SMILE: Mutual Information Learning for Integration of Single Cell Omics Data

Yang Xu[1], Priyojit Das[1], Rachel Patton McCord[2*]

[1]UT-ORNL Graduate School of Genome Science and Technology, University of Tennessee, Knoxville, TN [2]Biochemistry & Cellular and Molecular Biology Department, University of Tennessee, Knoxville, TN

* Correspondence: rmccord@utk.edu (R.P.M.)

Abstract

There is an increasing need for single cell omics data to be integrated across sources, types, and features of data. We present a deep clustering algorithm that learns discriminative representation for single-cell data via maximizing mutual information, SMILE (**S**ingle-cell **M**utual **I**nformation **Le**arning). SMILE successfully integrates multi-source single-cell transcriptome data and can also integrate single-cell ATAC-seq, RNA-seq, DNA methylation, and Hi-C data generated by the latest joint profiling technologies.

Main

As increasing numbers of single-cell sequencing datasets are available, data integration becomes a key research domain for understanding a complex cellular system from different angles, but also presents a new challenge for data analysis. In single-cell transcriptomics, batch effects are usually a prominent variation when comparing data from multiple sources and removing batch effect is a critical step for revealing biologically relevant variation. Besides integrating single-cell transcriptome data, integration of multimodal single-cell data is becoming important as technological breakthroughs make it possible to capture multiple data types from the same single cell. For example, sci-CAR and SHARE-seq can simultaneously profile chromatin accessibility and gene expression for thousands of single cells.[1, 2] scMethyl-HiC and sn-m3C-seq can profile DNA methylation and 3D chromatin structure at the same time at single-cell resolution.[3, 4] However, these data types do not naturally share the same feature space: transcriptomes are described using genes as features, while chromatin accessibility is reported across all intergenic spaces. Therefore, integration of multimodal data becomes even more challenging. To address the challenges above in a single method, we designed a deep clustering model, SMILE, that learns discriminative representation for data integration in an unsupervised manner. In our approach, we restructured cells into pairs, and we aimed to maximize the similarity between the paired cells in the latent space. Because of this cell-pairing design, SMILE extends naturally into integration of multimodal single-cell data, where data from two sources (RNA-seq/ATAC-seq or Methyl/Hi-C) exist for each cell and thus form a natural pair. We demonstrated that SMILE can effectively project RNA-seq/ATAC-seq data, as well as Methyl/Hi-C data, into the same latent space and achieve data integration.

We first explain the architecture of SMILE (Fig. 1a). The main component is a multi-layer perceptron (MLP) used as an encoder that projects cells from the original feature space X to representation z. To achieve this goal, SMILE relies on maximizing mutual information between X and z. Mutual information measures the dependency of z on X. If we maximize the dependency, we can end up using low-dimension z to represent the high-dimension X. Contrastive learning is

one approach to maximize mutual information, and it usually requires pairing one sample with a positive or negative sample. Then, the goal is to maximize similarity between the positive pair and dissimilarity between the negative pair in the representation *z*. Due to the lack of labels, pairing samples is a challenging task. However, treating a sample itself as its positive sample and any other cells as negative samples can be a shortcut for reframing the data into pairs, and Chen et al. demonstrated that such a simple framework can effectively learn visual representation for images.[5] In single-cell data, we adopt the same framework to pair each cell in a dataset to itself. To prevent each pair from being completely the same, we add gaussian noise to each cell. Then, maximizing mutual information becomes forcing each cell to be like itself, regardless of the added noise, and dissimilar to any other cells. We find that learning a representation in which each noise-added cell is like its noise-added pair also leads to collective similarity of cells within the same cell-type (Fig. 1b). To implement this in the neural network, we used noise-contrastive estimation (NCE) as the core loss function to guide the neural network to learn.[6] We did not directly apply NCE on representation *z*, but further reduced *z* to a 32-dimension output and *K* pseudo cell-type probabilities, by stacking two independent one-layer MLPs onto the encoder. A one-layer MLP generating a 32-dimension vector will produce rectified linear unit (ReLU) activated output, and the other will produce probabilities of pseudo cell-types with SoftMax activation. Finally, NCE was applied on the 32-dimension output and pseudo probabilities, independently. These two one-layer MLPs produce two independent outputs which both contribute to the representation learning of the encoder. Once trained, the encoder serves as a feature extractor that projects data from the original space *X* to a low-dimension representation *z*. Further implementation detail about NCE and the neural network can be found in **Methods**.

