Contrastive Cycle Adversarial Autoencoders for Single-cell Multi-omics Alignment and Integration

Xuesong Wang[1,4], Zhihang Hu[1], Tingyang Yu[2,3], Ruijie Wang[1], Yumeng Wei[1], Juan Shu[5], Jianzhu Ma[6], and Yu Li [*1,4]

[1]Department of Computer Science and Engineering, CUHK, Hong Kong SAR, China
[2]Department of Mathematics, CUHK, Hong Kong SAR, China
[3]Department of Information Engineering, CUHK, Hong Kong SAR, China
[4]The CUHK Shenzhen Research Institute, Hi-Tech Park, Nanshan, Shenzhen, 518057, China
[5]Purdue University, West Lafayette, IN 47907, United States
[6]Institute for Artificial Intelligence, Peking University, Beijing, 100871, China

Abstract

Muilti-modality data are ubiquitous in biology, especially that we have entered the multi-omics era, when we can measure the same biological object (cell) from different aspects (omics) to provide a more comprehensive insight into the cellular system. When dealing with such multi-omics data, the first step is to determine the correspondence among different modalities. In other words, we should match data from different spaces corresponding to the same object. This problem is particularly challenging in the single-cell multi-omics scenario because such data are very sparse with extremely high dimensions. Secondly, matched single-cell multi-omics data are rare and hard to collect. Furthermore, due to the limitations of the experimental environment, the data are usually highly noisy. To promote the single-cell multi-omics research, we overcome the above challenges, proposing a novel framework to align and integrate single-cell RNA-seq data and single-cell ATAC-seq data. Our approach can efficiently map the above data with high sparsity and noise from different spaces to a low-dimensional manifold in a unified space, making the downstream alignment and integration straightforward. Compared with the other state-of-the-art methods, our method performs better in both simulated and real single-cell data. The proposed method is helpful for the single-cell multi-omics research. The improvement for integration on the simulated data is significant.

Keywords: Single-cell multi-omics, Multi-omics integration, Cycle autoencoders, Contrastive learning, scRNA-seq and scATAC-seq

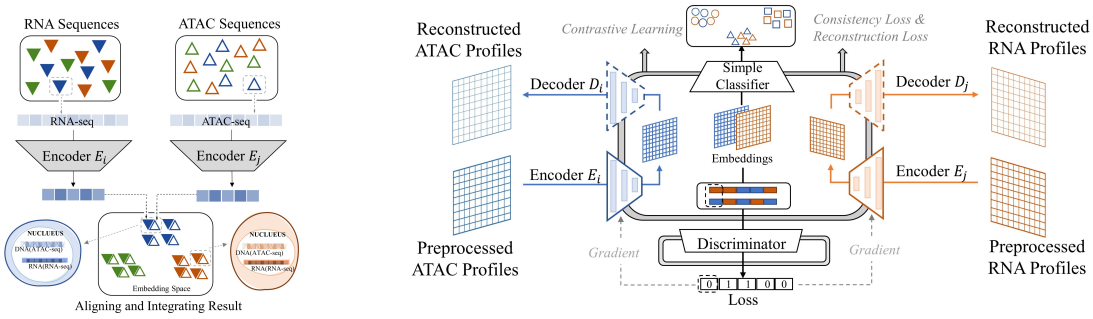
*Corresponding Author. Email: liyu@cse.cuhk.edu.hk

1 Introduction

Multi-modality data, which describe an object from different perspectives, help us understand the world more comprehensively [1]. Such data are typical and popular in biology, which are referred as multi-omics data. That is, we can measure a cell from different aspects (omics) to provide a more comprehensive insight into the cellular system. To achieve that, we should obtain multi-omics data related to the same cell, which is not a trivial task. Although multi-omics profiling approaches for the same set of single cells have become available [2], such as single-cell RNA sequencing (scRNA-seq) and single-cell Assay for Transposase Accessible Chromatin sequencing (scATAC-seq), which describe the same cell from different perspectives, the experiments are usually done at different time points for a set of cells. Consequently, we have the multi-omics data for a group of cells, but the correspondence between different modalities for a single cell is missing (Figure 1a). More specifically, we want to obtain the high-throughput paired multi-omics data for every single cell, which is referred as *alignment*. On the other hand, even for data within the same modality, the data distribution can be inconsistent because of the subtle differences in measurement processes [3], such as measurement time or equipment used, which is referred as *batch effects*. They should also be considered when we are studying the correspondence among different modalities. Considering such batch effects, we need to integrate different multi-omics data from the same batch, which is called *integration*. The above tasks of alignment and integration are very useful and interesting, but difficult, considering the distribution shifting within and across different modalities and the sparsity and high dimension of the single-cell data [4, 5, 6].

Some computational methods have been proposed to deal with this important but difficult problem, aligning and integrating data from different omics. However, the most popular works in the single-cell field focus on integrating datasets of the same modality [7, 8, 9], which are not suitable for different modalities. People usually integrate and align multi-omics data in the learned low-dimensional embedding space using dimension reduction techniques, such as Principal Component Analysis (PCA) [10, 11] and nonlinear successors of the classic Canonical Correlation Analysis (CCA) [12]. The typical examples are Seurat [9] and Deep Classic Canonical Correlation (DCCA) [13]. Seurat relies on the linear mapping of PCA and aligns the manifold vector based on linear methods Mutual Nearest Neighbors (MNNs) and CCA, which weaken its ability to handle nonlinear geometrical transformations across cellular modalities [14]. DCCA can be effective for nonlinear transformation benefiting from deep learning, but according to the results of our experiments, it is not robust enough when the signal-to-noise ratio (SNR) is low. We also tried Maximum Mean Discrepancy (MMD) [15] replacing CCA in the embedding space, but the performance is also not good enough. Several methods requiring no correspondence information were derived under various advanced machine learning techniques, such as Pamona [14], MATCHER [16], MMD-MA [17], UnionCom [18], SCOT [19]. Although these methods are unsupervised and achieve integrative performance with encouraging results [19], there are still other additional conditions required. For examples, MMD-MA and UnionCom, Pamona need the user specify several hyperparameters, while MATCHER and SCOT require the assumption that all datasets share the same underlying structure across cellular modalities [14]. Selecting hyperparameter values can be difficult and may need prior information and such assumption can be ineffective confronting dataset-specific cell types/structures across the single-cell datasets [14]. Deep learning methods are promising to provide alignment and transfer learning between datasets [20, 21]. Deep generative models, such as cycleGANs [22], MAGAN [23], RadialGAN [24] and starGAN [25], are used to learn a nonlinear mapping from one domain to another in unsupervised way. But the above transitions are almost within the same modality and can be disturbed by noise or sparsity in the data [12]. The scenario of multi-omics translation and alignment is much more complicated. Some other works indeed propose methods to align multi-omics data with multiple autoencoders [26, 27, 28]. However, such methods also can be seriously affected by noise or sparsity, which is a fundamental characteristic of single-cell data.

