Introduction:

The advent of single-cell technologies, specifically single-cell RNA sequencing (scRNA-seq) and single-cell Assay for Transposase-Accessible Chromatin sequencing (scATAC-seq) has revolutionized our understanding of cellular heterogeneity and the complex mechanisms of biological systems. These techniques offer unprecedented insights into the transcriptomic and epigenomic landscapes at the single-cell level, enabling researchers to unravel the intricacies of cellular functions and states in various biological contexts.

One of the most significant opportunities presented by these technologies is the ability to dissect the complex cellular compositions of tissues, providing a detailed atlas of cell types and states. This fine-grained resolution has been particularly transformative in identifying previously uncharacterized cell types in tissues like the human brain, where scRNA-seq revealed novel interneuron subtypes with unique transcriptional signatures [1]. Similarly, scATAC-seq has enabled the identification of lineage-specific chromatin accessibility profiles, uncovering regulatory elements critical to hematopoietic stem cell differentiation [2]. Furthermore, integrative analysis of scRNA-seq and scATAC-seq data, as demonstrated in studies of immune cells, has provided new insights into how transcription factors modulate cell states and immune responses [3].

However, these opportunities come with challenges. The large amount of data generated by single-cell techniques poses significant computational and analytical hurdles. For example, scRNA-seq datasets often suffer from batch effects and sparsity, requiring advanced algorithms for data integration and denoising, such as Harmony for batch correction [4]. Meanwhile, the integration of scATAC-seq and scRNA-seq data introduces complexities in linking chromatin accessibility to gene expression, which tools like Cicero attempt to resolve by constructing co-accessibility networks [5].

Recent advancements in bioinformatics have led to the development of various tools and algorithms tailored for single-cell data analysis. Methods like Seurat for scRNA-seq and ArchR for scATAC-seq have become foundational in this field, enabling researchers to identify transcriptional programs and chromatin accessibility landscapes at single-cell resolution [6, 3]. These tools have facilitated discoveries such as regulatory interactions governing T-cell activation and differentiation, advancing our understanding of immune cell plasticity [7].

Building upon the transformative capabilities of single-cell technologies, this project aims to address a key challenge in single-cell data analysis: the effective integration of scRNA-seq and scATAC-seq datasets. Integration is critical for combining complementary transcriptomic and epigenomic information, enabling a comprehensive understanding of cellular states and regulatory mechanisms. Despite advancements in computational tools, integrating such datasets remains complex due to batch effects, high dimensionality, and modality-specific characteristics. To tackle these challenges, we systematically evaluate three state-of-the-art methods—Optimal Transport, Triplet Loss, and Single-cell Deep Metric Learning (scDML)—for their ability to harmonize data while preserving biological fidelity. By applying these methods to the widely used PBMC 10k dataset and conducting detailed assessments through clustering metrics and marker gene analysis, this study seeks to highlight the strengths and limitations of each approach.

Methodology:

Dataset:

The PBMC 10k v3 dataset from 10x Genomics is a single-cell RNA sequencing dataset comprising gene expression profiles of approximately 10,000 Peripheral Blood Mononuclear Cells (PBMCs). PBMCs are critical components of the immune system, including diverse cell types such as T cells, B cells, NK cells, and monocytes, making them ideal for studying immune function and cell-type heterogeneity. This dataset was generated using the 10x Genomics Chromium platform and processed with Cell Ranger v3.0, ensuring high-quality data with well-filtered cell barcodes and features. The data is provided in an HDF5 format, containing a filtered feature-barcode matrix that captures the gene expression levels for high-quality single cells. This dataset is widely used for testing bioinformatics pipelines, exploring immune cell diversity, and benchmarking scRNA-seq data analysis tools.[1]

Data-preprocessing:

For scRNA-seq, we use a preprocessed and annotated dataset containing 13 distinct cell types, including immune cells such as B cell progenitors and various T cell subsets. The dataset undergoes rigorous quality control, normalization, and clustering using Seurat, which ensures accurate cell-type annotations and robust data processing. These steps provide a solid foundation for reliable and interpretable downstream analyses.

For scATAC-seq, we process the peak matrix, which identifies open chromatin regions across individual cells, using Signac.[2] A key step in this process is the transformation of the peak matrix into a gene activity matrix, which links chromatin accessibility to potential gene expression. This transformation involves associating accessible genomic regions—such as promoters, enhancers, and gene bodies—with their corresponding genes. In essence, the gene activity matrix transforms chromatin peaks into gene-associated features, summarizing accessibility data at the gene level. To accurately capture transcriptional regulation, promoter regions are extended upstream by up to 2 kilobases, based on the assumption that accessibility in these regions correlates with transcriptional activity. This step provides a proxy for gene expression, allowing chromatin accessibility data from scATAC-seq to be represented in a feature space directly comparable to scRNA-seq data. The primary motivation for transforming scATAC-seq data into a gene activity matrix is to align it with scRNA-seq data in a meaningful way. This transformation allows the two distinct modalities to share a common feature space based on genes, enabling a biologically interpretable comparison. Although the data originates from different modalities, the integration process simulates horizontal integration by treating both datasets as if they were of the same modality. Shared genes serve as anchors for alignment, ensuring consistency and comparability across datasets while preserving biological relevance. By creating a unified dataset through this alignment, we enable tools such as Optimal Transport, Triplet Loss, and scDML to be benchmarked effectively, providing a robust comparison of their ability to integrate multi-omics data while retaining critical biological signals.