It is now common to find multiple single-cell transcriptomics datasets for the same tissue or biological system generated by different techniques or research groups. A standard clustering analysis often fails to identify cell-types, but instead only detects differences between experimental batches. In contrast, SMILE directly learns a representation that is not confounded by batch effect and can be combined with common clustering methods for cell-type identification. We tested batch-effect correction in human pancreas data, human peripheral blood mononuclear cell (PBMC) data, and human heart data.[7-14] In this study, we compared SMILE with Harmony, which has been reported as a top method for batch-effect correction.[15, 16] We showed that SMILE has comparable performance to Harmony in these 3 systems (Fig. 1b-d) and both removed batch effects and recovered cell-types identified in original reports (Fig. 1e). Meanwhile, the integrated representations learned by SMILE and Harmony are classification-friendly for label transferring (Fig. 1f). One advantage of SMILE in contrast to Harmony is that SMILE can perform batch-effect correction without needing explicit batch information to be provided. This is useful when batch information is complicated by other differences, as is the case for data collected from different tissues: these experiments are different both biologically and in technical batch effects. By applying SMILE to 6 tissues in Human Cell Landscape, we show that SMILE can find a common representation for cell types that are conserved across these tissues (Fig. 1g and Supplementary Fig. 1).[17] This is similar to an ability recently reported for the MARS approach.[18]

To apply SMILE to joint profiling data, we modified SMILE into new architectures, pSMILE (Supplementary Fig. 2a) and mpSMILE (Fig. 2a). These new architectures contain two separate encoders (Encoder A and Encoder B). Encoder A projects RNA-seq or Methyl data into representation z^a, while Encoder B handles projection of ATAC-seq or Hi-C into representation z^b. We aim to learn z^a and z^b that will be confined in the same latent space, so we apply the same one-layer MLPs to each that further reduce them into 32-dimension output and probabilities of *K*

pseudo cell-types. Using the RNA/ATAC- or Methyl/Hi-C-joint data to train SMILE would be the same as self-paired data, except that we introduced two separate encoders. For pSMILE, we tested its performance on a simulated joint single-cell transcriptome dataset and two joint profiling datasets generated by SNARE-seq and sci-CAR (Supplementary Fig. 2b-c and Supplementary Fig. 3).[1, 19] Our pSMILE results indicate that RNA-seq data have a greater cell-type discriminative power than ATAC-seq, so we created the further modified mpSMILE with two duplicated Encoder As that share the same weights and are used to amplify the effect of RNA-seq datasets (Supplementary Fig. 4). In projecting joint data into the same space, SMILE accomplishes a similar purpose as Seurat_v3, but with more flexibility of input. Seurat_v3 implements canonical correlation analysis (CCA) to project both RNA-seq and ATAC-seq data into the same low-dimension space.[20] The use of CCA requires the two datasets to share the same features. In contrast, SMILE can work with datasets where the two data types involve entirely different features (e.g. genes vs. genomic bins). We applied mpSMILE to mouse skin data generated by SHARE-seq.[2] Since cell-type information in the original report is not accessible, we used the RNA-seq data to cluster cells into 21 groups, following the standard Scanpy procedure.[21] We then visualized mpSMILE results by UMAP. mpSMILE was able to project both ATAC-seq and RNA-seq data into the same space while discriminating between cell types (Fig. 2b and Supplementary Fig. 5). Compared to Seurat_v3, SMILE has higher F1 scores of transferring RNA-seq label to ATAC-seq and better multimodal integration in terms of silhouette score (Fig. 2c). To evaluate which biological factors drive the co-embedding we observe, we can set candidate genes from the dataset to zero and pass this altered data through the mpSMILE encoder to evaluate whether co-embedding is disrupted. Indeed, when we remove key differential genes, clusters are greatly disrupted in the co-embedding (Fig. 2d and Supplementary Fig. 6). Since this evaluation does not require re-training, this approach would allow rapid screening to detect which gene expression and chromatin accessibility features best explain the cell-type separation.

We next evaluated the applicability of SMILE to the integration of joint profiled Methylation and Hi-C data of mESC and NMuMG cells, and the human prefrontal cortex (PFC). Unlike integration of RNA-seq and ATAC-seq, it is difficult to match DNA methylation features to chromatin structure features. Therefore, using CCA for integration of Methyl and Hi-C as in Seurat would be a challenging task. In such a case, SMILE has the unique advantage of not requiring feature matching. We applied pSMILE in both mESC and NMuMG data and human PFC data. pSMILE can distinguish mESC and NMuMG cells but only revealed 5 major cell-types in human PFC (Fig. 1e and f). Then, we applied mpSMILE by using Methylation data in place of RNA-seq and Hi-C in place of ATAC-seq, because Methylation data recovers more distinct cell-types in Lee et al.[4] However, mpSMILE did not reveal more cell-types than pSMILE (Fig. 1f). We further projected Hi-C cells onto the tSNE space of CG methylation from the original study, but in a SMILE manner. Again, neuron sub-types did not align well with their CG methylation counterparts, though other cell-types did (Supplementary Fig. 7a). These results suggest that the Hi-C part of human PFC data lacks discriminative power to distinguish neuron sub-types as the original report showed.

For both types of joint-profiling data, the learned encoders could be used to project other single source datasets (e.g. ATAC-seq or Hi-C without paired RNA-seq or Methylation) into this same space to classify cell types and compare cell type distributions in different conditions. In summary, we demonstrated that SMILE removes batch effects across different single-cell transcriptome data as well as Harmony. We further applied SMILE to joint profiling data generated by SHARE-seq and sn-m3C-seq. SMILE saves users from extra feature engineering while integrating of multimodal single-cell data. Encoders learned by SMILE can be used to determine what biological

factors underlie the derived joint clustering and to transfer cell type labels to future related experiments.