In general, there are four major challenges in multi-omics alignment. Firstly, although we have a large amount of unaligned multi-omics data, the aligned data is very scarce. Secondly, the single-cell omics data is of very high dimension and extremely sparse. For example, the unprocessed scRNA-seq data contain around 25,000 genes (features) while, in the data matrix, 40% to 80% of the values are 0. Furthermore, as we have discussed, the data are highly noisy [4, 5, 6]. Finally, although both scRNA-seq data and scATAC-seq data describe the cell status, they contain very different information, and the mapping between the two kinds of data is highly complicated. To promote the single-cell multi-omics data analysis, we propose a framework based on Contrastive cycle adversarial Autoencoders (Con-AAE), which can resolve the above challenges precisely (Figure 1b). Con-AAE uses two autoencoders to map the two modal data into two low-dimensional manifolds, forcing the two spaces as unified as possible. We use two novel loss terms to achieve that. The first term is called adversarial loss. That is, we combine GAN with autoencoders, forcing the two autoencoders to produce unified embedding to deceive the discriminator, which is designed to distinguish whether two embedding factors are from the same modality or not. However, only



(a) Problem setting

(b) Framework

Figure 1: **(a)** scRNA-seq and scATAT-seq are two sequencing technologies, which measure different aspects of the same cell. The two experiments are usually done at different time points. We aim at identifying the correspondence between the two kinds of data from the same set of cells. **(b)** The Con-AAE framework uses two autoencoders to map the two kinds of sequence data into two low dimensional manifolds, forcing the two spaces as unified as possible with the adversarial loss and latent cycle-consistency loss. For the alignment task, we train the models without pairwise information but consider the data noise explicitly by utilizing self-supervised contrastive learning. For the integration task, we train the framework with labeled data.

using the adversarial loss may lead to model collapse. To avoid the problem, we further propose a novel latent cycle-consistency loss. In a nutshell, the embedding produced by the scRNA-seq encoder will go through the scATAC-seq decoder and encoder to produce another cycled embedding, for which we can check the consistency between the original embedding the cycled embedding. In addition to the above two loss terms, for the alignment task, we train the models without pairwise information but consider the data noise explicitly by taking advantage of self-supervised contrastive learning. For the integration task, we train the framework with labeled data. We extensively perform experiments on two real-world datasets and a group of simulated datasets. The comprehensive experiments on both the simulated datasets and real-world datasets show that our method has better performance and is more robust than the other state-of-art methods.

2 Methods

In this section, We give our framework in details below with Fig. 1b which illustrates the whole pipeline. To start with, we formalize the alignment problem as,

$$\forall (r, a) \in \mathscr{T}, f(a) = h(r). \tag{1}$$

We denote (r,a) as a pair of RNA-seq and scATAC-seq taken from the same cell. We would like to find two mappings f and h such that for any aligned {r,a} pairs in $\mathscr{T}$, f would maps the scRNA-seq profile and scATAC-seq profile to a shared embedding space. Due to the limitations of available real-world aligned data, we are actually working on an unsupervised problem and the results are evaluated on a few available aligned pairs.

The integration problem could be justified as a classification problem, and the objective results to finding the corresponding batch of each scRNA-seq or scATAC-seq profile. The groundtruth labels are available, and, therefore, it can be trained in a supervised way. For $\forall x \in$ {scRNA-seq}$\cup$ {scATAC-seq}, we want to train a classifier g such that

$$g(x) = label(x). \tag{2}$$

Our main model is built upon the framework of Adversarial Auto-Encoders [29] specialized for modality task, integrating our novel embedding consistency module and contrastive training process [30]. The intuition behind is that multi-omics from a single-cell data should obtain commonality, and their mappings could live in a unified low-dimensional manifold, which therefore makes alignment and integration task more accurate.

2.1 Adversarial Auto-Encoders

The usage of adversarial auto-encoders aims to map different omics into a unified latent manifold while able to reconstruct these different aspects. Therefore, as shown in Figure 1b, we are using a coupled set of encoders $\{E_i, E_j\}$[31] to map {scATACs-eq,scRNA-seq} into manifolds $\{Z_i, Z_j\}$, and decoders $\{D_i, D_j\}$ could decode the embedded manifolds back to the original distribution. The reconstruction loss is defined as follows,

$$\begin{aligned} L_{recon} &= \mathbb{E}_{x\sim p_{rna}} d(x, D_j(E_j(x))) \\ &+ \mathbb{E}_{x\sim p_{atac}} d(x, D_i(E_i(x))), \end{aligned} \tag{3}$$

whereas d stands for indicated distance in the embedding space. Discriminator $\mathscr{D}$ tries to align these embedded manifolds and works in the sense that input $x \in Z_j$, $\mathscr{D}(x) = 1$ or $x \in Z_i$, $\mathscr{D}(x) = 0$.

$$\begin{aligned} L_{adv} &= \mathbb{E}_{x\sim p_{rna}}[\log \mathscr{D}(E_j(x))] \\ &+ \mathbb{E}_{x\sim p_{atac}}[\log(1 - \mathscr{D}(E_i(x)))]. \end{aligned} \tag{4}$$

The above losses L_{recon} and L_{adv} are trained together with the same weights.

2.2 Latent Cycle-Consistency Loss

The backbone framework enforces the embedding manifolds to gradually align. However, a critical problem underlying is that since RNA and ATAC data are sparse in a high dimensional domain, the training procedure above only aligns and trains on those regions where the data exist.

For instance, if a region A in the embedding space around $E_j(x'), x' \in$\{scRNA-seq\} does not involve any existing $E_i(x), x \in$\{scATAC-seq\}. Then, neither the decoder D_i nor the encoder E_i is trained on A, thus they would not compute in a "reverse" mapping way, and the result of $E_i D_i E_j(x)$ would be unreasonable or may not lie on the aligned manifold. This critical problem causes the difficulty of inferring from scRNA-seq profile to scATAC-seq profile directly.

Therefore, we introduce a latent consistency loss shown in Figure 2a [22][32] to resolve this problem,

$$\begin{aligned} L_{cyc} &= \mathbb{E}_{x\sim p_{rna}}[d(E_j(x), E_i D_i E_j(x))] \\ &+ \mathbb{E}_{x\sim p_{atac}}[d(E_i(x), E_j D_j E_i(x))]. \end{aligned} \tag{5}$$

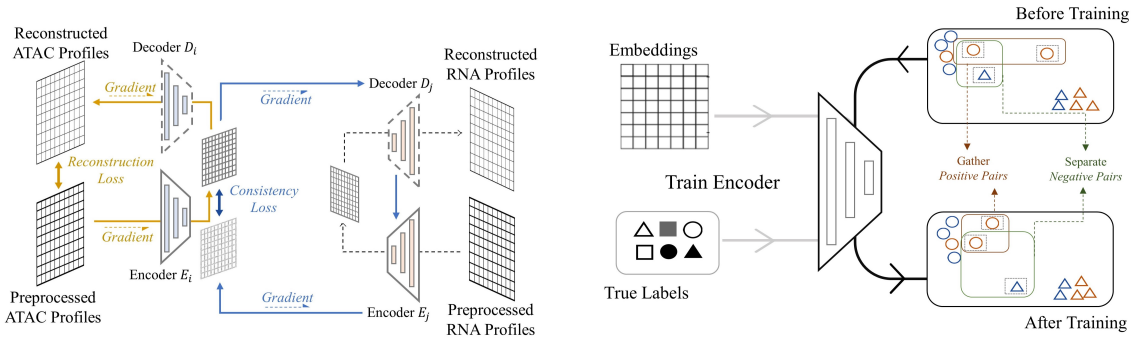
L_{cyc} aims to train the set of encoder-decoder on the domain of where different omics data may not exists, which enforces the smoothness and consistency on those regions. In this way, we could compare the embedding of $E_j(x), x \in$\{scATAC-seq\} directly with the existing scRNA-seq embedding around it.

