sCIN: A Contrastive Learning Framework for single-cell Multi-omics Data Integration

Amir Ebrahimi1 , Alireza Fotuhi Siahpirani2\*, Hesam Montazeri2\*

1 Department of Biotechnology, College of Science, University of Tehran, Tehran, Iran

2 Department of Bioinformatics, Institute of Biochemistry and Biophysics,  University of Tehran, Tehran, Iran

\* Correspondence to Hesam Montazeri hesam.montazeri@ut.ac.ir and and Alireza Fotuhi Siahpirani [a.fotuhi@ut.ac.ir](mailto:a.fotuhi@ut.ac.ir)

# Abstract

The rapid advancement of single-cell omics technologies, such as scRNA-seq and scATAC-seq, has changed our understanding of cellular heterogeneity and regulatory mechanisms. However, integrating data from these modalities remains a significant challenge due to distributional discrepancies and different feature spaces, making it hard to use comprehensive sources of knowledge about biological systems. To address this, we present sCIN (single-cell Contrastive INtegration), a novel framework that integrates paired single-cell multi-omics datasets into a shared low-dimensional latent space. sCIN utilizes modality-specific encoders and a contrastive learning approach to simultaneously learn latent representations for each modality and match cells with the same identities while removing technology-specific biases. The framework was carefully evaluated on three real-world datasets to prevent test data leakage. Results show that sCIN outperforms baseline and state-of-the-art models based on cell-type identification and modality alignment in multi-omics integration tasks. Moreover, we benchmarked sCIN and other models in two unpaired settings (with and without optimal transport). These findings position sCIN as a high-potential platform, capable of extending to a wide range of downstream applications such as cell-type annotation and disease subtyping, advancing the integration and analysis of single-cell multi-omics data.

**Keywords:** single-cell multi-omics, data integration, contrastive learning, neural networks, multi-modal learning

# Introduction

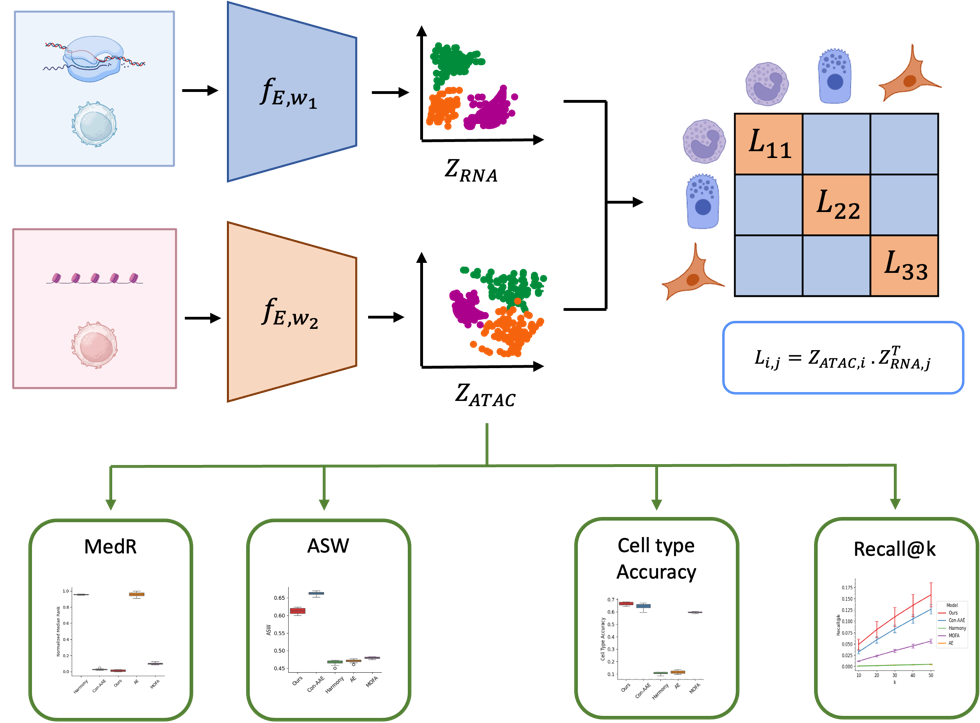
The development of single-cell omics technologies has transformed our ability to explore complex biological systems, providing a clearer understanding of their phenotypes across different omics layers and cell types. For instance, single-cell RNA sequencing (scRNA-seq) allows us to measure gene expression by capturing the abundance of messenger RNA (mRNA) molecules for each gene in thousands, or even millions, of individual cells [1]. Moreover, Single-cell assay for transposase-accessible chromatin with high throughput sequencing (scATAC-seq) focuses on identifying open chromatin regions, which are critical for gene regulation since they are hotspots for transcription factors (TFs) and other regulatory elements [2]. Since Transcription Factors (TFs) and other regulatory elements are more active in open chromatin regions, scATAC-seq provides additional information on gene regulation [3]. In addition to scATAC-seq, other technologies help map chromatin accessibility [4–6] and DNA methylation patterns [7,8] each adding unique insights into the molecular mechanisms that control gene regulation.