After preprocessing, cells with more than 5,000 total peak counts are retained, refining the scATAC-seq dataset to 7,866 high-quality cells with robust chromatin accessibility signals. This ensures the inclusion of only high-quality data, reducing noise and improving reliability. The refined dataset is optimized for

[1] https://www.10xgenomics.com/platforms/chromium?utm_medium=display&utm_source=google&utm_term=dis-goog-2023-11-website-page-chr-crx-chromium-pmax-tofu-8027&utm_content=website-page&utm_campaign=701KW00000192lpYAA&gad_source=1

[2] https://vipcca.readthedocs.io/en/latest/tutorials/vipcca_atac_tutorial.html

integration with scRNA-seq data, providing a unified platform to investigate regulatory dynamics and cellular processes. By emphasizing biological relevance and working within a shared feature space, this vertical integration strategy facilitates a deeper understanding of gene regulation and cellular networks.

First, logarithmic normalization stabilizes variance across genes and identifies highly variable genes separately in each dataset. The intersection of these HVG sets is then taken to create a shared set of highly variable genes common across all datasets. Each dataset is subsequently filtered to include only this shared set of HVGs, ensuring that all datasets are aligned with the same set of features. The aligned datasets are then integrated by concatenating them along the cell axis. This process ensures that all datasets share a consistent feature set while preserving unique cell identifiers to distinguish their origins. Any missing genes in a dataset are padded with zeros, enabling seamless integration.

Following integration, the combined dataset is scaled and subjected to principal component analysis to reduce dimensionality, simplifying the complex data into principal components. Clustering is performed using the Leiden algorithm, grouping cells based on gene expression profiles. The resulting clusters are visualized with UMAP to provide insights into the cellular heterogeneity and relationships. Finally, differential expression testing identifies genes that vary significantly between clusters, uncovering molecular factors that distinguish cell types or states.

Optimal Transport:

To integrate single-cell ATAC-seq and RNA-seq data, we employ an **entropy-regularized optimal transport (OT) framework**, a principled method for aligning datasets from different modalities by treating each as a probability distribution. ATAC-seq captures chromatin accessibility, indicating regulatory activity, while RNA-seq measures gene expression, reflecting functional outcomes. Integration of these datasets is challenging due to differences in their underlying data structures, batch effects, and noise. Optimal transport addresses these challenges by computing a coupling matrix, which represents a probabilistic alignment of cells between the two datasets based on shared biological features, such as genes. This probabilistic alignment is particularly suitable for single-cell data, where high variability and noise often make rigid one-to-one mappings biologically unrealistic.

The OT framework minimizes a biologically informed cost function, which quantifies dissimilarity between cells based on shared features, such as gene activity, while enforcing marginal constraints to preserve the overall distributions of the original datasets. To enhance biological relevance, we preprocess the data by constructing k-nearest neighbor (k-NN) graphs that capture local structures within each dataset. These graphs are then refined using shortest-path distances, ensuring that the cost function reflects biologically meaningful connectivity rather than raw pairwise distances. Entropy regularization is introduced into the optimization process to ensure smoothness in the coupling matrix, allowing cells to align flexibly with multiple potential counterparts. This regularization also enhances robustness to noise and improves computational efficiency, making the framework scalable to large datasets.

To solve the entropy-regularized optimal transport problem, we utilize the Sinkhorn algorithm, which iteratively adjusts the coupling matrix to minimize the cost function while satisfying marginal constraints. The resulting coupling matrix provides a biologically informed alignment between cells in ATAC-seq and RNA-seq datasets. Following this, barycentric projection is applied to map cells from one dataset into the feature space of the other, aligning them into a unified space that preserves intrinsic biological relationships while correcting for batch effects. This integration facilitates downstream analyses such as clustering, visualization, and cell-type annotation. By fine-tuning parameters, such as the number of neighbors in the k-NN graph and regularization strengths, we ensure that the alignment is both robust and interpretable, providing a reliable framework for integrating multi-omics single-cell data. [8]

INSCT: Batch-Aware Triplet Sampling

INSCT integrates scRNA-seq and scATAC-seq datasets using a triplet loss function, designed to learn a unified embedding space that minimizes batch effects while preserving biological relationships. It dynamically defines triplets (anchor, positive, negative) using Mutual Nearest Neighbors (MNN) and k-Nearest Neighbors (KNN). MNNs identify inter-batch positive cells that are transcriptionally similar but from different batches, while KNNs identify intra-batch positive cells for anchors lacking MNNs, ensuring comprehensive coverage of cell types. Negative cells are randomly sampled within the same batch, leveraging their inherent transcriptional dissimilarity. This batch-aware approach forces the model to align cells across batches, capturing both local and global transcriptional similarities. The neural network architecture, derived from ivis, consists of three dense layers with Alpha Dropout to prevent overfitting and outputs a two-dimensional embedding optimized for visualization and analysis. Triplet sampling is performed dynamically at each epoch, ensuring that training captures diverse relationships without constructing explicit graphs or clusters. INSCT's lightweight implementation and efficient KNN computation make it scalable for large datasets, focusing on direct alignment through dynamic triplet optimization. [9]

scDML: Hierarchical Clustering and Metric Learning

scDML builds upon traditional triplet loss by introducing a hierarchical clustering step informed by graph-based relationships to address batch effects comprehensively. The method begins by constructing a joint graph combining intra-batch KNN and inter-batch MNN relationships, capturing both local and global similarities. A similarity matrix derived from the graph guides hierarchical clustering, iteratively merging clusters across batches to create batch-corrected clusters that emphasize biological variation. After this preprocessing, scDML refines the embedding space by applying metric learning with triplet loss. In contrast to INSCT, triplets are now defined based solely on biological distinctions: anchor-positive pairs are sampled from the same updated cluster, while negatives are drawn from different clusters. This ensures the embedding reflects biological coherence without residual batch effects. The neural network architecture uses principal component inputs and dense layers to output a compact 32-dimensional embedding optimized through hard triplet prioritization. This refined embedding captures subtle biological variations while eliminating batch influences, providing a robust foundation for clustering, visualization, and downstream analyses. [10]

Evaluation:

Figure 1 illustrates the significant batch effect, evident from the clear separation of clusters by omics type. Among the three integration methods, Optimal Transport(Figure 2) demonstrates the best performance, effectively integrating both omics datasets and minimizing batch-induced separation. In contrast, the other two tools(Figure 2,3) show only limited integration, with clusters displaying partial mixing of omics but failing to achieve the same level of alignment as OT.

