertainty, and its superior performance in complex integration tasks noted in prior benchmarks.

4.3 BBKNN (Batch Balanced k-Nearest Neighbors)

BBKNN is a graph-based method that performs a lightweight correction. It constructs a k-nearest neighbor graph separately for each batch and then identifies “mutual nearest neighbors” (MNNs) across batches. The final graph is created by connecting cells to these MNNs, effectively forcing connections across batches. We selected BBKNN for its computational efficiency, minimal distortion of the original data structure, and its design to maximize the preservation of biological variance.

4.4 Scanorama

Scanorama is also a mutual nearest neighbor (MNN)-based method, but it operates by identifying MNNs in a high-dimensional feature space and then performs a panoramic stitching of these overlapping subspaces to create a globally aligned embedding. It is designed to be scalable. We included Scanorama for its robustness, efficiency in handling large datasets, and its reputation for maintaining global data structure during integration.

All methods were implemented using their standard Python packages (harmonypy, scvi-tools, bbknn, scanorama) with default parameters unless otherwise specified, ensuring a fair and practical comparison.

5 Experiments, Results, and Discussion

5.1 Experimental Setup & Evaluation Metrics

The performance of a batch integration method is judged by its ability to balance two core, often competing, objectives: the removal of non-biological technical noise and the preservation of genuine biological signal. Our evaluation employs a suite of metrics categorized accordingly.

- **Batch Effect Removal Metrics:** These quantify the success in minimizing the influence of the batch variable. Ideal values indicate that cells are well-mixed irrespective of their technical origin. Key metrics in this category include ASW_{batch}, PCR, ILISI, NMI_{batch}, and ARI_{batch} cluster.
- **Biological Conservation Metrics:** These assess how well the true biological structure (e.g., cell-type identity) is maintained after integration. Ideal values indicate that biologically similar cells remain close while distinct populations remain separate. This category includes metrics such as graph connectivity, ASW_{label}, cLISI, and the Davies-Bouldin index.

By analyzing results across both categories, we can identify whether a method successfully achieves integration, merely obscures biological variation, or inadequately corrects for batch effects.

5.2 Results and Discussion

5.2.1 GSE132044 (Technical Batch Challenge)

Benchmarking on the GSE132044 dataset, characterized by strong technical batch effects across diverse sequencing protocols, reveals a fundamental trade-off inherent to batch integration. The overall performance landscape, summarized in Figure 1, shows that methods excelling at removing batch effects tend to compromise the preservation of global biological structure, and vice versa.

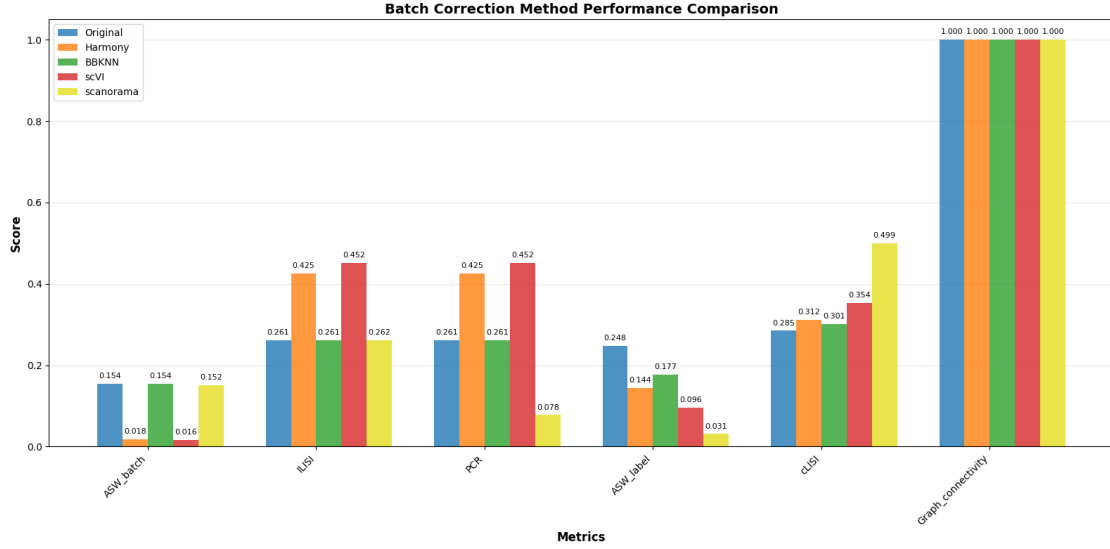


Figure 1: Comprehensive evaluation of batch integration methods on the GSE132044 dataset across six key metrics. Higher scores are better for ILISI, Graph Connectivity, ASW_label, and cLISI. Lower scores are better for ASW_batch and PCR.

