Thesis Project: Single-cell Multi-omics Data Integration through Deep Learning

Chaozhong Liu

1. Introduction

Single-cell omics technology is being rapidly developed to measure individual cellular changes at the epigenome, genome, and transcriptome levels. However, omics data from different levels, termed multi-omics data, are still analyzed separately. Separate analysis cannot link each independent result from all levels, generating only pieces of information without an integrated explanation for biological questions. Therefore, we aim to develop a single cell multi-omics integration tool, to fill this gap, extract biological information, and solve biological questions.

2. Literature Review

Integrating single-cell omics data is an exciting frontier that remains a challenge in molecular biology. To integrate different omics data types, two main strategies are proposed, the experimental approach that profiles multiple omics data simultaneously, and computational approaches merging independent omics datasets. With the low throughput and high cost of experimental approaches[1], the development of computational methods is of increasing importance to integrating multi-omics datasets.

The task is illustrated in **Figure 1**. In general, we have multiple independent datasets from different techniques, like scRNA-seq, scATAC-seq, etc. By applying computational methods, we expect to finally get an integrated dataset merging all identities and features. However, these datasets have different features (genes, peaks, etc.) and different cells without any correspondence information.

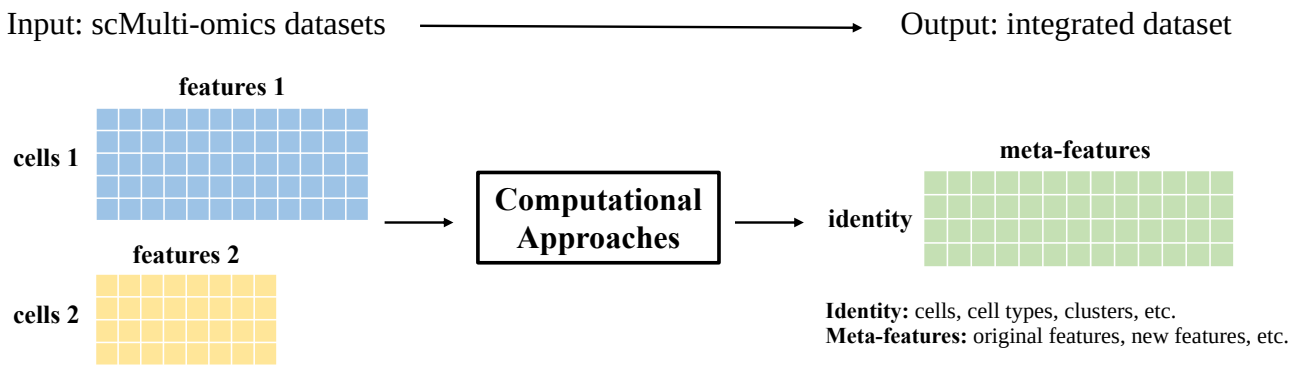

Figure 1. Task description.

From reviewing current methods, embedding datasets into a shared latent space is the general integration strategy. This strategy places great importance on a suitable embedding algorithm. And because of distinct features and cells, it is a key step to achieve, in which all the algorithms differ. Current methods can be grouped into two categories: (1) mapping features to a common feature set, and then do embedding; (2) do embedding with distinct features and cells without mapping. Below is a brief review of the mainstream algorithms.

2.1 Methods involving empirical feature mapping

2.1.1. Seurat

Seurat[2] is a popular algorithm used by researchers when they analyze their single-cell data. It applies joint CCA to project single-cell RNA-seq data (reference set) and single-cell ATAC-seq data (query set) into a common latent space (Figure 2. A). In this latent space, Seurat finds the mutual nearest neighbors between cells in different datasets, which is also called anchors (Figure 2. B). Anchors will then be scored and filtered by the consistency between latent space and high-dimensional space (Figure 2. C). With scored anchors, Seurat generates the weight matrix ***W*** that can be understood as measuring the strength of association between each

query set cell and each anchor pairs (Figure 2. D). Finally, datasets are integrated by moving each cell of the query set to reference set in the feature space (Figure 2. E). The scale and direction of this movement ***C*** is determined by all anchors and the weights on that cell being moved. But all before applying joint CCA, reference set and query set need to have a common set of features. Usually, it is done by empirical feature mapping to convert ATAC-seq peaks to gene-level features called gene activity. For example, one of this empirical mapping is to sum up the counts of all peaks that lie within 3kb upstream of the gene promoter regions. Only when the mapping is done, can Seurat implement the joint CCA algorithm.

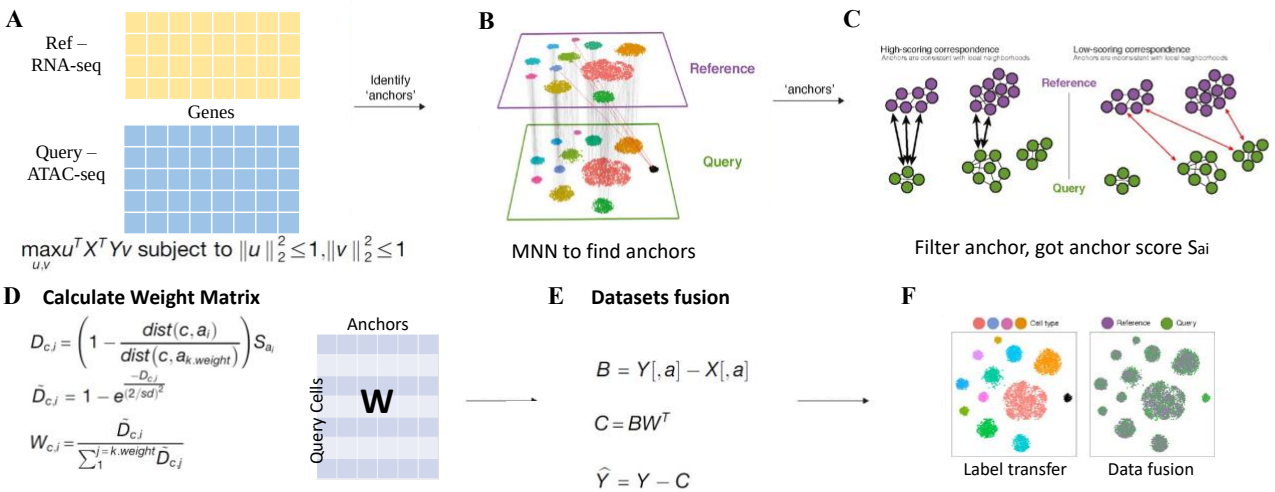

Figure 2. Overview of Seurat.