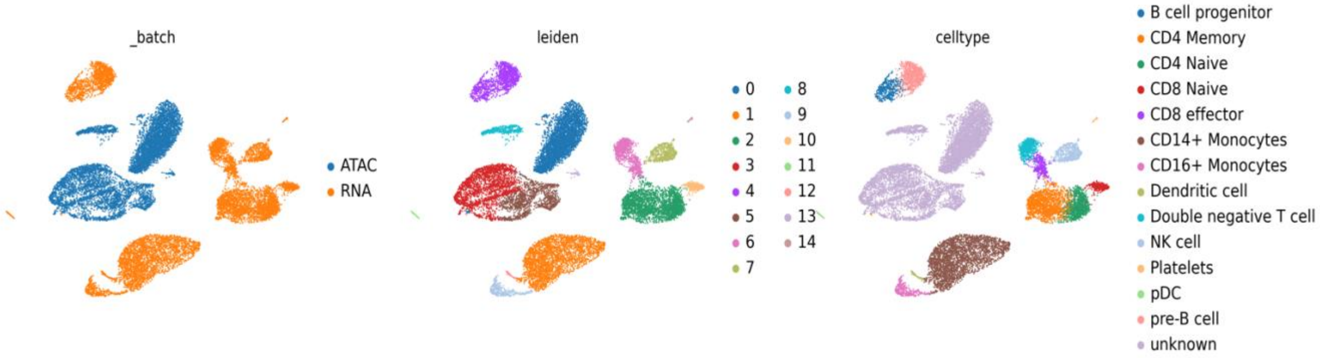


Figure 1 UMAP projection of unintegrated data

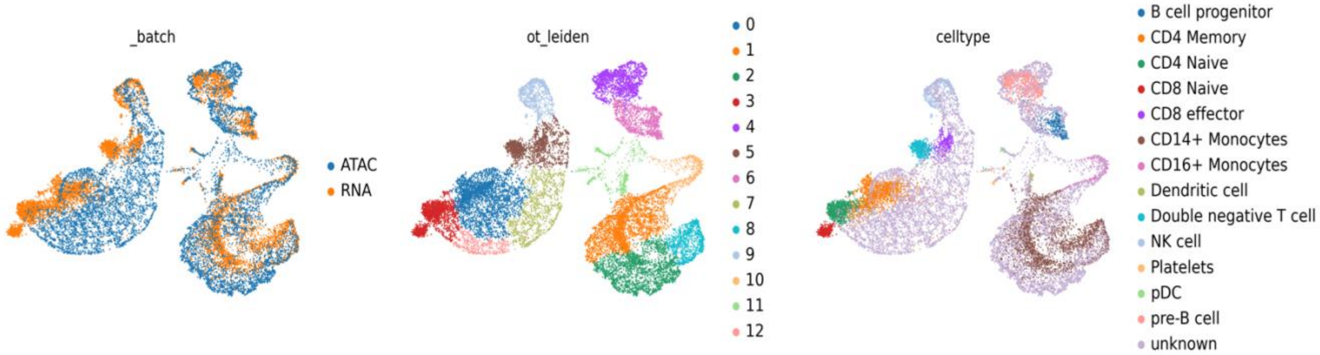


Figure 2 UMAP projection of optimal transferred data

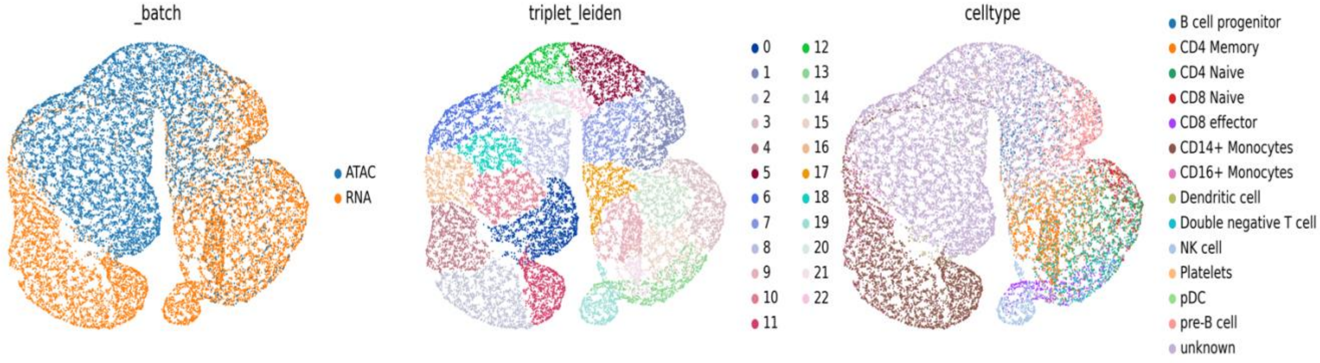


Figure 3 UMAP projection of integrated data using triplet loss

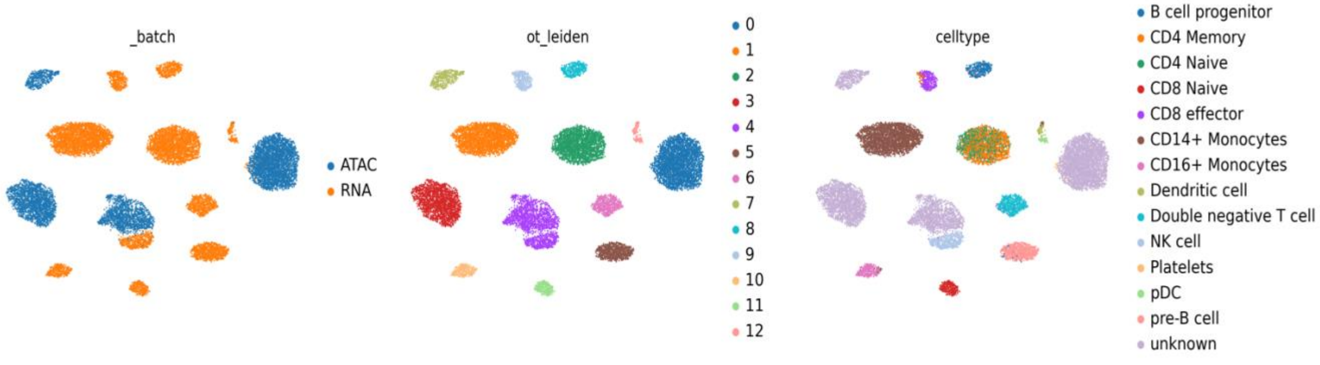


Figure 4 UMAP projection of integrated data using scDML

Metric	Raw	Triplet Loss	Optimal Transport	scDML
ASW Batch	0.320330	0.147360	0.019590	0.258388
ASW Cell Type	0.278077	0.279886	0.040106	0.609408
ARI Batch	0.943689	0.036149	0.000128	0.081027
ARI Cell Type	0.310411	0.140147	0.191035	0.357269
NMI Batch	0.905499	0.024198	0.000009	0.227707
NMI Cell Type	0.696635	0.486641	0.503241	0.728474

Table 1 evaluation metrics

Average Silhouette Width (ASW) Analysis

The ASW assesses how effectively individual cells align with their assigned clusters by comparing their similarity to points within the same cluster (cohesion) against their similarity to points in the nearest neighboring cluster (separation). For batch-specific analysis, a high ASW score suggests that cells cluster tightly by batch, reflecting poor integration dominated by technical artifacts. Conversely, a lower ASW score after integration indicates effective batch mixing, where clustering is influenced more by biological signals than batch-specific characteristics. For cell type analysis, a high ASW reflects cohesive clustering of cells based on their biological type, ensuring good cell type purity, while a lower ASW suggests potential overlap or poor distinction between cell types.

From Table 1, the initial ASW score for batches in raw data was 0.320330, highlighting batch-specific clustering. Post-integration, Optimal Transport achieved the lowest ASW for batches at 0.019590, effectively eliminating batch-specific clustering. Triplet Loss and scDML also reduced the ASW for batches to 0.147360 and 0.258388, respectively, reflecting improved batch mixing, though not as extensively as Optimal Transport. For cell types, the raw data had an ASW of 0.278077, indicating moderate clustering of cell types. Post-integration, scDML achieved the highest ASW for cell types at 0.609408, demonstrating its ability to preserve biological distinctions. Conversely, Optimal Transport resulted in a much lower ASW for cell types at 0.040106, suggesting that while batch effects were minimized, biological distinctions between cell types were less pronounced.

Adjusted Rand Index (ARI) Analysis