Gene expression is a complex process involving multiple stages, including transcription, post-transcriptional regulation, translation, and post-translational modifications. To better understand these processes, there has been a growing demand for sequencing technologies capable of simultaneously measuring multiple molecular features of a cell. Such technologies provide a more comprehensive view of cellular function and reduce batch effects that can arise from separate experiments. For instance, SHARE-seq [9] and SNARE-seq [10] simultaneously profile the transcriptome and chromatin accessibility of a cell population by integrating DNA fragmentation (via transposase enzymes) and mRNA reverse transcription. CITE-seq [11] combines antibody-based tagging with reverse transcription to capture both transcriptome and cell surface protein profiles within the same cells. However, despite the vast amounts of high-throughput data generated by these methods, analyzing each data modality independently can lead to fragmented insights and an incomplete understanding of the biological system. One approach to overcoming this challenge is to integrate the different modalities of omics data for analysis and inference. However, this “integration task” is non-trivial due to disparities in the distributions and feature spaces of the various modalities.

Several computational methods have been developed in recent years to address this problem. Most of these approaches aim to learn a low-dimensional joint representation of multiple modalities using dimensionality reduction techniques such as Principal Component Analysis PCA[12], Canonical Correlation Analysis(CCA)[13–15], or more advanced non-linear methods like deep neural networks[16–18]. For instance, Seurat[19] applies PCA to transform data linearly and uses Mutual Nearest Neighbors and CCA to align embeddings in a shared latent space. While effective, this approach is limited in capturing non-linear relationships between modalities. Another example is MOFA [20], which employs matrix factorization to decompose input data matrices from different modalities into a shared matrix of factors for samples and weight matrices for each modality. Meanwhile, Harmony [21] uses fuzzy K-Nearest Neighbor (KNN) clustering to group cells based on cell types while maximizing batch diversity within each cluster, offering a robust method for batch effect correction and modality alignment.