2.1.2 Liger

Same as Seurat, Liger[3] also requires an empirical feature mapping before applying this algorithm. The workflow of Liger is illustrated in **Figure 3**. First, Liger applies integrative nonnegative matrix factorization (iNMF) to identify shared and dataset-specific metagenes across datasets. Then datasets are embedded into a shared factor neighborhood (SFN) graph by following steps: (1) Build k-nearest neighbor graphs separately for each dataset using the H factor loadings; (2) Annotate each cell *i* with *F(i)*, the highest H factor loading; (3) For each cell *i*, collect the ‘‘factor neighborhood’’ vector *FN(i)* by computing a histogram of *F(i)* for each of its *k* nearest; (4) Calculate Manhattan distance between pairs of cells (*i, j*); (5) Connect pairs of cells with low distance to generate the SFN. Finally, Liger performs Louvain community detection on the graph to jointly identify cell clusters across datasets.

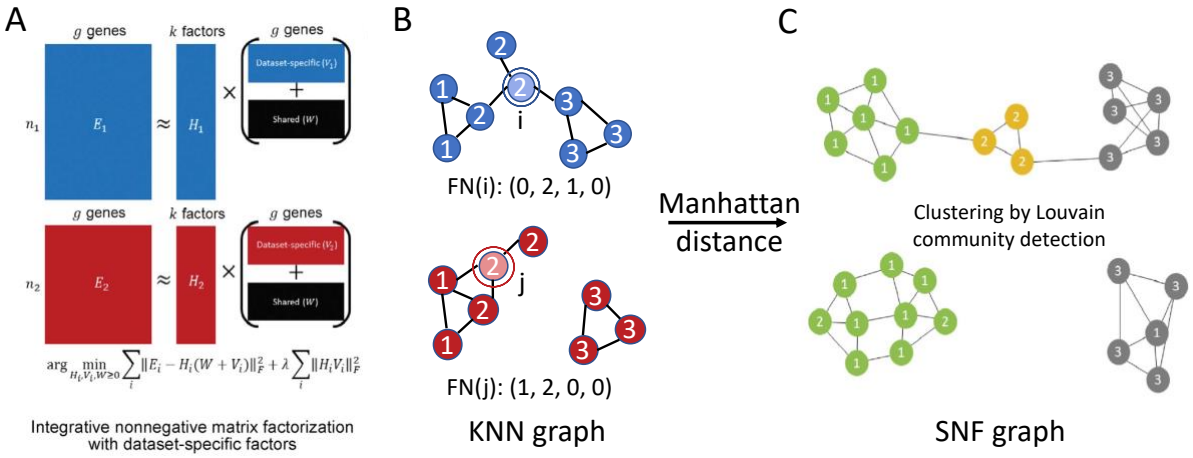

Figure 3. Overview of Liger.

2.2 Methods without empirical feature mapping

2.2.1 bindSC

The input of bindSC[4] are omics data from different platforms, ***X*** and ***Y*** showed in **Figure 4A**. Same as Seurat, bindSC also applies CCA to perform embedding. But instead of using empirical mapping to match features, bindSC initializes mapping by empirical rules, transferring ***Y*** to ***Z***. But then bi-CCA will capture correlated variables in both cells and features (rows and columns), during which this feature mapping ***Z*** is optimized by the machine in each iteration. In the co-embedding space derived from bi-CCA, bindSC clusters cells between two modalities and finally, constructs pseudo-cell level multi-omics profiles as the final integrated dataset.

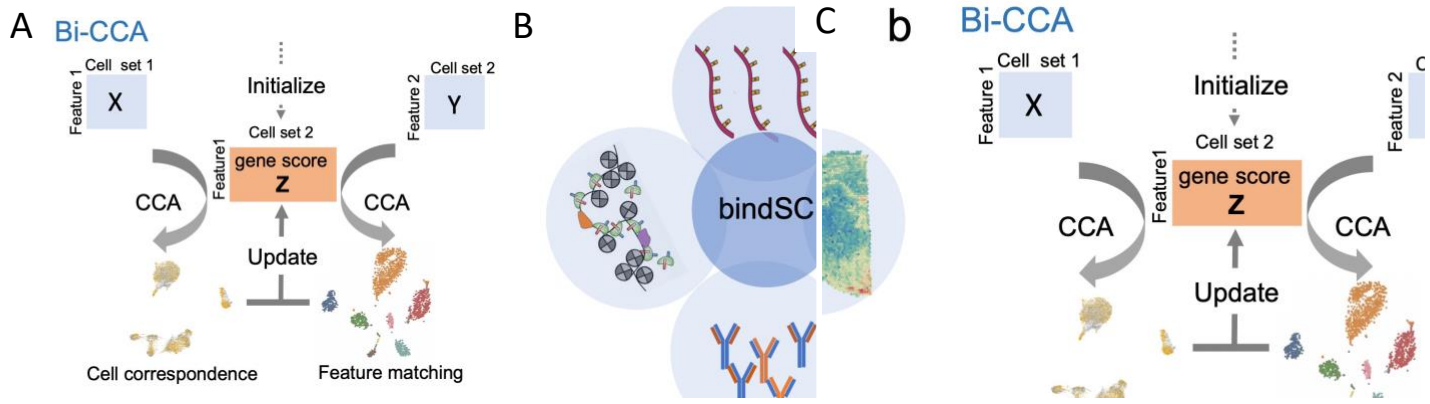

Figure 4. Overview of bindSC.