The Adjusted Rand Index (ARI) evaluates the similarity between two clustering assignments, measuring how well cells are grouped relative to predefined labels. For batch-specific analysis, a high ARI score indicates clustering aligned with batch labels, reflecting batch effects, while a low ARI score indicates successful integration where clustering is no longer driven by batch effects. For cell type analysis, a high ARI indicates that clustering aligns well with true biological cell type labels, reflecting preserved biological fidelity after integration.

In Table 1, the raw data showed a high ARI for batches at 0.943689, indicating strong batch effects. Post-integration, Optimal Transport achieved the lowest ARI for batches at 0.000128, signifying near-complete elimination of batch effects. Triplet Loss and scDML reduced the ARI for batches to 0.036149 and 0.081027, respectively, reflecting substantial batch effect mitigation but less so than Optimal Transport. For cell types, the raw data had an ARI of 0.310411, showing moderate biological clustering. scDML improved the ARI for cell types to 0.357269, the highest among the methods, indicating strong preservation of biological distinctions. Optimal Transport had a lower ARI for cell types at 0.191035, suggesting a trade-off where batch correction was prioritized over cell type fidelity.

Normalized Mutual Information (NMI) Analysis

Normalized Mutual Information (NMI) quantifies the similarity between two clustering outcomes. For batch-specific analysis, a high NMI reflects clustering dominated by batch labels, indicating poor integration, while a low NMI suggests effective batch correction with clustering independent of batch labels. For cell type analysis, a high NMI signifies that clustering aligns with true biological cell type categories, demonstrating preserved biological fidelity.

From Table 1, the raw data had a high NMI for batches at 0.905499, reflecting batch-driven clustering. Optimal Transport achieved the lowest NMI for batches at 0.000009, highlighting its success in removing batch effects. Triplet Loss and scDML reduced the NMI for batches to 0.024198 and 0.227707, respectively, indicating moderate to significant batch effect correction. For cell types, the raw data showed an NMI of 0.696635, indicating reasonable biological clustering. Post-integration, scDML achieved the highest NMI for cell types at 0.728474, emphasizing its ability to retain biological fidelity. Optimal Transport and Triplet Loss had lower NMI scores for cell types at 0.503241 and 0.486641, respectively, reflecting their limited ability to preserve biological clustering while correcting batch effects.