2.3 Supervised Contrastive Loss

The above framework works in an unsupervised manner such that the embedded latent manifolds of multi-omics align properly. A simple classifier trained on our latent space could achieve rather good accuracy on the integration task.

We could further improve our work on both tasks taking advantage of the ground-truth cell type labels. The cell type label could refer to biological cell types, or the label of data batches collected from different time or platform. Following the idea of contrastive learning [33, 34], we employ a contrastive loss in embedding space to the training procedure. It enforces smaller In-Batch distance and larger Between-Batch distance. In-Batch refers to different modalities data collected from the same batch, vice versa. We equally treat both modalities in contrastive training, which benefits the alignment task in the sense that multi-omics of the same single-cell data should obviously belong to the same batch. We show that lowering the In-Batch distance indeed improves the alignment accuracy in the below ablation studies. On the other hand, contrastive training benefits integration task by enabling the decision boundary smoother and more robust.

In practice, we first encode data from two modalities to the embedding space. Define the embedding by $z \in \mathbb{Z}$. Given z^a as anchor vector in latent space, we select an z^p such that $argmax_{z^p}\{d(z^a, z^p)\}, label\{z^a\} = label\{z^p\}$, which is named hard positive. The intuition of hard positive is to find a vector furthest from the



(a) Illustration of latent cycle-consistency loss.

(b) Illustration of latent contrastive loss.

Figure 2: **(a)** The embedding produced by the first encoder will go through the second decoder and encoder to produce another cycled embedding, for which we can check the consistency between the original embedding the cycled embedding. **(b)** The contrastive loss minimizes the distance between positive pairs and maximizes the distance between negative pairs. This loss makes our method more robust to noise.

anchor within same cluster. Similarly, we have z^n as hard negative such that $argmin_{z^n}\{d(z^a, z^n)\}, label\{z^a\} \neq label\{z^p\}$. z^n is defined as the closest vector that from a different cluster. The objective immediately follows,

$$d(z^a, z^p) + \alpha < d(z^a, z^n), \forall (z^a, z^p, z^n) \in \mathbb{Z}. \tag{6}$$

Above, α is the margin defined accordingly by us. Thus, by the contrastive loss, we tend to optimize,

$$\begin{aligned} L_{con} &= \mathbb{E}_{z^a \sim Z}[d(z^a, z^p) - d(z^a, z^n) + \alpha]. \\ L_{con} &= \mathbb{E}_{x \sim \{RNA\}}[d(E_j(x), z^p) - d(E_j(x), z^n) + \alpha] \\ &\quad + E_{x \sim \{ATAC\}}[d(E_i(x), z^p) - d(E_i(x), z^n) + \alpha]. \end{aligned} \tag{7}$$

Figure 2b shows after training, instances within the same batch are pushed towards each other, and those from the different batches are pushed away. Thus, the decision boundary of the labels tends to be smoother and robust, which also benefits the alignment task.

2.4 Training Procedure

In the above sections, we proposed several losses related to different objectives. Following the training procedure of Generative Adversarial Nets [35], we adopt a two-stage training scheme where L_{adv} and $L_{recon}, L_{cyc}, L_{con}$ are trained separately as the pseudo-code in Algorithm 1.

Algorithm 1 Training Procedure

```
while numbers of training iterations do
    while k_1 steps do
        sample mini-batch {x_1, x_2, ..., x_m} from {scRNA-seq}
        sample mini-batch {y_1, y_2, ..., y_m} from {scATAC-seq}
        Search positives and negatives z^a, z^p for each x_1, ..., y_m.
        Update E_i, E_j, D_i, D_j by descending its stochastic gradient (1/m) ∇L_recon + L_cyc + L_con − L_adv
    end while
    while k_2 steps do
        sample mini-batch {x_1, x_2, ..., x_m} from {scRNA-seq}
        sample mini-batch {y_1, y_2, ..., y_m} from {scATAC-seq}
        Update Discriminator 𝒟 by descending its stochastic gradient (1/2m) ∇L_adv
    end while
end while
```

In this way, the Discriminator $\mathscr{D}$ competes against the encoder-decoder E_i, E_j, D_i, D_j until the training ends and reaches the equilibrium.

3 Experimental Setup

3.0.1 Real-world Dataset.

We use two sets of single-cell multi-omics data generated by co-assays. The first dataset is generated using the sci-CAR assay [3]. For the single-cell ATAC-seq data, we download the processed data from [26], which was computed as described in [3], resulting in a matrix of 1791×815. For the single-cell RNA-seq data, we pick the genes with $q-value > 0.05$ from the genes being differentially expressed [3], which forms a 1791×2613 matrix. Such paried data was collected from human lung adenocarcinoma-derived A549 cells corresponding to 0-, 1-, or 3-hour treatment with DEX, so we have three batches here and the batch label information is available.

We denote the results of SNAREseq [36] assay as the second datasets, which also consists of chromatin accessibility and gene expression. The data was collected from a mixture of human cell lines: BJ, H1, K562, and GM12878. We reduce the dimension of the data by PCA. The resulting matrix for ATAC is of size 1047×1000 and 1047×500 for gene matrix. Annotation information for BJ, H1, K562, and GM12878 is generated by the code provided by author, so labels of four batches is available.

3.0.2 Simulated Datasets.

We simulate several datasets of different sizes, which contain 1200, 2100, 3000, and 6000 cells, respectively. For the single-cell RNA-seq data, we utilize three Gaussian distributions with different parameters to generate three batches, and the feature dimension is 1000. For the single-cell ATAC-seq data, we train an autoencoder with the simulated RNA-seq data and map them to 500 dimensions as ATAC-seq data. After that, we randomly set around 40% of features to 0 for RNA-seq data since the real-world RNA-seq data matrix is very sparse. Considering the inevitable mismatch in the experiment, we also randomly set around 10% mismatches in the datasets and shuffle all the pairs. Furthermore, we add noise to them with the SNR equal to 5, 10, 15, 20, and 25, along with the version without noise. Then, we have 24 simulated datasets here, and their parameters are close to those of the real-world dataset. Surely, the real-world multi-omics are more complicated than the simulated data, but the experimental results show that our method is sufficient to distinguish the performance of different methods.

3.0.3 Evaluation criteria.

We utilize two existing manners [26] to evaluate integration and alignment, respectively. (a) The fraction of cells whose batch assignment is predicted correctly based on the latent space embedding. (b) $recall@k$, *i.e.*, the proportion of cells whose true match is within the k closest samples in the embedding space (in ℓ_1-distance) or in the original space for methods that do not generate an embedding space. (a) is for integration, while (b) is for alignment.