2.2.2 MMD-MA

Maximum mean discrepancy-based manifold alignment (MMD-MA)[5,6] aligns single-cell datasets from different modalities in a common shared latent space, without requiring any correspondence information of either cells or features. The idea behind MMD-MA and its objective function:

$$\min_{\alpha_1,\alpha_2} \mathrm{MMD}(K_1\alpha_1, K_2\alpha_2)^2 + \lambda_2(\mathrm{dis}(\alpha_1) + \mathrm{dis}(\alpha_2)) + \lambda_1(\mathrm{pen}(\alpha_1) + \mathrm{pen}(\alpha_2))$$

is to create a latent space where the maximum mean discrepancy (MMD) between the data sets is minimized while maintaining the underlying structure of each data set. The MMD term:

$$\mathrm{MMD}^2\left(\left\{u_1^{(1)},\ldots,u_{n_1}^{(1)}\right\},\left\{u_1^{(2)},\ldots,u_{n_2}^{(2)}\right\}\right) = \frac{1}{n_1^2}\sum_{i,j=1}^{n_1} K_M(u_i^{(1)},u_j^{(1)}) - \frac{2}{n_1 n_2}\sum_{i=1}^{n_1}\sum_{j=1}^{n_2} K_M(u_i^{(1)},u_j^{(2)}) + \frac{1}{n_2^2}\sum_{i,j=1}^{n_2} K_M(u_i^{(2)},u_j^{(2)})$$

ensures the discrepancy between inter-datasets difference and inner-dataset difference is minimized; While the second term:

$$\mathrm{dis}(\alpha_I) = ||K_I - K_I\alpha_I\alpha_I^\top K_I^\top||_2^2,$$

minimizes the distortion between input space and the latent space. The way MMD-MA circumvents feature mapping is to generate similarity matrices, ***K1*** and ***K2***, as input instead of the raw matrices. So cells can be aligned in the latent space with no requirement for features.

2.2.3 UnionCom

UnionCom[7] extends the generalized unsupervised manifold alignment (GUMA) algorithm to solve the co-embedding problem in single-cell data. Its workflow is shown in **Figure 5**. It applies a similar strategy as MMD-MA to avoid the need for feature correspondence. Instead of using similarity matrices as input, UnionCom uses geometrical distance matrices $\boldsymbol{K_x}$ and $\boldsymbol{K_y}$ (Figure 5B). Next, the task is to re-scale $\boldsymbol{K_x}$ and match cells between $\boldsymbol{X}$ and $\boldsymbol{Y}$ one-to-one by point matching matrix $\boldsymbol{F}$. By defining the objective function, the scale factor and the $\boldsymbol{F}$ matrix are optimized, from which the cell correspondence can be inferred (Figure 5C). Finally, to co-embed the datasets into a shared latent space, UnionCom applies t-SNE with distance penalty ensuring matched cell in $\boldsymbol{F}$ is close to each other (Figure 5D).

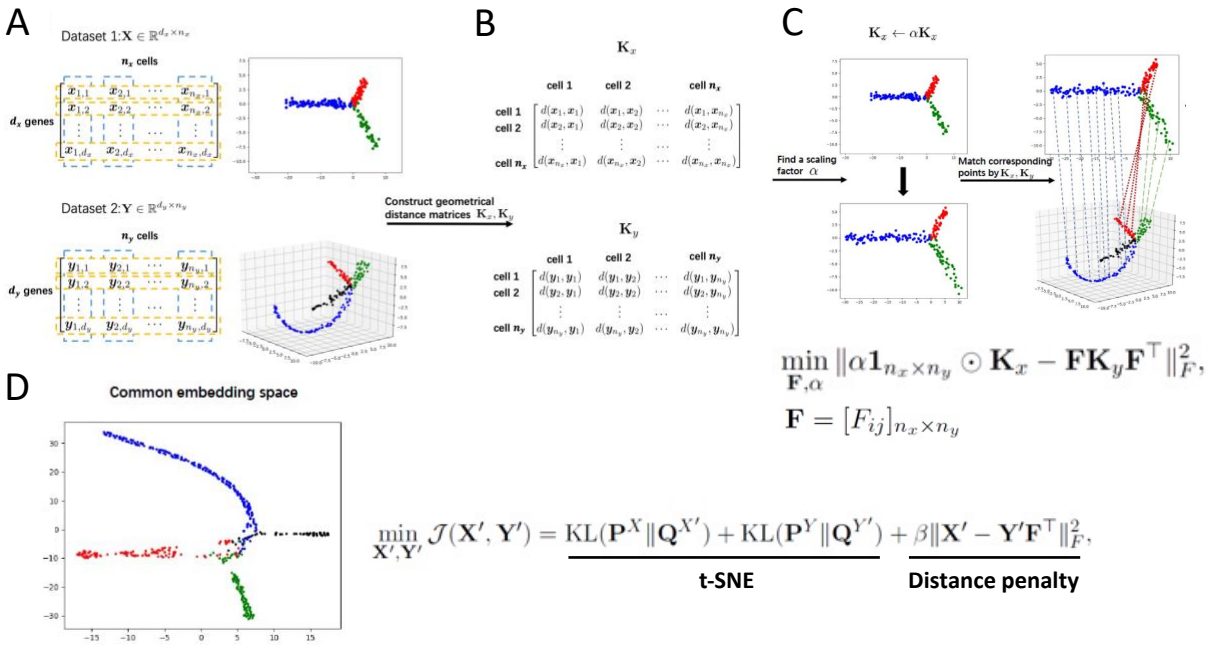

Figure 5. Overview of UnionCom

2.3 Discussion about integration methods

Above are five representative algorithms for integrating single-cell multi-omics data. Both ways of integration have their pros and cons. Integration by assuming feature correspondence (feature mapping), like Seurat and Liger, will be easy to interpret the integration result and extract biological explanation. For example, aligned cells between transcriptomic data and chromatin activity data allow the detection of links between transcriptome factor and gene expression. But the embedding done by methods like this can easily be distorted if the empirical mapping doesn't hold, which is still unclear in this field.