Overall Comparison Across Metrics

Optimal Transport emerged as the most effective method for mitigating batch effects, achieving the lowest ASW, ARI, and NMI for batches, indicating near-complete elimination of batch-specific clustering. However, this came at the cost of cell type preservation, as evidenced by its lower ASW, ARI, and NMI scores for cell types. In contrast, scDML balanced batch effect removal and cell type fidelity, achieving the highest scores for cell type metrics while moderately reducing batch effects. Triplet Loss demonstrated a balanced but less pronounced performance across all metrics, showing moderate improvements in batch effect removal and cell type preservation.

Marker gene analysis:

In the optimal transport integrated data, the preservation of biological characteristics was validated through the analysis of cell type-specific marker gene expression, presented in Figure 5. The analysis showed consistent expression patterns within cell types, indicating that cells with similar biological properties were accurately grouped in the integrated embedding. Each cell type, identified by unique marker gene expression, was distinctly mapped to specific regions in the embedding, confirming that the optimal transport integration successfully retained the biological significance of the data.

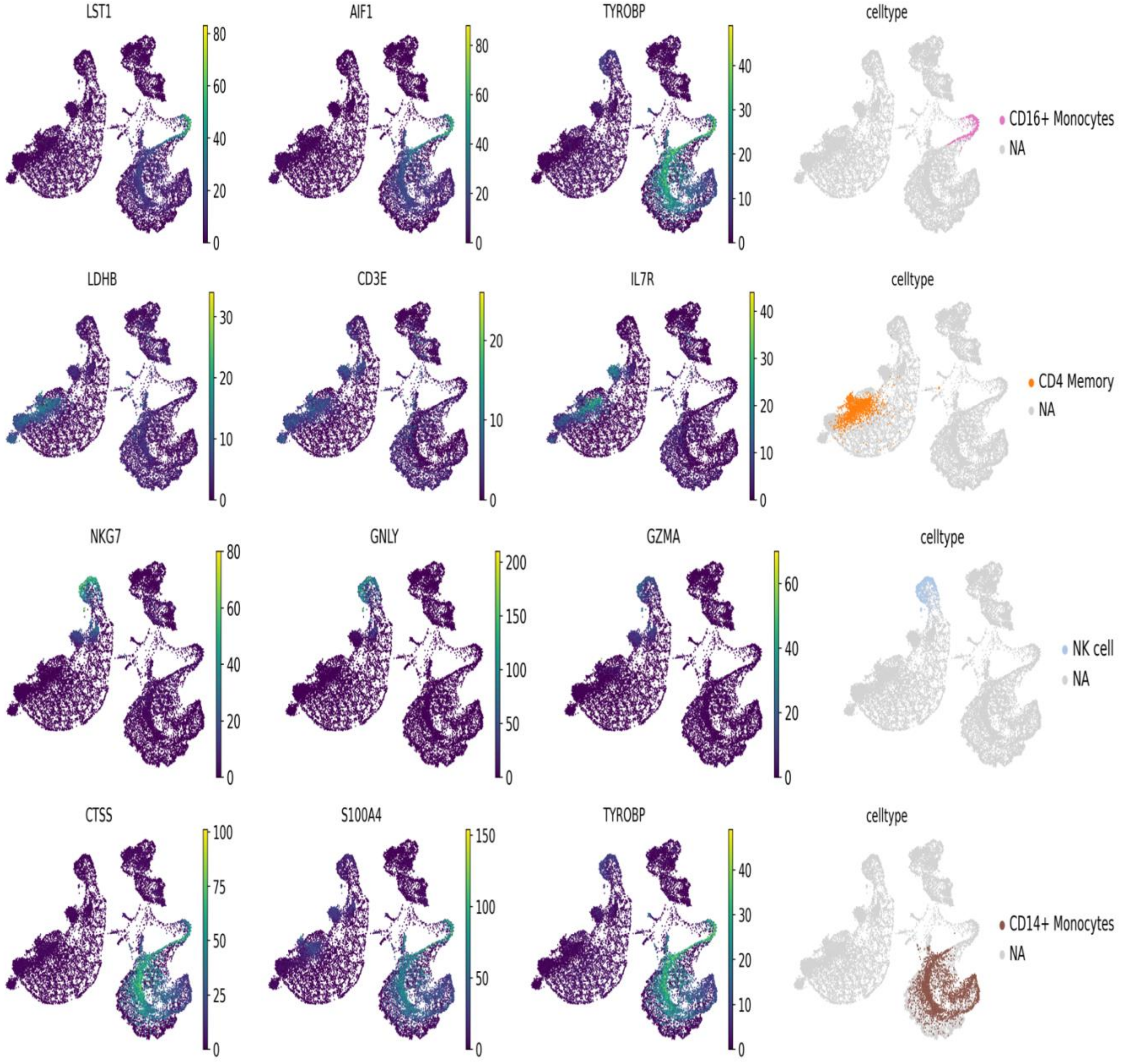


Figure 5 Top three marker gene for four sampled cell type from OT transformed data.

The conservation of biological properties in the data integrated via triplet loss was validated by evaluating the expression patterns of cell type-specific marker genes, as shown in Figure 6. This study revealed consistent expression patterns across cells of the same kind, confirming that the integration strategy successfully grouped cells with similar biological features. The triplet loss approach effectively localized cell types to areas inside the embedding space, comparable to the findings obtained with optimal transport integration.