The methods naturally separated into two groups based on their performance profiles. scVI and Harmony formed the first group, demonstrating the most effective batch correction with the lowest ASW_batch scores. However, this came with a noticeable reduction in ASW_label scores, indicating diminished separation between biological cell types. Conversely, BBKNN showed minimal batch correction, leaving ASW_batch nearly unchanged from the original data, but best preserved the original biological clustering (ASW_label). Scanorama presented a more complex profile, achieving an excellent PCR score but performing poorly on ASW_label, suggesting its integration approach may over-correct in this context.

A key observation is the divergent behavior of local versus global biological conservation metrics. For instance, scVI achieved a high cLISI score despite a low ASW_label, highlighting that biological information can be preserved locally in cell neighborhoods even when global cluster separation is blurred—a nuanced insight for evaluating integration success.

5.2.2 GSE201333 (Biological Batch Challenge)

We evaluated four batch integration strategies on four immune-related tissues from Tabula Sapiens (spleen, lymph node, bone marrow, thymus). For each tissue, we randomly subsampled 5,000 cells and treated donor identity as the batch covariate. We compared four embeddings: the uncorrected tissue UMAP (baseline), the scVI latent space, a Scanorama-like UMAP, and a post-integration UMAP.

Across tissues, the baseline embedding consistently shows strong donor effects, with batch labels aligning more closely with latent structure, indicating insufficient mixing. After integration, all three corrected embeddings—scVI, Scanorama-like, and UMAP-based integration—substantially reduce donor-driven variation, improving batch-mixing metrics while maintaining high biological signal according to ASW_label and graph connectivity.

Among the integrated methods, the Scanorama-like embedding most consistently achieves the best balance between batch mixing and cell-type preservation, offering strong donor integration while maintaining clearly separable biological clusters. The scVI latent space effectively removes batch effects but can slightly blur fine-grained cell-type boundaries, whereas the baseline preserves those boundaries at the expense of much stronger donor dependence.