2.3.1 Correlation between modalities after empirical feature mapping

To check whether the empirical mapping is a good representation of the ground truth, we apply it to SNARE-seq[10] adult mouse brain dataset, where transcriptome and chromatin activity are simultaneously profiled in each cell, indicating we have the ground truth. Results are summarized in **Figure 6**. The empirical mapping we choose is the sum of all peaks 2kb upstream of the promoter region to represent the gene activity (Figure 6A). Due to the sparsity of single-cell data (Figure 6B), we filtered cells by the number of genes that are both detected in RNA-seq and ATAC-seq data. And when calculating Spearman correlation, only detected genes are used. Figure 6C-D summarized the correlation of all 8448 pairs, showing that only a weak correlation is found, and the number of genes is limited. One typical pattern is also shown in Figure 6E. Besides the absolute value of correlation, we also studied the relative correlation comparing the true pair and random pairs. Figure 6F summarizes the ranks of true pairs among random pairs. It is expected that the majority of cells are of high ranks (high value on the left

of the histogram) if the feature mapping is reasonable. However, many of the ranks are even worth than random pairs, indicating this empirical mapping doesn't work well. What's more, the high sparsity also shades the true information and adds noises to the data.

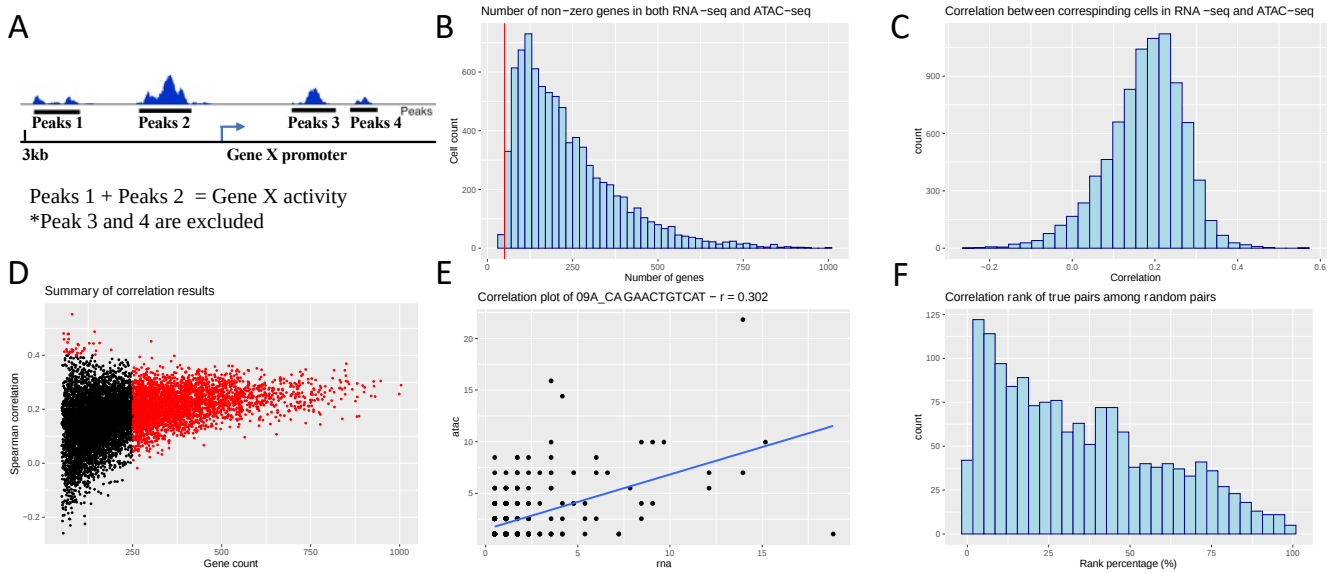

Figure 6. Correlation between transcriptome and gene activity generated from feature mapping.

The second kind of method avoids using empirical mapping to various extents and showed promising embedding results when tested in multi-omics single-cell data. But the results are less interpretable. Inter-plays between modalities cannot be inferred from a model of this kind. Or in other words, feature mapping is still unclear. Besides, manifold alignment methods like MMD-MA and UnionCom rely heavily on the basic assumption that cells of single-cell multi-omics datasets are sampled from similar intrinsic low-dimensional structures embedded in the data. But this assumption can easily be wrong in a real problem. For example, single-cell RNA-seq done from neural tissues consists mainly of non-neuronal cells while single-cell ATAC-seq done consists mainly of neuronal cells. This discrepancy will surely make the intrinsic data structure different and break the assumption.

Hypothesis and Aims

Hypothesis: With well-designed deep learning model, the machine can learn the complex feature mapping between omics data while performing the integration task.

There are three main aims of the thesis project.

Aim 1: To develop a neural network framework for single-cell multi-omics data integration. We will build an interpretable deep learning model that can (i) perform integration between single-cell omics data, (ii) learn the feature mapping between omics, and (iii) be easily interpreted to extract the inter-play between omics.

Aim 2: To develop a bioinformatic tool that can extract cell type specific features from multi-omics data. Neural network models are often treated as a black box. However, with well-designed model and tool available to interpret neural networks, we hypothesize that our model can extract biological information from the joint modeling of multi-omics data. Thus, we will develop user-friendly bioinformatic tools to help interpret the trained model and integration results.

Aim 3: To dissect the interplay of epigenome and transcriptome during Alzheimer's Disease Progression using single cell multi-omics data. We will apply the model to scRNA-seq and scATAC-seq data from an AD mouse model. We will then validate our model with discovered markers to compare with known disease progression markers such as APP and PSEN1 and perform Gene Ontology Annotation. We will also interpret the trained model to understand the interplay of transcriptome and chromatin accessibility during disease progression.

Significance

Systematic view combining multi-omics is important to understand diseases

Complex diseases like cancer, heart disease, and Alzheimer's disease are caused by multiple and still unknown factors in the biological system. Many more units in multiple molecular layers play their role in the disease progression than a simple Mendelian single-gene disorder, resulting in both aetiological and clinical heterogeneity. This heterogeneity complicates diagnosis, treatment, and the design and testing of new drugs[8]. With the development of high-throughput technologies, measuring the multiple omics data at the single-cell level has explained part of the heterogeneity from the perspective of cell-type differences. However, the analysis of only one data type is limited to correlations rather than the real complicated causative changes involving multiple layers like genome, epigenome, proteome, etc. Integration of different omics data types can elucidate potential causative changes that lead to disease, or the treatment targets, that can be then tested in further molecular studies[9].