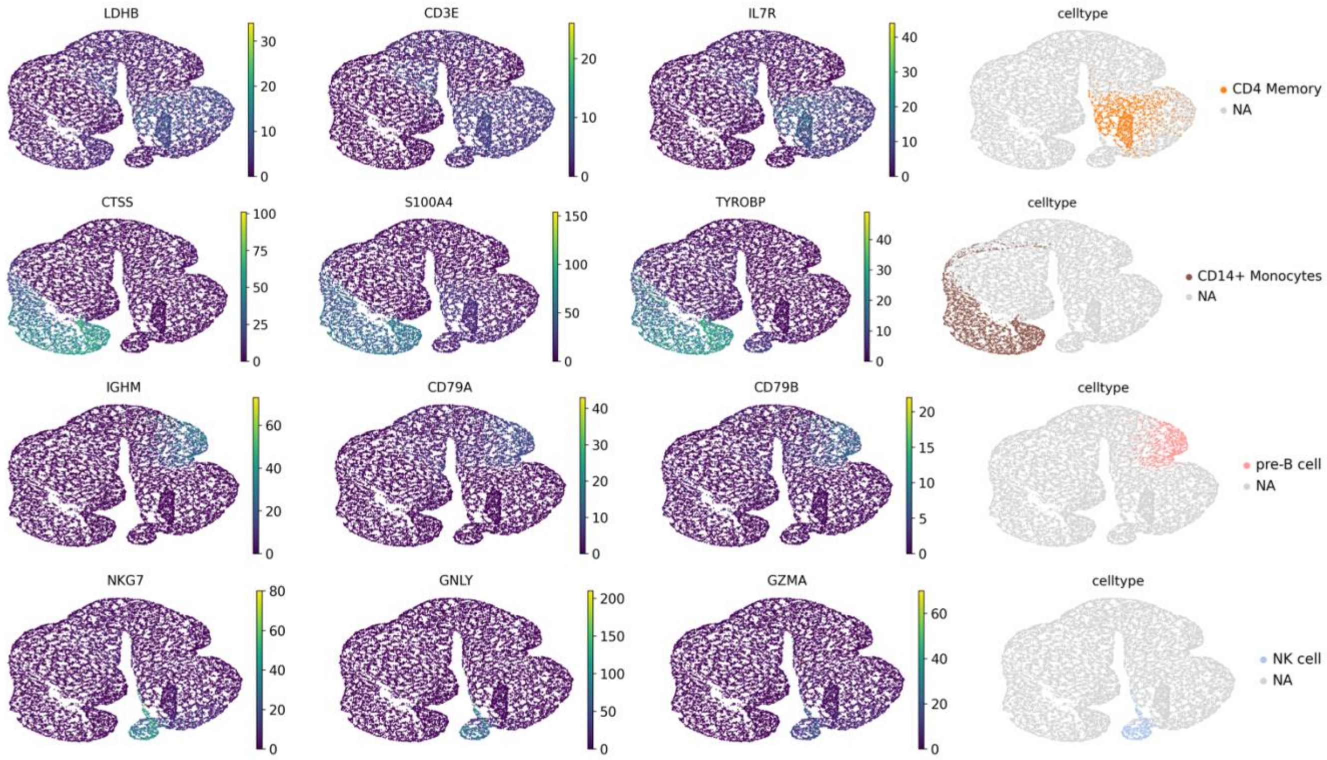
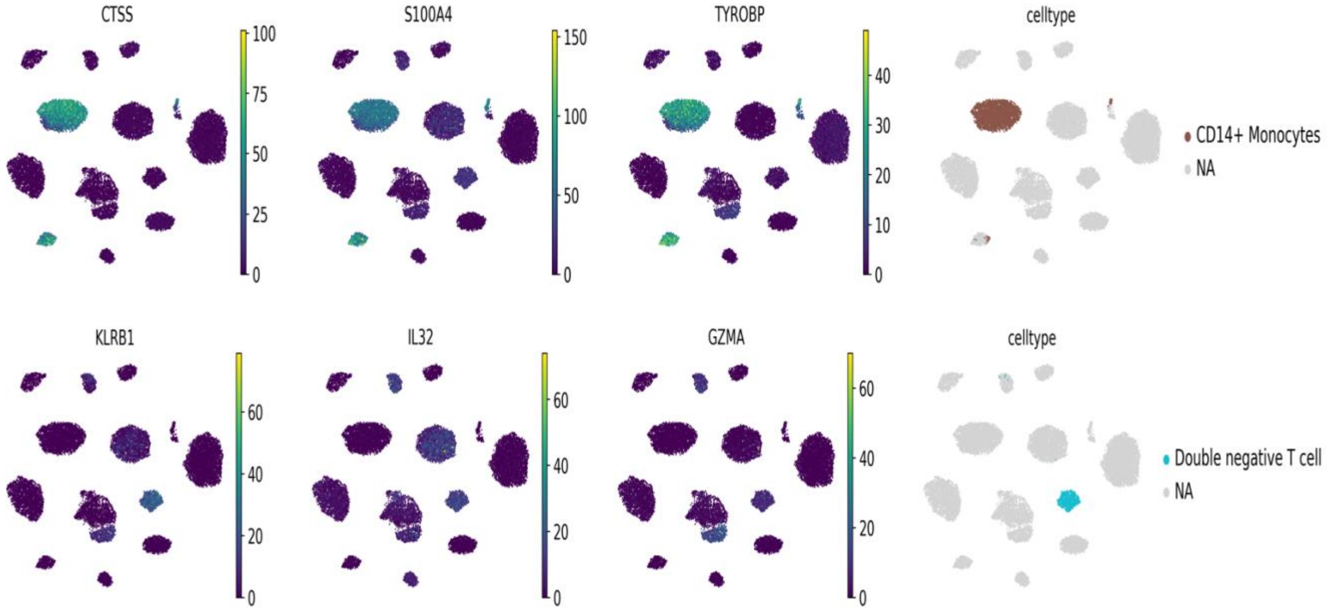


Figure 6 Top three marker gene for four sampled cell type from triplet loss integrated data.