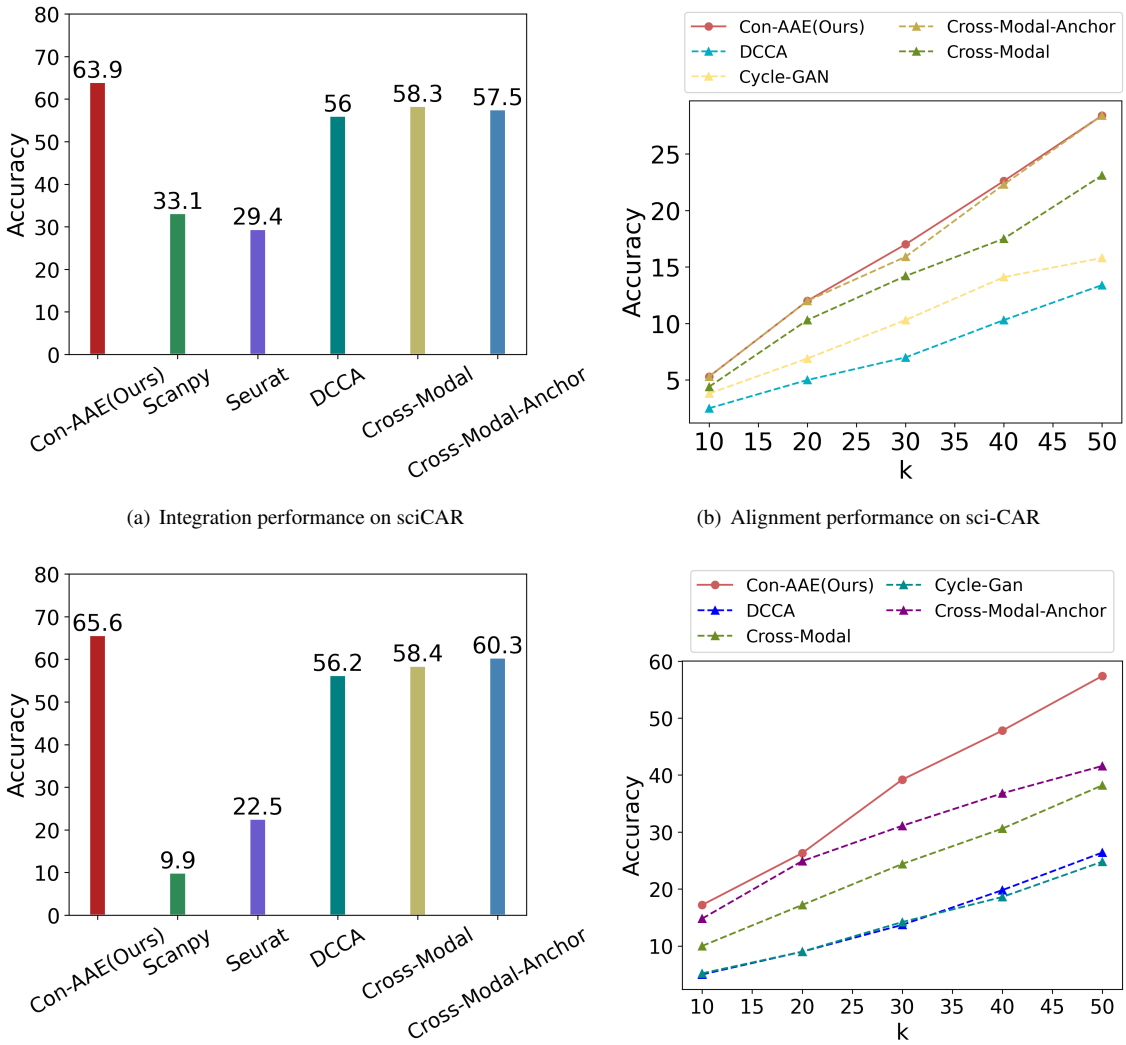
4 Results

4.1 Compared with SOTA.

Instead of assuming all datasets share the same underlying structure or specifying parts of hyperparameters like some traditional machine learning mathods [19, 16, 18, 17, 14, 9], we obtain more information from datasets with partial correspondence information (batch label or cell types label). We selected several state-of-art methods based on deep learning like us, including cross-modal [26], cross-modal-anchor (pairwise information added), DCCA [13], cycle-GAN [22]. Moreover, we also compared our method with two classic machine learning methods of integration, which are Scanpy[37] and Seurat[9]. These two methods are under the assumption that all datasets have the same features, so we applied PCA to make datasets have same features before using them. We applied Con-AAE and these methods on the simulated datasets and the real-world dataset, respectively.

4.1.1 On Simulated Datasets.

The extensive experimental results are shown in Figure 4. As shown in left part of Figure 4, Con-AAE performs better than all the other methods in most cases on the integration task, regardless of data size and SNR. Regarding



(a) Integration performance on sciCAR

(b) Alignment performance on sci-CAR

(c) Integration performance on SNAREseq

(d) Alignment performance on SNAREseq

Figure 3: Con-AAE compared with SOTA methods on the real-world datasets. Con-AAE has the best performance on both criteria. Note that the identification of cell pairwise correspondences between single cells is termed "anchor"[9]. Cross-modal-anchor indicates that "anchor" information is provided when training Cross-modal.

alignment, we evaluate different methods using $recall@k$, whose results are shown in right part of Figure 4. Again, Con-AAE is consistently better than the other competing methods. Notice that our method's performance is very consistent against different data sizes and noise levels, while the other methods may perform well on some settings but badly for the other settings. The results indicate that Con-AAE is robust and stable enough to have the potential to handle the complicated single-cell multi-omics alignment and integration problems with a low SNR ratio.

4.1.2 On Real-world Dataset.

Eventually, we care about the methods' performance on the real-world dataset the most, although the real-world dataset with ground-truth information is very limited. Still, Con-AAE shows superior performance. On the sci-CAR datasets. Con-AAE outperforms the other methods by up to 34.5% on the integration task, as shown in Figure 3a. For alignment, Con-AAE always has better performance than all the other methods no matter what k is (Figure 3c). On the SNAREseq datasets, more obviously, Con-AAE also has dominant performance on each evaluation metric. The improvement on the integration task is up to 55.7% (Figure 3b). On the other hand, the performance on $recall@k$ is better than others no matter what k is (Figure 3d).