Single cell omics data integration remains a challenge

Discussed in the literature review above, there are still gaps to be filled to integrate single-cell omics data. Methods like Liger and Seurat require feature correspondence (common feature set), so an empirical feature mapping is inevitable. Due to data sparsity and no optimal mapping, the data will be distorted to some extent. BindSC optimized the feature mapping process by bi-CCA but the feature mapping is not thoroughly interpreted or validated. Methods involving manifold alignment doesn't require feature correspondence, but the model is less biological explainable, and the performance is limited by the weak assumption. In a word, there is still a need for an algorithm that can optimize the feature mapping between datasets while performing integration task.

Progress and experimental results to date

1. Baseline model to validate the feasibility of deep learning models

To test whether neural network is capable of the co-embedding and alignment task, and also to learn the behavior of NN in doing integration task, we built a baseline encoder neural network model.

Architecture of the neural network and objective functions. To align cells and remove the modality effect when integrating multi-omics datasets, we created a model based on scDGN neural network framework, including three modules, as shown in **Figure 7B**. The encoder module is used to represent datasets into a common lower dimensional space and contains two fully connected layers which produce the hidden features $x' = f_{e2}(f_{e1x}(x; \theta_{e1x}); \theta_{e2})$ for omics data X or $y' = f_{e2}(f_{e1y}(y; \theta_{e1y}); \theta_{e2})$ for omics data Y, where θ represents the parameters in these layers. The label classifier, $f_{lx}(x'; \theta_{lx})$ for X and $f_{ly}(y'; \theta_{ly})$ for Y, ensures the inner structure of datasets remains in the common space. The goal of the domain discriminator $f_d(x'; \theta_d)$ and $f_d(y'; \theta_d)$ is to determine whether a pair of inputs ((x_i,x_j), (x_i,y_j), (y_i,x_j), (y_i,y_j)) are from the same domain or not. The overall objective function to be minimized is:

$$E = L_l(f_{lz}(z_1'; \theta_{lz}), lz) + \lambda L_d(f_d(z_1'; \theta_d), f_d(z_2'; \theta_d)),$$

where $z_1, z_2 \in \{x, y\}$, λ can control the trade-off between the goals of domain invariance and higher classification accuracy. Inspired by Siamese networks, the domain loss or adversarial loss adopts a contrastive loss for a pair $\boldsymbol{z_1}$ and $\boldsymbol{z_2}$, where $\boldsymbol{z} \in \{x, y\}$:

$$L_d(f_d(x'; \theta_d), f_d(y'; \theta_d)) = U(1 - \cos(f_d(z_1'), f_d(z_2'))) + (1 - U)\max\{0, \cos(f_d(z_1'), f_d(z_2')) - m\},$$

where U=0 indicates the two cells are from the same modality but different cell types (positive pairs), and U=1 indicates that they are pre-defined anchors (negative pairs, anchor is defined in the next part), $cos(\cdot)$ is the cosine embedding loss, and m is the margin that indicates the prediction boundary.

Overall, the aim of the objective function is to minimize the label classification loss and the domain/adversarial loss. In this way, modality effect will be removed in the encoder-generated co-embedding space while data structure is kept.

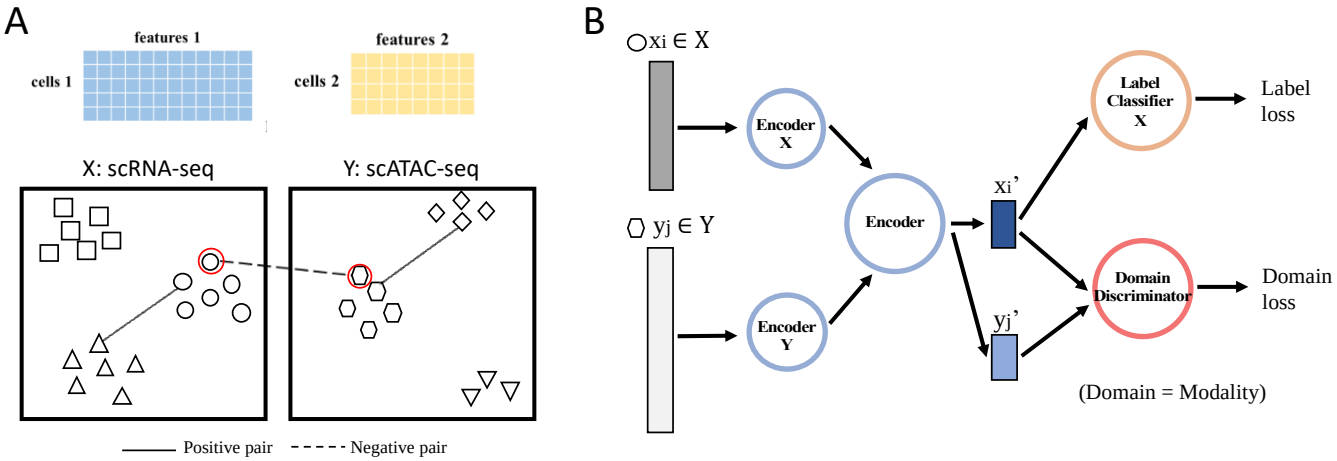

Figure 7. Baseline model overview.

Strategy to make the neural network an unsupervised model. Applying a supervised method like the neural network to the biological problem is being criticized for the need for training, limiting the generalization of its utility. Here we made the model an unsupervised one by strategy described as follows.

The true labels needed for training the model are cell type labels so that the label classifier can be trained, and the value of U in the domain discriminator can be determined. However, these true labels are, in fact, our goal of integrating multi-omics dataset and is not available most of the time. To integrate datasets without any true label, we modified the label classifier to learn each dataset's UMAP clustering result. For domain discriminator, U=0 represents pairs from the same dataset in different clusters. And U=1 stands for pairs taken as anchors (cell pairwise correspondences between single cells across datasets[2]). These anchors are selected by diagonalized

CCA followed by Mutual Nearest Neighbors and anchor filtering, applied in Seurat. During training, AMANDA will learn to separate cells with different clusters while merging anchors as close as possible. With this, no true label is needed for the model, and we can integrate any datasets without training beforehand.