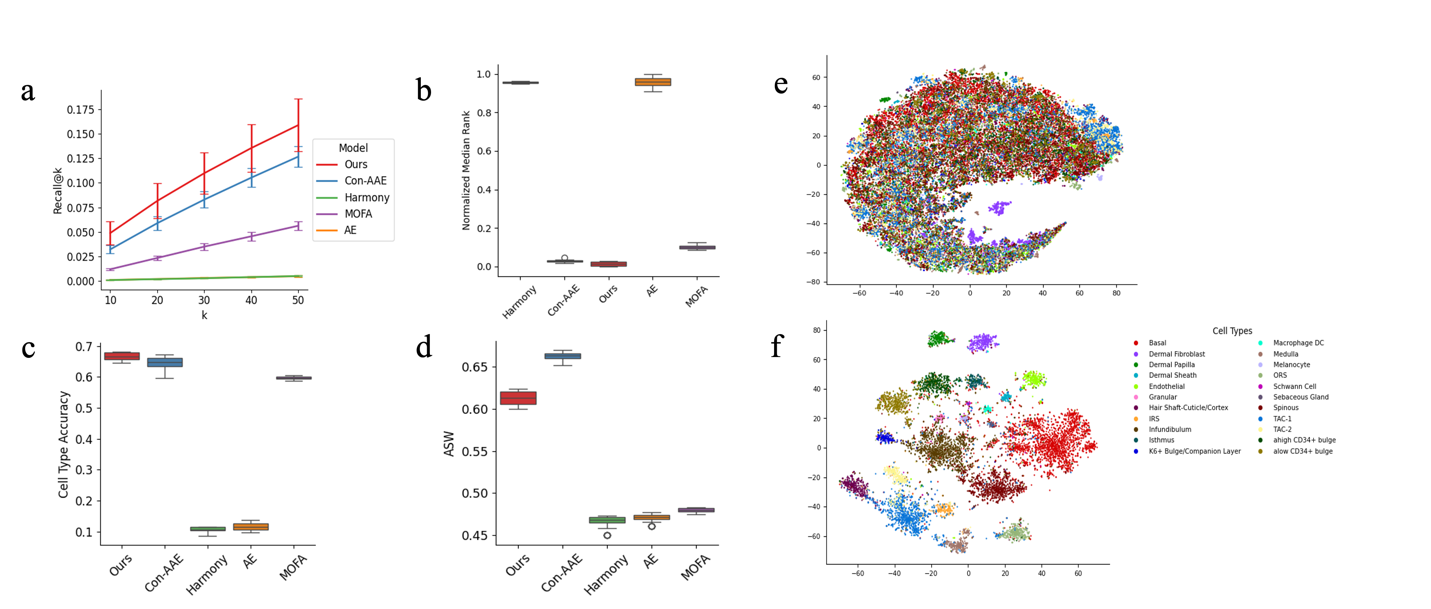
Additionally, there are a lot of efforts in developing multi-modal deep learning frameworks due to their ability to learn a hierarchical representation of each modality and handle high-dimensional data through mapping various modalities’ feature spaces into a unified and smaller subspace [22–26]. For instance, scglue adopts a graph variational autoencoder to uncover the interactions between biological features(e.g., genes, accessible regions, methylated regions, etc) and integrates them with the joint latent embeddings of modalities learned by modality-specific variational autoencoders [18]. GCN-SC [27] uses mutual nearest neighbors to link cells across datasets, constructs a mixed graph, and adjusts query datasets via GCN, followed by non-negative matrix factorization for dimension reduction and visualization. MultiVI [28] is a variational autoencoder-based framework that integrates scRNA-seq and scATAC-seq data by learning modality-specific latent representations through deep neural network encoders, correcting for batch effects, and aligning these representations into a shared latent space. Although these methods show promising results and achieve good performance, they often suffer from sensitivity to low signal-to-noise ratio(SNR) and sparsity, which are the key attributes of single-cell datasets[29]. Moreover, some state-of-the-art models, specifically deep learning frameworks, overcomplicate the model architecture, which compromises interpretability.

Here, we present sCIN, a simple yet powerful framework that projects different modalities of paired single-cell omics datasets into a shared low-dimensional latent space that preserves the cell type heterogeneity while removing the technology effects. Particularly, sCIN consists of two modality-specific encoders that learn omics-specific representations in reduced dimensions. Then, the model minimizes the distance between cells with the same identity (i.e., cell type) while maximizing the distance between cells with unmatched identities through a contrastive loss function. We have tested the integration ability of our model on three real-world datasets. These datasets contain genes, chromatin accessibility, and protein features profiled simultaneously from the same set of cells. The results demonstrate that sCIN outperforms some state-of-the-art methods that use various strategies in the integration task.