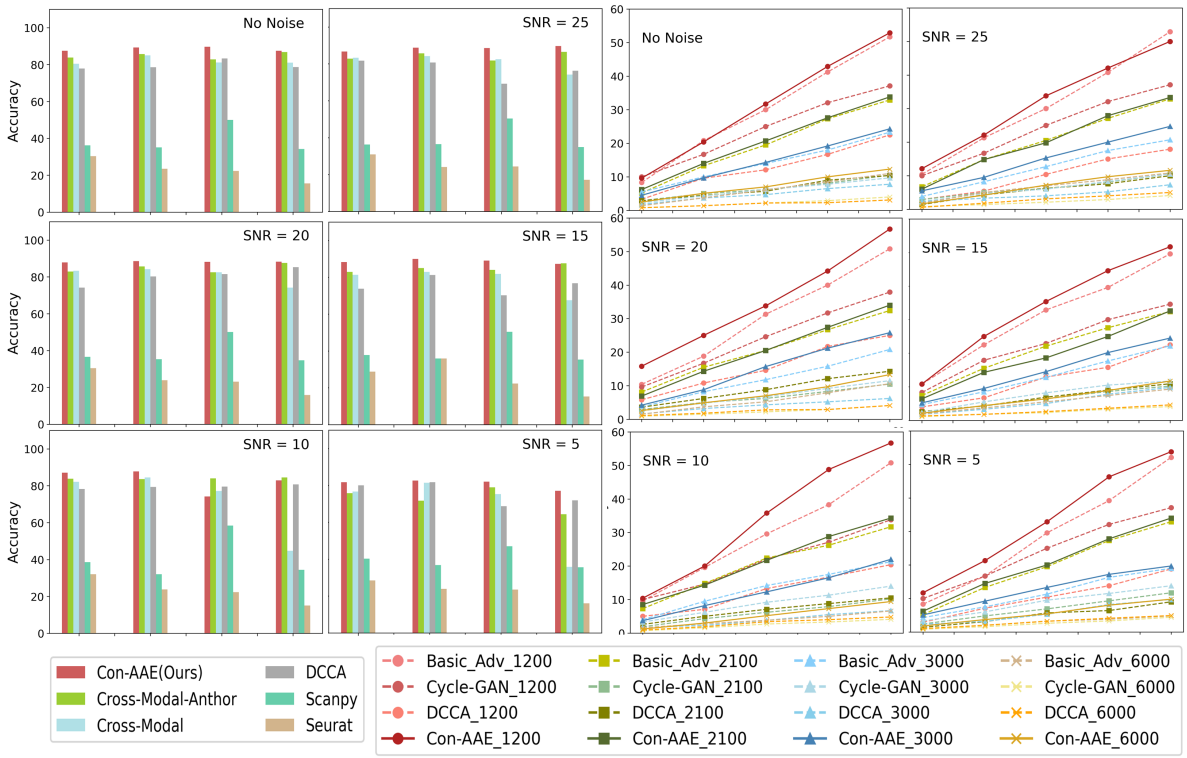


Figure 4: Con-AAE compared with SOTA methods on simulated datasets. Con-AAE almost always has the best performance with the change of data size and SNR, while the other methods struggle with SNR or data size changing. Because there is no embedding space in Cycle-GAN, we are unable to evaluate it on integration. cross-modal-anchor is a supervised method (pairwise information provided), so we do not compare with it on alignment.

4.2 Ablation Studies

We perform comprehensive ablation studies on the Sci-CAR dataset, and the results shows the effectiveness of different components.

There are three parts in Table 1. The first part indicates there is no adversarial loss in embedding space. The second part indicates an MMD loss [38] instead of an adversarial loss. And the last part indicates whether there is an adversarial loss in the embedding space. Clearly, all the items in the third part are better than the corresponding items of the other two parts, demonstrating that the adversarial loss works better than MMD loss on this problem.

There are five items in each part of Table 1. The first one is the basic framework, as mentioned before. The anchor one means pairwise information provided, which indicates that it is a supervised learning problem instead of an unsupervised one. cyc and contra denote latent cycle-consistency loss and contrastive loss, respectively. As shown in the table, the addition of cyc and contra improves the model to some extent. Apparently, Con-AAE has the best performance. Latent cycle-consistency loss and contrastive loss alone can improve the performance to some degree, but Con-AEE is more robust and has better scalability.

Impressively, Con-AAE has better performance even compared to some supervised methods with the pairwise information provided. Within Table 1, we compare our method with the anchor methods. For such methods, we train the basic coupled autoencoders using the pairwise information as the supervision. For Con-AAE, we still perform unsupervised learning using cycle-consistency loss and contrastive loss. Even without the supervised information, Con-AAE can still outperform the basic supervised anchor methods consistently on both accuracy and $recall@k$. It suggests that cycle-consistency loss and contrastive loss can force our method to learn a unified latent space for the two kinds of single-cell omics data, making the alignment and integration much easier. We also tried to combine Con-AAE with the pairwise information. The supervised information can help our method further, but the degree is very slight. We suppose that in the real data, the pairwise information may contain noise, which is common in the single-cell field. Because of the contrastive loss, which makes Con-AAE a robust method, such weak supervision does not help our method too much.

Table 1: Ablation Study of different components in Con-AAE and Comparison with other methods. Basic refers to Coupled AEs plus Simple classifier; Adv refers to adversarial loss; mmd refers to mmd loss; anchor refers to pairwise information added; cyc refers to latent cycle-consistency loss; contra refers to contrastive loss.

Method	Integration ACC	Recall@k k=10	Recall@k k=20	Recall@k k=30	Recall@k k=40	Recall@k k=50
Basic	56.1	3.9	9.2	13.4	18.1	20.6
Basic_anchor	60.8	5.3	12	15.9	22.3	28.4
Basic_cyc	56.7	6.7	11.7	15.6	19.8	24.5
Basic_contra	58.9	6.9	12.5	15.6	21.7	27.3
Basic_contra_cyc	62.2	7.8	10.8	15.9	21.2	25.9
Basic_mmd	57.5	4.7	9.4	12.5	17.3	21.2
Basic_mmd_anchor	56.4	5.3	11.4	16.7	21.2	24.3
Basic_mmd_cyc	60.8	3.9	10	16.2	21.2	25.1
Basic_mmd_contra	57.5	5.8	12.8	15.6	24	27
Basic_mmd_contra_cyc	57.8	4.5	9.5	13.7	19.8	24.8
Basic_adv	58.3	4.4	10.3	14.2	17.5	23.1
Basic_adv_anchor	61.7	5	10.3	15.9	21.2	26.5
Basic_adv_cyc	58.7	5	10.3	14.2	19.8	24.3
Basic_adv_contra	60.61	4.4	9.4	13.6	18.4	25.4
Con-AAE	**63.9**	**5.3**	**12**	**17**	**22.6**	**28.4**