2. Preliminary results

The baseline model successfully integrated the SNARE-seq cell mixture dataset. With the model constructed, we tried to integrate the snRNA-seq and scATAC-seq data from SNARE-seq[10] where transcriptome and chromatin activity are simultaneously profiled in each cell. Fine-tuned network is being trained for 100 epochs with a learning rate of 1e-3. Figure X shows the performance of our model and a comparison with Seurat. UMAP was done based on the PCA results from a 32-length hidden layer vector to represent the co-embedded space. Figure 8A shows that almost all the cells from the two datasets are merged as they are supposed to be, and different clusters are well-separated.

By applying a KNN classifier on the first 3 PCs (as they explained most of the variance) to transfer snRNA-seq cell labels to snATAC-seq cells and then validating with the true label, we achieved the mean label transfer accuracy of 92%, higher than the performance of Seurat. All these have proved that our baseline model works for this task.

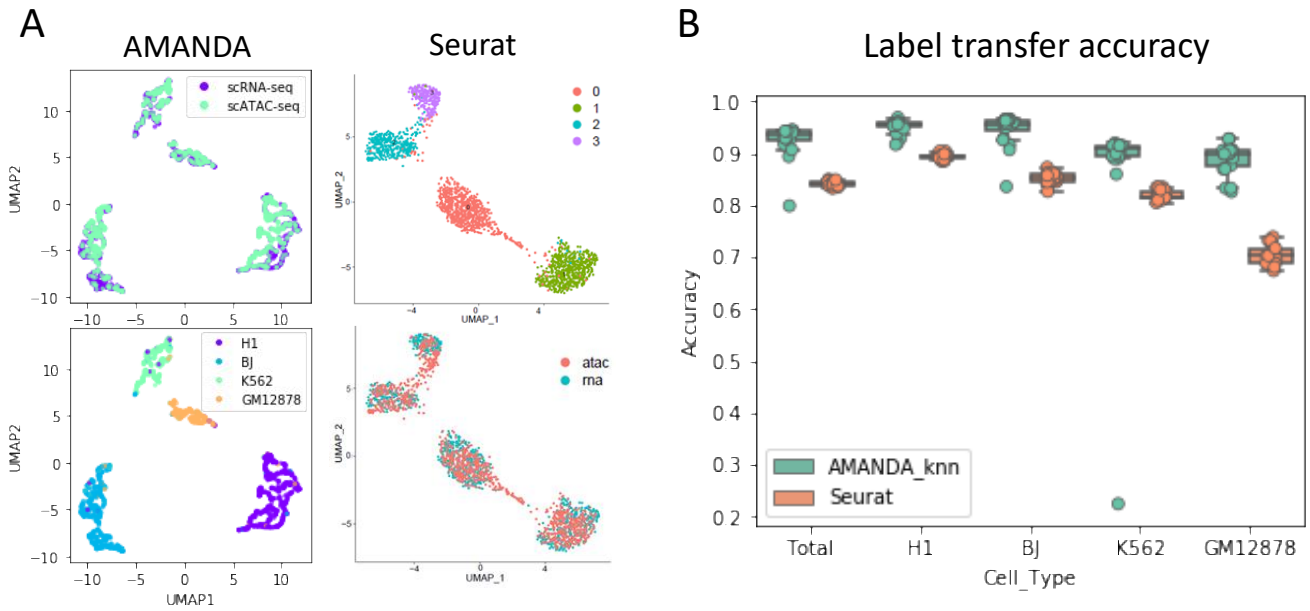

Figure 8. Performance comparison between the baseline model and Seurat.

Activation Maximization has found potential marker genes and motifs. Interpretability is also an important property that we expect our integration tool should have. Activation maximization has been a popular technique to decode the neural network and detect important features. We applied AM to the well-trained baseline model to generate the prototype transcriptome and chromatin accessibility of each cell type. By differential analysis comparing the 4 prototypes, we were able to find the potential marker genes and motifs. Then we validated the marker genes with two studies[11,12] that profile the transcriptomes of cell lines and defined a set of markers. Figure 9A shows the heatmap of prototype gene expression patterns, with rows annotated by cell lines and columns indicating which cell line do these marker genes belong to according to the two studies. With the differential accessible regions found from prototypes, we used Homer to annotate and predict potential motif and TFs, then validated with another independent study. Figure 9B shows that cell-type-specific transcription factors and motifs are significantly detected only in that cell type. We conclude that AM can to help interpret the trained baseline model and made it possible to extract the inter-play between omics data.

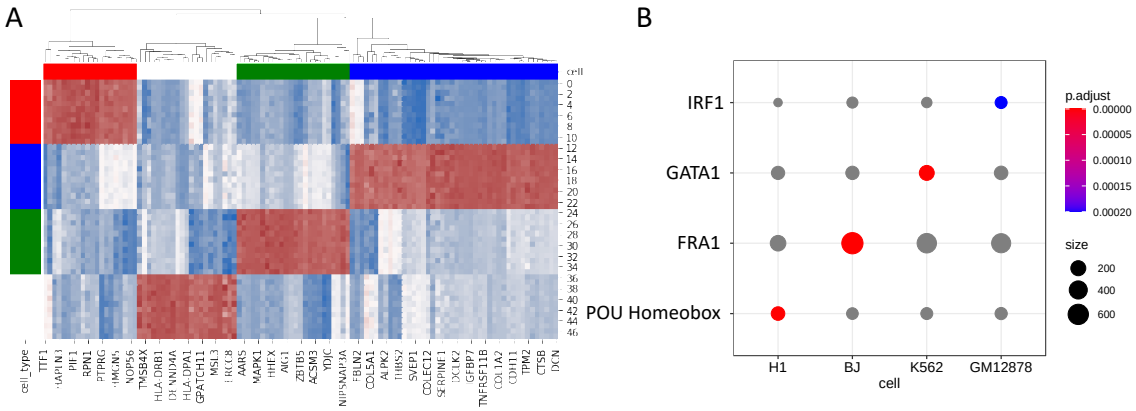

Figure 9. Activation Maximization has found important genes and motifs for cell types.