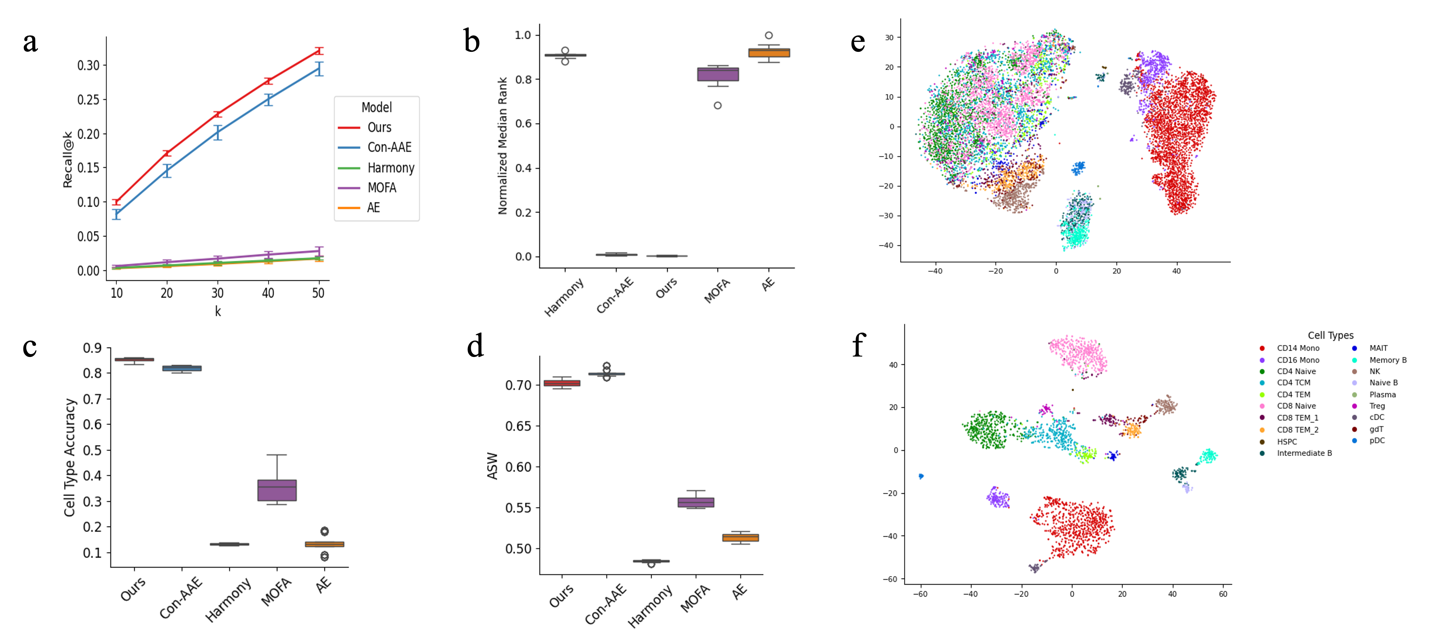


**Fig. 1-** **sCIN workflow.** The framework uses modality-specific encoders to extract latent representations for each modality. A contrastive loss function is applied to align these representations, capturing relationships between biological covariates. Performance is evaluated using hold-out integrated embeddings.