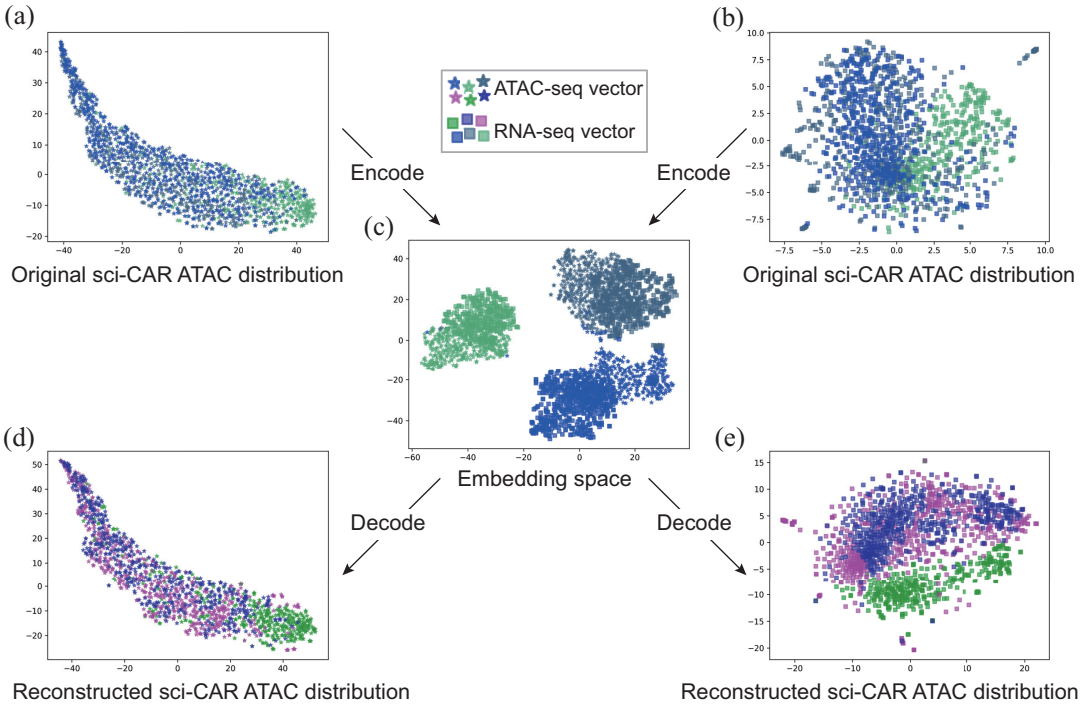


Figure 5: Visualization of the sci-CAR data in the original spaces, shared embedding space, and the reconstructed spaces. The star represents the scATAC-seq data while the square means the scRNA-seq data. Different colors refer to different batch labels, corresponding to 0-, 1-, 3- hours treatment with DEX. For easy viewing, we used different colors when reconstructing data. The distribution of reconstructed data is still distinguishable.

4.3 Visualization

To visualize the integration and alignment, we utilize t-SNE to project the data from the embedding space to the 2D space. As we can see in Figure 5 and 6, for the two real-world datasets, we projected the scATAC-seq data and scRNA-seq data to a shared embedding space. And within the space, the different modality (with different shapes) data from the same batch (with the same color) form into clusters, suggesting that our method indeed learned a latent space where the data from different omics can be integrated easily. Furthermore, we used the decoder of each side to translate the embedding vector back to the original space. The resulted data distribution is quite close to the original distribution, which indicates that our framework indeed learned the underlying features of the data

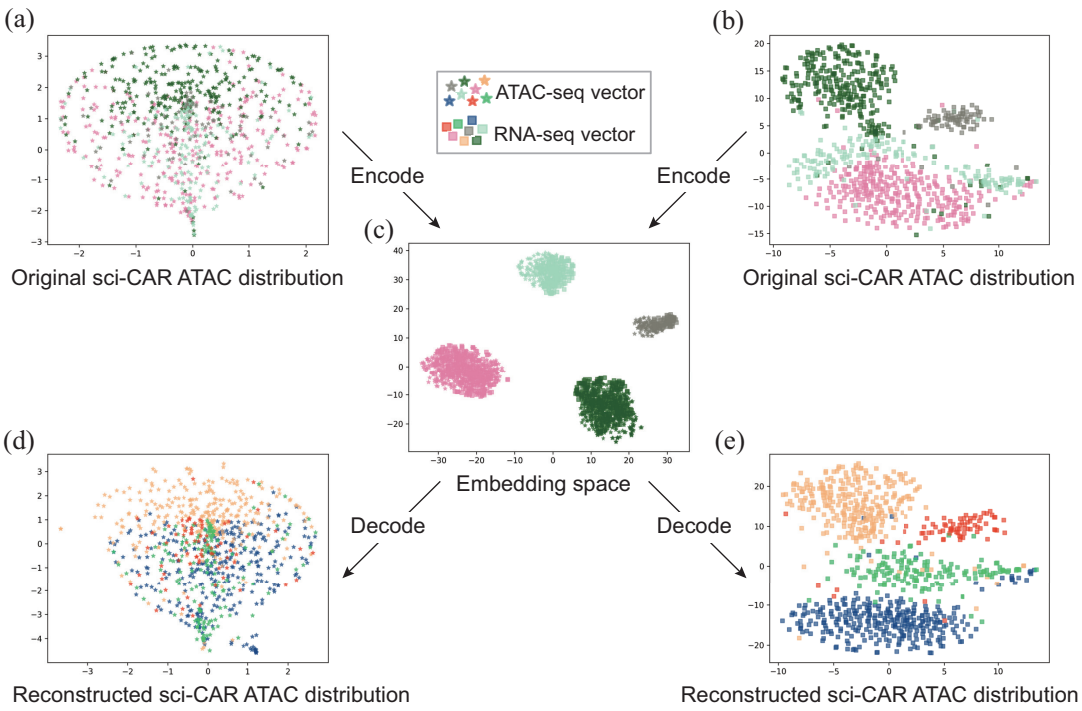


Figure 6: Visualization of the SNAREseq data in the original spaces, shared embedding space, and the reconstructed spaces. The star represents the scATAC-seq data while square means the scRNA-seq data. Different colors refer to the H1, BJ, K562, and GM12878 human cells. We can see that the scATAC-seq data and scRNA-seq data from the same cell type are tightly combined in the embedding space, And also, we utilize different colors representing reconstructed data, which also has an impressive effect.

and removed the redundant features effectively in the process of encoding.

5 Discussion

In this paper, we propose a novel framework, Con-AAE, aiming at integrating and aligning the multi-omics data at the single-cell level. On the one hand, our proposed method can map different modalities into the embedding spaces and overlap these two distributions with the help of an adversarial loss and a novel latent cycle-consistency loss. On the other hand, we apply a novel self-supervised contrastive loss in the embedding space to improve the robustness and scalability of the entire framework. Comprehensive experimental results on both the simulated datasets and the real datasets show that the proposed framework can outperform the other state-of-the-art methods for both the alignment and the integration tasks. Detailed ablation studies also dissect and demonstrate the effectiveness of each component in the framework. Our method will be helpful for both the single-cell multi-omics research and the general multi-modality learning tasks in computational biology.

For future work, we aim to extend our work from a two-domain task to a multiple-domain study, allowing it to integrate and align multiple modalities. Besides integration and alignment between sequence modalities, we intend to perform our method on different kinds of data, including but not limited to images, geometrical spatial structure, *etc*. Obviously, it is very interesting to investigate the spatial transcriptomics data. We will also develop methods for translating modalities. By doing so, we hope to build a system that could be utilized in medical, biological, and other fields.

6 Acknowledgements

This work was supported by a grant by The CUHK Shenzhen Research Institute. Thanks to Yixuan Wang for providing partial exquisite images.

References

[1] Baltrušaitis, T., Ahuja, C. & Morency, L.-P. Multimodal machine learning: A survey and taxonomy. *IEEE transactions on pattern analysis and machine intelligence* **41**, 423–443 (2018).

[2] Gala, R. *et al.* Consistent cross-modal identification of cortical neurons with coupled autoencoders. *Nature Computational Science* **1**, 120–127 (2021).