3. Discussion about the baseline model

Our baseline model has validated the feasibility of neural networks and especially encoder in completing the integration task. Decent label transfer accuracy has been reached, and the model has learned the key genes and open regions for cell types. However, the model does have several limitations to be improved.

First, it relies on the anchors from Seurat as the pseudo true label. This is the biggest limitation of our current baseline model. As mentioned in the literature review part, a bad feature mapping could distort the data structure. Due to that, the anchors predicted by Seurat are not that reliable. Figure 10A shows that fraction of anchor pairs that are from the same cell type is low. And we cannot filter anchors by the scores given by Seurat as shown in Figure 10B.

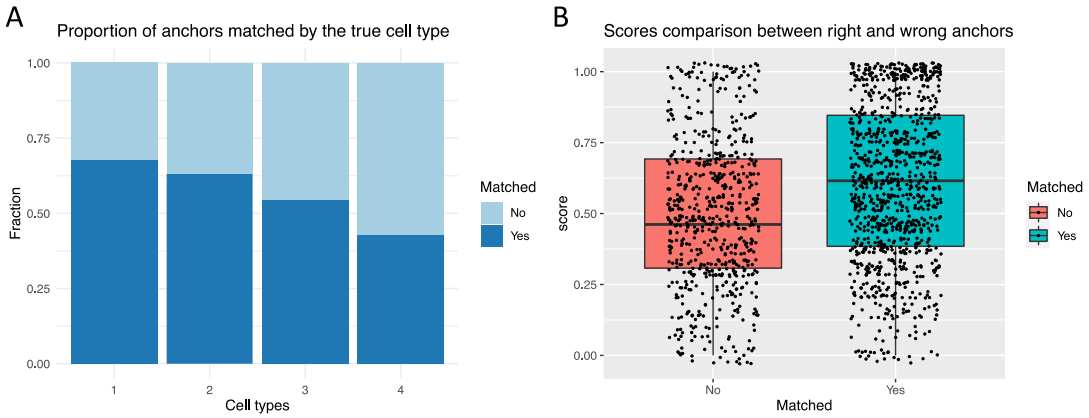

Figure 10. Anchor quality summary.

Second, although our model could be interpreted by using activation maximization, we still cannot get the whole picture of feature mapping but only some relationships between genes and open regions. There is still space to improve the baseline model.

Third, the robustness of our model is also unsatisfactory. This neural network needs to be fined tuned to achieve a decent performance and it's hard to avoid overfitting. Besides, the prototype derived from activation maximization is not that robust either. We tried to train 10 baseline models and apply AM to derive the prototype and markers. But the marker overlap among the 10 models is few. All these indicate that we need to improve the robustness of our new model.

Specific Aims and future plan

Specific Aim: Build a One-shot Learning model through Siamese Neural Network embedded with feature mapping between ATAC-seq and RNA-seq

To fill the current gaps in the realm of single-cell multi-omics integration (unclear feature mapping and sparsity), and improve our baseline model (relies on Seurat anchor, hard to generalize), we will build a new Siamese network fully utilizing the strength of one-shot learning. The network will be trained with the ground truth label from the SNARE-seq dataset. We believe the trained model can be applied in any other dataset for two reasons. First, one-shot learning is trained to judge the similarity between pairs rather than predicting any categories, makes it quite general so long as it's still applied to a similar task. Second, we believe the feature mapping between chromatin accessibility and transcriptome is quite general a problem regardless of the cell types or tissues. Once the model is trained, it should be capable of mapping ATAC-seq peaks to genes in other datasets.

Architecture of the Siamese network

Overview. The model receives simultaneously one cell from the snRNA-seq data and another from the scATAC-seq data as the input, denoted as $x \in \mathbb{R}^{1\times G}$, $y \in \mathbb{R}^{1\times P}$. G is the number of genes and P is the number of bins derived from cutting the genome into 5k bp windows and sum all reads within each window. Y will go through the feature mapping module first to get $x2 \in \mathbb{R}^{1\times G}$. Then x and $x2$ will go through several fully-connected layers to get the final Siamese loss.

Feature mapping module with attention mechanism embedded. Figure 11 illustrates the structure of this module. Convolutional blocks with pooling first reduce the spatial dimensionality from N to N/width/2^{L2}, creating C different channels. Self-attention pooling will detect the local interaction among genomic regions during training. Then multi-heads attention (MHA) blocks will capture long-range interactions across the whole genome, such as enhancer and promoter. Cropping block trims M positions on each side to ensure each output position will have at least M positions of genomic context on both sides. Finally, a convolutional layer with G channels will generate a vector for each gene, which goes through feed forward layers, giving the final output $x2 \in \mathbb{R}^{1\times G}$. This whole module is inspired by the Transformer[13] and its derivative Enformer[14].

Siamese module. x and $x2$ will go through several feed forward layers to reduce the dimension, ending up with $x' \in \mathbb{R}^{1\times G'}$, $x2' \in \mathbb{R}^{1\times G'}$. Finally, Siamese loss will be calculated as follows:

$$L(x', x2', U) = (1-U)D(x', x2')^2 + U * \max\{0, m - D(x', x2')^2\},$$

U is the label indicating whether the two cells are corresponding pairs (U=0) or not (U=1). D() defines the distance between x' and $x2'$. During training, the loss will be minimized.