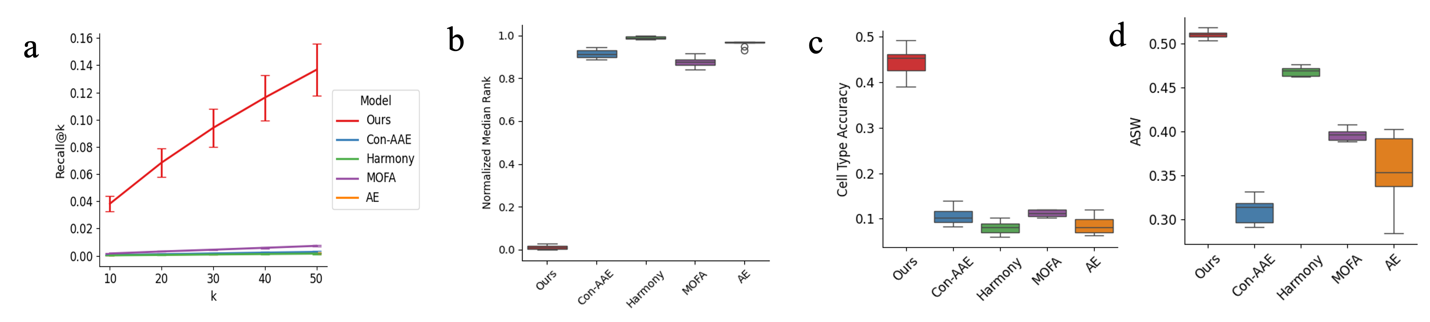
# Results



**Fig. 2- a)** The comparison of Recall@k metrics between different models for K = 10, 20, 30, 40, and 50 for the SHARE-seq dataset. Our method outperforms all values of K. **b)** It shows the mean of the Median Rank metric for different models for the SHARE-seq dataset(Gene expression and chromatin accessibility modalities). **c)** Cell Type Accuracies between models that show the percentage of closest embeddings from different modalities having the same cell types. **d)** Comparison of ASW score across models shows the clusters’ quality.  **e)** t-SNE embeddings of the original PBMC dataset. **f)** t-SNE embeddings of the integrated datasets.



**Fig. 3- a)**The comparison of Recall@k metrics between different models for K = 10, 20, 30, 40, and 50 for the 10X PBMC Multiome dataset. Our method outperforms all values of K. **b)** It shows the mean of the Median Rank metric for different models for the PBMC dataset(Gene expression and chromatin accessibility modalities). **c)** Cell Type Accuracies between models that show the percentage of closest embeddings from different modalities having the same cell types. **d)** Comparison of ASW score across models shows the clusters’ quality. **e)** t-SNE embeddings of the original PBMC dataset. **f)** t-SNE embeddings of the integrated datasets.



**Fig. 4-** **a)** The comparison of Recall@k metrics between different models for K = 10, 20, 30, 40, and 50 for the CITE-seq dataset. Our method outperforms all values of K. **b)** It shows the mean of the Median Rank metric for different models for the CITE dataset(Gene expression and ADT modalities). **c)** Cell Type Accuracies between models that show the percentage of closest embeddings from different modalities having the same cell types. **d)** Comparison of ASW score across models shows the clusters’ quality.

Unpaired results

## Methods

The preprocessing steps for RNA-seq, ATAC-seq, and ADT (Antibody-Derived Tags) count matrices followed a standardized approach using the Scanpy pipeline [30] pipeline. For RNA-seq data, the count matrix was normalized by total counts across all genes to ensure that all cells had the same total counts. A log(. + 1) transformation was then applied, followed by scaling to achieve zero mean and unit variance. Batch correction was performed using the ComBat method [31] implemented in Scanpy (scanpy.pp.combat; Pederson 2012). The same procedure was applied to the scATAC-seq accessibility matrix, assuming it contained raw counts. For ADT data, preprocessing also adhered to the Scanpy pipeline, involving normalization, log transformation, and scaling.

# Our framework

Our method employs contrastive learning [[32,33]](https://paperpile.com/c/utNs0T/xgRs+Bkwb) as one of the most powerful representation learning algorithms to find an integrated representation of two different modalities of data. We used two encoders to generate two modality-specific embeddings. The architectures of the encoders are the same, consisting of three layers. The first linear layer maps the input data into a hidden dimension (256 for SHARE-seq and 10x datasets and 64 for CITE-seq dataset) followed by a batch normalization layer. The ReLU function activates the last linear layer to capture non-linear relationships between features, resulting in latent embeddings(128 for SHARE-seq and 10x datasets and 16 for CITE-seq dataset).

To preserve cell type heterogeneity while integrating technological batches, we tailored the contrastive learning algorithm to consider the paired cells from both modalities as positive pairs while all other cells with different cell types represent negative pairs. The algorithm's goal is to minimize the distance between positive pairs while keeping negative pairs far apart.