Figure 7 demonstrates the preservation of cell-type-specific marker gene expression following data integration. High expression levels of marker genes are confined to their respective cell-type clusters within the embedding space, indicating that the integration effectively maintained the biological characteristics of each cell type.



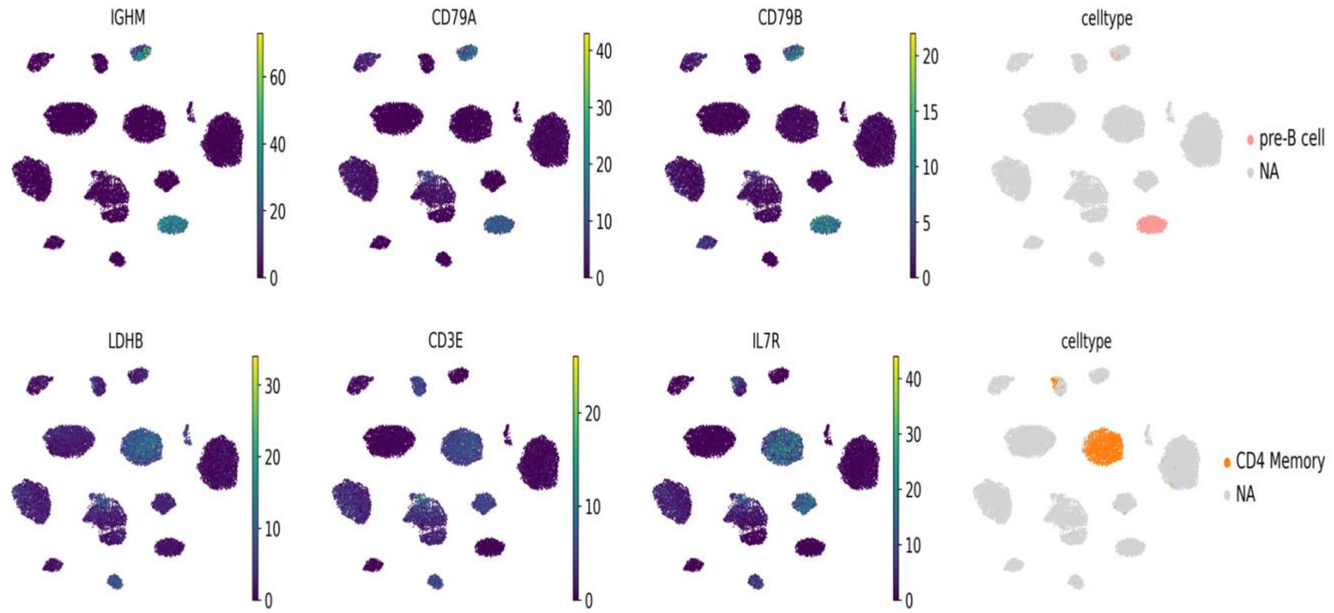


Figure 7 Top three marker gene for four sampled cell type from scDML integrated data.

Summary and Conclusion

In this study, we tackled the pervasive challenge of omics data integration in single-cell RNA sequencing data integration by evaluating three advanced computational methods: Optimal Transport, Triplet Loss, and Single-cell Deep Metric Learning. These methods, representative of cutting-edge approaches in data harmonization, were systematically assessed for their ability to align data from disparate batches while preserving critical biological insights. The integration quality was comprehensively evaluated using robust metrics such as Average Silhouette Width (ASW), Adjusted Rand Index (ARI), and Normalized Mutual Information (NMI), coupled with marker gene analyses to validate the retention of biological features. Our findings demonstrate distinct strengths and trade-offs among the three integration methods, each excelling in specific aspects of batch effect mitigation and cell-type preservation.

The **Optimal Transport** method emerged as the most effective for batch effect removal, achieving the lowest batch-specific ASW (0.019590) and ARI (0.000128), indicating a near-complete elimination of batch-driven clustering. These results highlight its superior ability to harmonize data from different batches into a cohesive embedding, effectively minimizing batch-specific biases. Marker gene analysis further validated its biological relevance, as cell-type marker genes were consistently expressed within their corresponding clusters, confirming that the method retained essential biological characteristics. However, the method showed limitations in preserving fine-grained cell-type distinctions, with lower ARI (0.191035) and NMI (0.503241) scores for cell types. This suggests that while Optimal Transport prioritizes batch effect mitigation, it may sacrifice some resolution in distinguishing between cell types.

The **scDML** approach demonstrated a notable strength in preserving cell-type integrity. It achieved the highest ARI (0.357269) and NMI (0.728474) scores for cell types, reflecting its ability to retain and enhance biological clustering. While scDML was slightly less effective in reducing batch effects, as evidenced by higher batch-specific ASW (0.258388) and ARI (0.081027) scores compared to Optimal Transport, it still showed substantial improvements over raw data. This trade-off suggests that scDML is particularly suitable for studies prioritizing the identification and analysis of distinct cellular identities and transitions, as it excels in maintaining biologically meaningful clusters across batches.

The **Triplet Loss** method offered a balanced performance, providing moderate success in both batch effect mitigation and cell-type preservation. Its batch-specific ASW (0.147360) and ARI (0.036149) scores indicate significant improvements in batch alignment compared to the raw data, though not to the extent achieved by Optimal Transport. Similarly, its cell-type-specific ARI (0.140147) and NMI (0.486641) scores suggest reasonable retention of cell-type identity, though not as strong as scDML. Marker gene analysis confirmed that Triplet Loss maintained biologically coherent clusters, making it a versatile method for applications requiring a balance between batch effect correction and biological accuracy.