[3] Cao, J. *et al.* Joint profiling of chromatin accessibility and gene expression in thousands of single cells. *Science* **361**, 1380–1385 (2018).

[4] Klein, A. M. *et al.* Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. *Cell* **161**, 1187–1201 (2015).

[5] Macosko, E. Z. *et al.* Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell* **161**, 1202–1214 (2015).

[6] Buenrostro, J. D. *et al.* Single-cell chromatin accessibility reveals principles of regulatory variation. *Nature* **523**, 486–490 (2015).

[7] Butler, A., Hoffman, P., Smibert, P., Papalexi, E. & Satija, R. Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nature biotechnology* **36**, 411–420 (2018).

[8] Trong, T. N. *et al.* Semisupervised generative autoencoder for single-cell data. *Journal of Computational Biology* **27**, 1190–1203 (2020).

[9] Stuart, T. *et al.* Comprehensive integration of single-cell data. *Cell* **177**, 1888–1902 (2019).

[10] Bersanelli, M. *et al.* Methods for the integration of multi-omics data: mathematical aspects. *BMC bioinformatics* **17**, 167–177 (2016).

[11] Argelaguet, R. *et al.* Multi-omics factor analysis—a framework for unsupervised integration of multi-omics data sets. *Molecular systems biology* **14**, e8124 (2018).

[12] Stanley III, J. S., Gigante, S., Wolf, G. & Krishnaswamy, S. Harmonic alignment. In *Proceedings of the 2020 SIAM International Conference on Data Mining*, 316–324 (SIAM, 2020).

[13] Andrew, G., Arora, R., Bilmes, J. & Livescu, K. Deep canonical correlation analysis. In *International conference on machine learning*, 1247–1255 (PMLR, 2013).

[14] Cao, K., Hong, Y. & Wan, L. Manifold alignment for heterogeneous single-cell multi-omics data integration using pamona. *bioRxiv* (2020).

[15] Borgwardt, K. M. *et al.* Integrating structured biological data by kernel maximum mean discrepancy. *Bioinformatics* **22**, e49–e57 (2006).

[16] Welch, J. D., Hartemink, A. J. & Prins, J. F. Matcher: manifold alignment reveals correspondence between single cell transcriptome and epigenome dynamics. *Genome biology* **18**, 1–19 (2017).

[17] Singh, R. *et al.* Unsupervised manifold alignment for single-cell multi-omics data. In *Proceedings of the 11th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics*, 1–10 (2020).

[18] Cao, K., Bai, X., Hong, Y. & Wan, L. Unsupervised topological alignment for single-cell multi-omics integration. *Bioinformatics* **36**, i48–i56 (2020).

[19] Demetci, P., Santorella, R., Sandstede, B., Noble, W. S. & Singh, R. Gromov-wasserstein optimal transport to align single-cell multi-omics data. *BioRxiv* (2020).

[20] Li, Y. *et al.* Deep learning in bioinformatics: Introduction, application, and perspective in the big data era. *Methods* **166**, 4–21 (2019).

[21] Li, H. *et al.* Modern deep learning in bioinformatics. *Journal of molecular cell biology* **12**, 823–827 (2020).

[22] Zhu, J.-Y., Park, T., Isola, P. & Efros, A. A. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, 2223–2232 (2017).

[23] Wang, R., Cully, A., Chang, H. J. & Demiris, Y. Magan: Margin adaptation for generative adversarial networks. *arXiv preprint arXiv:1704.03817* (2017).

[24] Yoon, J., Jordon, J. & Schaar, M. Radialgan: Leveraging multiple datasets to improve target-specific predictive models using generative adversarial networks. In *International Conference on Machine Learning*, 5699–5707 (PMLR, 2018).

[25] Choi, Y. *et al.* Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 8789–8797 (2018).

[26] Dai Yang, K. *et al.* Multi-domain translation between single-cell imaging and sequencing data using autoencoders. *Nature Communications* **12**, 1–10 (2021).

[27] Zhang, X. *et al.* Integrated multi-omics analysis using variational autoencoders: Application to pan-cancer classification. In *2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, 765–769 (IEEE, 2019).

[28] Ma, T. & Zhang, A. Integrate multi-omics data with biological interaction networks using multi-view factorization autoencoder (mae). *BMC genomics* **20**, 1–11 (2019).

[29] Makhzani, A., Shlens, J., Jaitly, N., Goodfellow, I. & Frey, B. Adversarial autoencoders. *arXiv preprint arXiv:1511.05644* (2015).

[30] Chen, T., Kornblith, S., Norouzi, M. & Hinton, G. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, 1597–1607 (PMLR, 2020).

[31] Hira, M. T. *et al.* Integrated multi-omics analysis of ovarian cancer using variational autoencoders. *Scientific reports* **11**, 1–16 (2021).

[32] Hu, Z. & Wang, J. T. L. Generative adversarial networks for video prediction with action control. In Seghrouchni, A. E. F. & Sarne, D. (eds.) *Artificial Intelligence. IJCAI 2019 International Workshops - Macao, China, August 10-12, 2019, Revised Selected Best Papers*, vol. 12158 of *Lecture Notes in Computer Science*, 87–105 (Springer, 2019). URL `https://doi.org/10.1007/978-3-030-56150-5\_5`.

[33] Schroff, F., Kalenichenko, D. & Philbin, J. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 815–823 (2015).

[34] Han, W. *et al.* Self-supervised contrastive learning for integrative single cell rna-seq data analysis. *bioRxiv* (2021).

[35] Goodfellow, I. *et al.* Generative adversarial nets. *Advances in neural information processing systems* **27** (2014).

[36] Chen, S., Lake, B. B. & Zhang, K. High-throughput sequencing of the transcriptome and chromatin accessibility in the same cell. *Nature biotechnology* **37**, 1452–1457 (2019).

[37] Wolf, F. A., Angerer, P. & Theis, F. J. Scanpy: large-scale single-cell gene expression data analysis. *Genome biology* **19**, 1–5 (2018).

[38] Bińkowski, M., Sutherland, D. J., Arbel, M. & Gretton, A. Demystifying mmd gans. *arXiv preprint arXiv:1801.01401* (2018).

A Reproducibility

A.1 Details of the experiments

The details of the network architecture are summarized in Table 2. The learning rate of the model is 0.0001, and the batch size is 32 for the real-world dataset, 100 for simulated datasets. We trained this model for 4000 epochs using the Adam optimizer with $\beta_1 = 0.5$, $\beta_2 = 0.999$, and weight decay is set to 0.0001. Using early stopping, we found that the best model usually appears in 1000 epoch or 2000 epoch. The activation function is LeakyReLU after each layer, followed with Batch Normalization.

Table 2: "d" refers to the dimension of the input data. 50 is the dimension of the embedding space. Note that the last hidden layer in encoder and the fast layer in decoder are 100. We set it to be 100 because the dimension is 2613 and 815 for the real-world dataset, 1000 and 500 for simulated datasets. If the dimension of the input data is larger, this number is supposed to be adjusted accordingly.