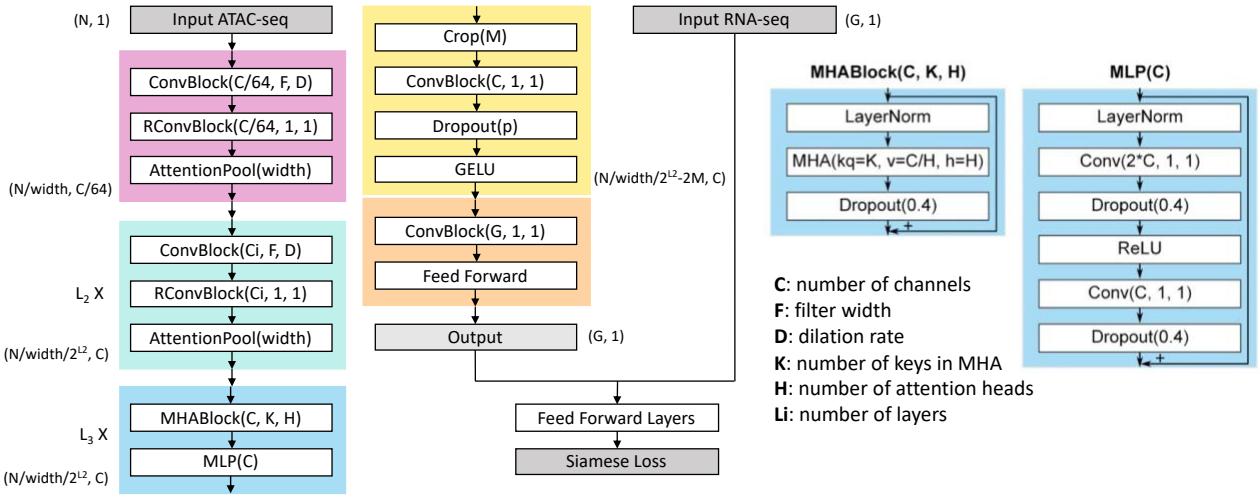

Figure 11. Architecture of the Siamese network.

Strategy to define the labels

The Siamese network is trained with **APN** triplets. Here in our case, the **A**nchor cell will be one from the scRNA-seq data, while the **P**ositive cell is the corresponding cell of anchor in scATAC-seq data, and the **N**egative pair could be any other cell from the scATAC-seq data. However, this will make the model hard to be trained. First, due to a large number of single cells, it will take unimaginable time to be trained with all possible APN combinations. Second, some cells are similar to each other, making the model either learn nothing or fail to be general. Thus, we need to modify the true label definition. Rather than the one-to-one cell correspondence, we take a sub-group of cells as the same identity, defined by the K Nearest Neighbors graph derived from the scRNA-seq data. By doing this, the learning burden is released, and the model is easier to be generalized. Besides, we believe that taking the K-Nearest-Neighbors as the same identity will let the neighbors compensating each other in undetected genes and, to some extent, solved the data sparsity problem.

Performance evaluation

The model will be trained on the SNARE-seq adult mouse brain data. When decent training performance is reached, it will be tested on the SNARE-seq newborn mouse brain data and SHARE-seq mouse skin data. The performance will be evaluated in four metrics. First, the accuracy of whether true pairs are predicted to be identical by the model. Second, cell type labels of cells in scATAC-seq data will be predicted by weighted max voting of the labels of its identical cells predicted in scRNA-seq data. It is the process of transferring scRNA-seq cell type label to scATAC-seq data. Then the label transfer accuracy will be calculated. Third, we will project transcriptome data, together with generated transcriptome data from scATAC-seq, into a co-embedding space. With the co-embedding result, silhouette score will be calculated to measure how well clusters are separated and modalities are merged.

Reference List

1. Tim Stuart, Rahul Satija. Integrative single-cell analysis. Nature Reviews Genetics 20(5):257-272 (2019).
2. Tim Stuart, Andrew Butler, Paul Hoffman, Christoph Hafemeister, Efthymia Papalexi, William M.MauckIII, Yuhan Hao, Marlon Stoeckius, Peter Smibert, Rahul Satija. Comprehensive Integration of Single-Cell Data. Cell 177, 1888–1902 (2019).
3. Joshua D. Welch, Velina Kozareva, Ashley Ferreira, Charles Vanderburg, Carly Martin, Evan Z. Macosko. Single-Cell Multi-omic Integration Compares and Contrasts Features of Brain Cell Identity. Cell 177, 1873–1887 (2019).
4. Jinzhuang Dou, Shaoheng Liang, Vakul Mohanty, Xuesen Cheng, Sangbae Kim, Jongsu Choi, Yumei Li, Katayoun Rezvani, Rui Chen, Ken Chen. Unbiased integration of single cell multi-omics data. bioRxiv 2020.12.11.422014 (2020).
5. Ritambhara Singh, Pinar Demetci, Giancarlo Bonora, Vijay Ramani, Choli Lee, He Fang, Zhijun Duan, Xinxian Deng, Jay Shendure, Christine Disteche, William Stafford Noble. Unsupervised manifold alignment for single-cell multi-omics data. bioRxiv 2020. 06. 13. 149195 (2020).
6. Jie Liu, Yuanhao Huang, Ritambhara Singh, Jean-Philippe Vert, William Stafford Noble. Jointly embedding multiple single-cell omics measurements. bioRxiv 644310 (2019)
7. Kai Cao, Xiangqi Bai, Yiguang Hong, Lin Wan. Unsupervised topological alignment for single-cell multi-omics integration. Bioinformatics 36, i48-i56 (2020).
8. AmanPreet Badhwar, G Peggy McFall, Shraddha Sapkota, Sandra E Black, Howard Chertkow, Simon Duchesne, Mario Masellis, Liang Li, Roger A Dixon, Pierre Bellec. A multiomics approach to heterogeneity in Alzheimer's disease: focused review and roadmap. Brain 143(5):1315-1331 (2020).
9. Yehudit Hasin, Marcus Seldin, Aldons Lusis. Multi-omics approaches to disease. Genome Biol 18(1):83 (2017).
10. Song Chen, Blue B Lake, and Kun Zhang. High-throughput sequencing of the transcriptome and chromatin accessibility in the same cell. Nat Biotechnol 37(12):1452-1457 (2019).
11. Bernstein BE, Stamatoyannopoulos JA, Costello JF, et al. The NIH Roadmap Epigenomics Mapping Consortium. Nat Biotechnol 28(10):1045-1048 (2010).
12. Uhlén M, Fagerberg L, Hallström BM, et al. Proteomics. Tissue-based map of the human proteome. Science 347(6220):1260419 (2015).
13. Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, Illia Polosukhin. Attention Is All You Need. arXiv:1706.03762 [cs.CL] (2017).
14. Žiga Avsec, Vikram Agarwal, Daniel Visentin, Joseph R. Ledsam, et al. Effective gene expression prediction from sequence by integrating long-range interactions. bioRxiv 2021.04.07.438649 ; doi: https://doi.org/10.1101/2021.04.07.438649.