In summary, the choice of integration method should align with the specific objectives of the analysis. **Optimal Transport** is ideal for rigorous batch effect removal and global alignment, **scDML** excels in preserving detailed cell-type information, and **Triplet Loss** provides a pragmatic balance between batch correction and biological fidelity. These findings underscore the importance of tailoring the integration approach to the unique requirements of the study.

Implications for Future Research

This study highlights the critical importance of selecting appropriate integration methods in single-cell data analysis. The performance trade-offs among the evaluated methods underscore that no single approach is universally superior across all metrics. Instead, the choice of integration technique should be tailored to the specific scientific question at hand, such as prioritizing batch effect mitigation or cell-type identity preservation.

Future research should focus on further optimizing these methods to address their limitations. For instance, while Optimal Transport delivered exceptional batch correction, its computational intensity may limit scalability for larger datasets. Developing more efficient algorithms or hybrid approaches that combine the strengths of Optimal Transport and scDML could offer a pathway to improved integration strategies. Similarly, scDML's performance in batch effect correction could benefit from enhancements in its underlying network architecture or the inclusion of additional loss functions tailored for multi-omic data.

Additionally, testing these methods on **diverse, large-scale datasets** from different biological systems and experimental contexts will be essential to evaluate their robustness and generalizability. Current benchmarks often focus on specific datasets, limiting insights into how these methods perform under varying conditions, such as datasets with rare cell populations, extreme batch effects, or high levels of noise. Standardizing evaluation metrics and creating comprehensive benchmark datasets that encompass diverse scenarios will further accelerate progress in the field. Moreover, integrating data from multiple modalities, such as scRNA-seq with scATAC-seq or proteomics data, represents a promising frontier. The development of integration frameworks that can seamlessly handle multi-omic datasets while maintaining biological accuracy will enable deeper insights into cellular systems.

In conclusion, this study not only demonstrates the potential of advanced computational methods for single-cell data integration but also emphasizes the need for thoughtful method selection based on specific analytical goals. By addressing the challenges posed by batch effects and leveraging these innovative techniques, researchers can unlock more accurate and biologically meaningful interpretations of complex single-cell datasets, paving the way for transformative discoveries in biomedical research.

[1] Lake, B., Chen, S., Sos, B. *et al.* Integrative single-cell analysis of transcriptional and epigenetic states in the human adult brain. *Nat Biotechnol* **36**, 70–80 (2018). https://doi.org/10.1038/nbt.4038

[2] Pan, Q., Zhou, H., Lu, Q. *et al.* History-independent cyclic response of nanotwinned metals. *Nature* **551**, 214–217 (2017). https://doi.org/10.1038/nature24266

[3] Granja, J.M., Klemm, S., McGinnis, L.M. *et al.* Single-cell multiomic analysis identifies regulatory programs in mixed-phenotype acute leukemia. *Nat Biotechnol* **37**, 1458–1465 (2019). https://doi.org/10.1038/s41587-019-0332-7

[4] Korsunsky, I., Millard, N., Fan, J. *et al.* Fast, sensitive and accurate integration of single-cell data with Harmony. *Nat Methods* **16**, 1289–1296 (2019). https://doi.org/10.1038/s41592-019-0619-0

[5] Pliner HA, Packer JS, McFaline-Figueroa JL, Cusanovich DA, Daza RM, Aghamirzaie D, Srivatsan S, Qiu X, Jackson D, Minkina A, Adey AC, Steemers FJ, Shendure J, Trapnell C. Cicero Predicts cis-Regulatory DNA Interactions from Single-Cell Chromatin Accessibility Data. Mol Cell. 2018 Sep 6;71(5):858-871.e8. doi: 10.1016/j.molcel.2018.06.044. Epub 2018 Aug 2. PMID: 30078726; PMCID: PMC6582963.

[6] Hao Y, Hao S, Andersen-Nissen E, Mauck WM 3rd, Zheng S, Butler A, Lee MJ, Wilk AJ, Darby C, Zager M, Hoffman P, Stoeckius M, Papalexi E, Mimitou EP, Jain J, Srivastava A, Stuart T, Fleming LM, Yeung B, Rogers AJ, McElrath JM, Blish CA, Gottardo R, Smibert P, Satija R. Integrated analysis of multimodal single-cell data. Cell. 2021 Jun 24;184(13):3573-3587.e29. doi: 10.1016/j.cell.2021.04.048. Epub 2021 May 31. PMID: 34062119; PMCID: PMC8238499.

[7] Stuart, T., Satija, R. Integrative single-cell analysis. *Nat Rev Genet* **20**, 257–272 (2019). https://doi.org/10.1038/s41576-019-0093-7

[8] Demetci P, Santorella R, Sandstede B, Noble WS, Singh R. SCOT: Single-Cell Multi-Omics Alignment with Optimal Transport. J Comput Biol. 2022 Jan;29(1):3-18. doi: 10.1089/cmb.2021.0446. PMID: 35050714; PMCID: PMC8812493.

[9] Simon, L.M., Wang, YY. & Zhao, Z. Integration of millions of transcriptomes using batch-aware triplet neural networks. *Nat Mach Intell* **3**, 705–715 (2021). https://doi.org/10.1038/s42256-021-00361-8

[10] Yu, X., Xu, X., Zhang, J. *et al.* Batch alignment of single-cell transcriptomics data using deep metric learning. *Nat Commun* **14**, 960 (2023). https://doi.org/10.1038/s41467-023-36635-5