	Encoder	Decoder
Input size	d	50
Hidden layer size	d, d, d, 100	100, d, d, d
Output size	50	d

The discriminator consists of 2 hidden layers with 50 and 100 nodes, respectively. The output size is 1. The simple classifier consists of 1 layer, and the ouput size is 3.

A.2 Implementation of Con-AAE

We have implemented the Con-AAE in Python 3.7.7 with Pytorch 1.0, whose experiments were run on Nvidia Tesla P100. The source code is available at https://github.com/kakarotcq/RNA-Seq-and-ATAC-Seq-mapping.

A.3 The hyperparameter tunning range

For all the methods, the weights of different loss terms are shown in Table 3. Parts of the weights should be adjusted in accordance of the reality. The weights of anchor loss and cycle-consistency loss are supposed to be low because the noise level of the real dataset is high, with a few mismatches, while contrastive loss should be adjusted to a higher value, and vice versa.

Table 3: Loss terms and their corresponding weights for training Con-AAE.

Loss function	Loss type	Weight
Reconstruction loss	MSELoss	10.0
Adversarial loss	MSELoss	10.0
Classifier loss	CrossEntropyLoss	1.0
Cycle-consistency loss	MSELoss	1.0, 5.0, 10.0
Contrastive loss	MarginRankingLoss	1.0, 5.0, 10.0
Anchor loss	MSELoss	0.05, 0.1, 1.0

B Extra experimental results

We also tried to train Con-AAE with the pairwise information, denoted as Con-AAE-anchor. The comparison between Con-AAE and Con-AAE-anchor is shown in Table 4. We can see that the performance of integration is almost the same while the performance of alignment is improved slightly as the pairwise information is added. In general, Con-AAE is a very stable unsupervised method, which is robust to noise. So the weak supervised pairwise information with a high level of noise does not improve the performance of Con-AAE very significantly. But still, this experiment suggests that Con-AAE is a flexible framework, which can incorporate supervised information.

We also conducted ablation studies on the simulated datasets. The quantitative results are shown in the Table 5 and Figure 7. We can see that the cycle-consistency loss and contrastive loss could improve the model in many cases, but not stably enough. Compared with that, Con-AAE almost always has the best performance with the change of data sizes and SNRs, while the other methods struggle when the SNRs or data sizes change.

Table 4: Comparison between Con-AAE and Con-AAE-anchor on the real dataset. Con-AAE is a very stable unsupervised method, which is robust to noise. So the weak supervision with a high level of noise does not improve the performance of Con-AAE significantly.

Method	Integration ACC	Recall@k=10	Recall@k=20	Recall@k=30	Recall@k=40	Recall@k=50
Con-AAE	63.9	5.3	12	17	22.6	28.4
Con-AAE-anchor	63.4	7	13.7	18.4	25.4	29.1

Table 5: Ablation study on the simulated datasets with various SNRs: basic plus adversarial loss (denoted as Basic_adv), basic plus adversarial loss and cycle-consistency loss method (denoted as Basic_adv_cyc), basic plus adversarial loss and contrastive loss (denoted as Basic_adv_contra) and basic plus adversarial loss, cycle-consistency loss, and contrastive loss (Con-AAE). The performance of Con-AAE is in bold. Clearly, Con-AAE's performance is very stable across different data sizes and SNRs, compared to the other baseline methods.

#Sample	Method	No Noise	SNR25	SNR20	SNR15	SNR10	SNR5	SD	AVG
1200	Basic_adv	80.4	83.3	83.3	81.3	82.1	76.7	2.47	81.18
	Basic_adv_cyc	83.3	83.3	83.8	83.3	84.2	82.3	0.64	83.36
	Basic_adv_contra	85.0	85.4	85.0	87.5	83.8	72.5	5.38	83.20
	Con-AAE	**87.5**	**86.7**	**87.9**	**88.3**	**87.1**	**81.7**	**2.43**	**86.53**
2100	Basic_adv	85.0	85.0	84.2	82.9	84.5	81.4	1.42	83.80
	Basic_adv_cyc	72.4	72.4	71.2	70.4	76.2	77.4	2.82	73.33
	Basic_adv_contra	91.0	90.0	91.9	91.4	86.2	72.1	7.63	87.10
	Con-AAE	**89.3**	**88.8**	**88.6**	**90.0**	**87.8**	**82.6**	**2.67**	**87.85**
3000	Basic_adv	81.1	82.5	82.5	81.8	77.2	75.3	3.06	80.07
	Basic_adv_cyc	85.8	86.8	87.0	86.5	84.0	80.6	2.47	85.12
	Basic_adv_contra	88.8	90.0	89.2	67.5	60.8	86.5	12.87	80.47
	Con-AAE	**89.7**	**88.6**	**88.2**	**89.1**	**74.2**	**82.0**	**6.12**	**85.30**
6000	Basic_adv	81.0	81.2	74.2	67.5	44.7	35.8	19.33	64.07
	Basic_adv_cyc	81.1	85.2	85.5	78.6	79.2	78.5	3.24	81.35
	Basic_adv_contra	90.8	90.9	91.3	90.0	81.9	67.6	9.43	85.42
	Con-AAE	**87.5**	**89.7**	**88.3**	**87.3**	**82.9**	**77.1**	**4.69**	**85.47**

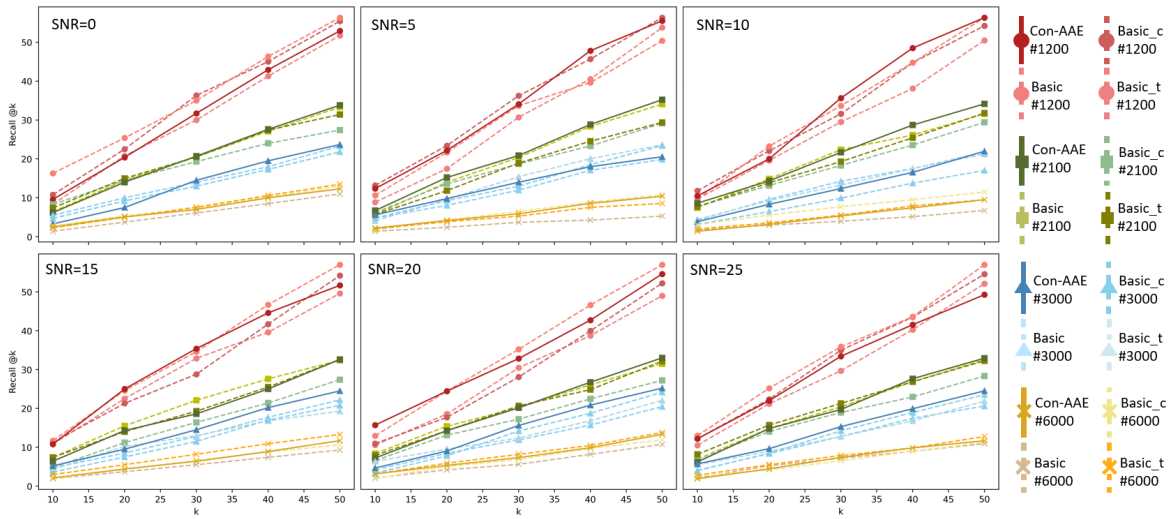


Figure 7: The alignment performance on simulated datasets with different components.