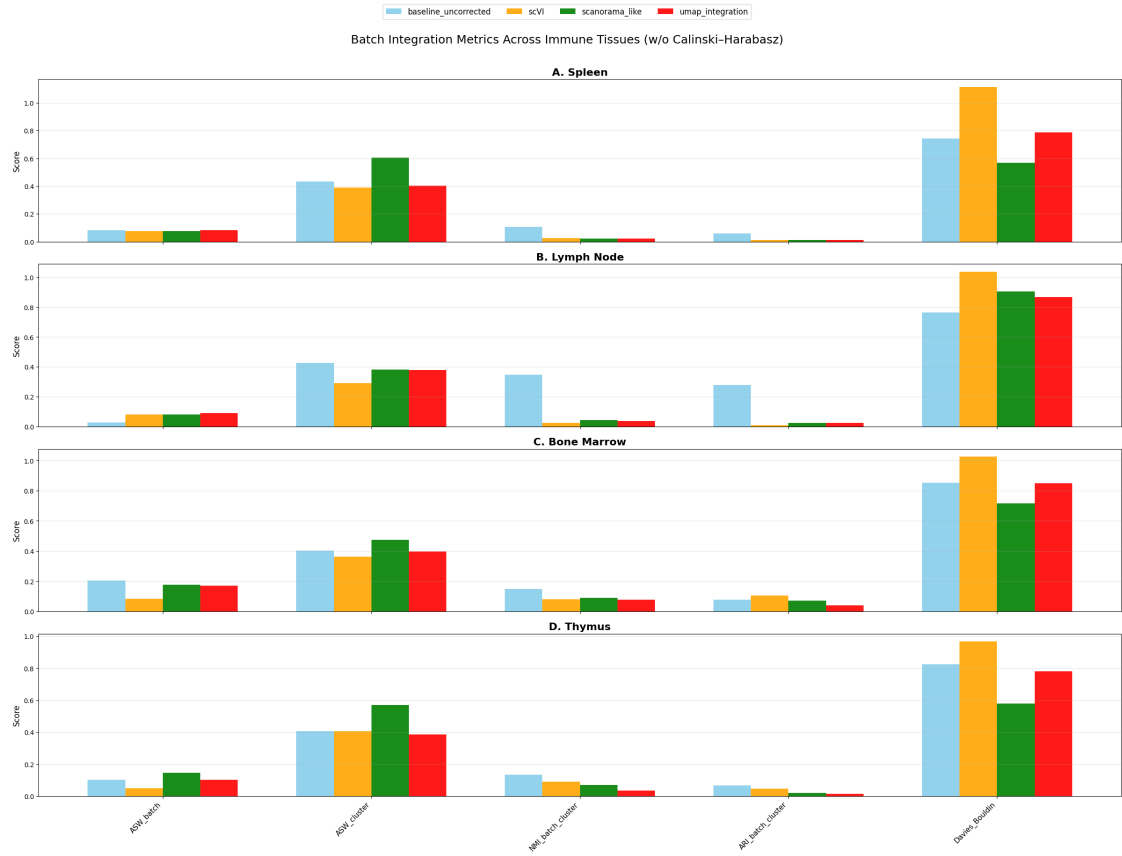


Figure 2: UMAP visualizations of immune tissue data before and after integration.

6 Conclusion and Future Work

This project established a framework for the context-aware benchmarking of scRNA-seq batch integration methods. By applying four leading algorithms to two distinct datasets representing major integration challenges, we demonstrated that method performance is not absolute but is intrinsically linked to data characteristics. Our analysis provides actionable guidance: for strong technical batch effects, consider scVI or Harmony; for preserving nuanced biological variation across similar batches, BBKNN or Scanorama may be preferable.

A primary limitation is the scope, restricted to four methods and two datasets. Given six more months, future work would focus on: 1) Expanding the benchmark to include newer methods (e.g., scANVI, MultiVI) and more diverse datasets (e.g., disease time courses, spatial transcriptomics); 2) Systematic hyperparameter tuning to distinguish between algorithmic limits and suboptimal parameter choices; 3) Developing a decision-support tool, such as a machine learning classifier that recommends the best method based on user-provided dataset features (size, batch strength, biological complexity). This would transform the project from a static comparison into a dynamic, community-informed resource for the field.

References

- Tabula Sapiens Consortium et al. The tabula sapiens: A multiple-organ, single-cell transcriptomic atlas of humans. *Science*, 376(6594):eabl4896, 2022.
- Jiarui Ding, Xian Adiconis, Sean K Simmons, Monika S Kowalczyk, Cynthia C Hession, Nemanja D Marjanovic, Travis K Hughes, Marc H Wadsworth, Tyler Burks, Lan T Nguyen, et al. Systematic comparison of single-cell and single-nucleus rna-sequencing methods. *Nature biotechnology*, 38(6):737–746, 2020.
- Brian Hie, Bryan Bryson, and Bonnie Berger. Efficient integration of heterogeneous single-cell transcriptomes using scanorama. *Nature biotechnology*, 37(6):685–691, 2019.
- Ilya Korsunsky, Nghia Millard, Jean Fan, Kamil Slowikowski, Fan Zhang, Kevin Wei, Yuriy Baglaenko, Michael Brenner, Po-ru Loh, and Soumya Raychaudhuri. Fast, sensitive and accurate integration of single-cell data with harmony. *Nature methods*, 16(12):1289–1296, 2019.
- Romain Lopez, Jeffrey Regier, Michael B Cole, Michael I Jordan, and Nir Yosef. Deep generative modeling for single-cell transcriptomics. *Nature methods*, 15(12):1053–1058, 2018.
- Krzysztof Polanski, Matthew D Young, Zhichao Miao, Kerstin B Meyer, Sarah A Teichmann, and Jong-Eun Park. Bbknn: fast batch alignment of single cell transcriptomes. *Bioinformatics*, 36(3):964–965, 2020.