Methods

Architecture of SMILE, p(paired)SMILE and mp(modified paired)SMILE

All three variants of SMILE have encoders as the main components for feature extraction. In SMILE, there is only one encoder. This encoder consists of three fully connected layers that have 1000, 512, and 128 nodes, respectively. Each fully connected layer is coupled with a BatchNorm layer to normalize output which is further activated by ReLu function. Different from SMILE, pSMILE and mpSMILE have another encoder that has the same structure as the one in SMILE but takes an input with different dimension. The use of two independent encoders (Encoder A and B) in pSMILE and mpSMILE is to handle inputs from two different data sources with different features, for example RNA-seq and ATAC-seq. Therefore, pSMILE and mpSMILE do not require extra feature engineering to match features for inputs from two different data sources. Modified from pSMILE, mpSMILE has a duplicated Encoder A, which shares the same weights. Using duplicated encoders takes advantage of the discriminative power of RNA-seq or Methyl to learn a more discriminative representation. This is because we observed a compromise between the source with more discriminative power and the source with less. We stack two independent fully connected layers to the encoder(s), which further reduce the representation *z* to a 32-dimension vector with ReLu activation and *K* pseudo cell-type probabilities with SoftMax activation. In our application of SMILE, we set *K* to 25, or in the case of semi-supervised learning, *K* can be set to the number of known cell types.

Cell pairing

SMILE takes paired cells as inputs. When using SMILE for integration of multi-source single-cell transcriptome data, we treat each cell itself as positive pair. To prevent the two cells in each pair from being completely the same, we add gaussian noise to differentiate them. Thus, the learning process will maximize the true similarities between the pair while minimizing the effect of noise. When using pSMILE and mpSMILE for integration of multimodal single-cell data, we pair a cell from RNA-seq/Methyl with its counterpart from ATAC-seq/Hi-C. Since joint profiling quantifies two aspects of one single cell, we know that one cell in RNA-seq/Methyl has a corresponding cell in ATAC-seq/Hi-C. When RNA-seq and ATAC-seq come from two separate studies, the user may need to pair cells of the same cell-type manually. Here, we suggest using “FindTransferAnchors” function in Seurat_v3 to generate cell pairs. Then, user can use these paired cells to train p/mpSMILE (Supplementary Fig. 8).

Loss function

a. **Noise-contrastive estimation (NCE).** The core concept of making cells resemble themselves resides in the use of NCE as the main loss function. In training, we divide a whole dataset into multiple batches, and each batch has *N* cells. For multi-source single-cell transcriptome data, we differentiate each cell into two by adding random gaussian noise. Therefore, there are $2N$ cells in one batch. In each batch, each cell has itself as positive sample and the rest of $2(N-1)$ cells as its negative samples. For joint profiling data and in an *N*-pair batch, one of *N* cells in RNA-seq/Methyl has its corresponding cell among *N* in ATAC-seq/Hi-C as the positive sample and the rest of $2(N-1)$ cells summed from both RNA-seq/Methyl and ATAC-seq/Hi-C as negative samples. Let $sim\,(u, v) =$

$u^T v/\|u\|\|v\|$ denote the dot product between L_2 normalized u and v. Then, NCE for a positive pair of examples (i, j) can be defined as (Eq. 1). In this study, we applied NCE to 32-dimension outputs (two matrices in a shape of *N* by 32) with τ as 0.05 and *K* pseudo cell-type probabilities (two matrices in a shape of *K* by *N*) with τ as 0.15.

b. **Mean squared error (MSE).** We use MSE as additional loss function in p/mpSMILE to push the representation of ATAC-seq/Hi-C to be closer to the representation of RNA-seq/Methyl in the latent space (Eq. 2). Of note, MSE alone is unable to drive the model to learn a discriminative representation (Supplementary Fig. 7b and Supplementary Fig. 9).

$Eq.1: loss_{i,j} = -\log \frac{\exp(sim(z_i, z_j)/\tau)}{\sum_{k=1\ (k\neq i)}^{2N} \exp(sim(z_i, z_k)/\tau)}$, where τ denotes a temperature parameter.

$Eq.2: loss_{MSE} = \frac{1}{N}\sum_{i=1}^{N}(z^a - z^b)$, where z^a and $z^b$are representations through Encoder A and B.

Evaluation of data integration

To evaluate batch-effect correction, we use Adjusted Rand Index (ARI) as the evaluation metric. We perform Leiden clustering to re-cluster cells using the batch-removed representation, and we compare the new clustering label with the cell-type label reported by the authors. For fair comparison of SMILE and Harmony, we use multiple resolutions to get the clustering label that has the highest ARI against authors' cell-type label, and report that ARI value as the performance of SMILE or Harmony in that data.

To evaluate integration of multimodal single-cell data, we select silhouette score as the metric. Because both SHARE-seq and sn-m3C-seq quantify two aspects for one single cell, we can use the clustering label identified through RNA-seq or cell-type label of Methyl as ground-truth for ATAC-seq or Hi-C. We, then, can calculate silhouette score to quantify how well mixed cells from these two aspects are, while distinct properties of cell-types are preserved. Meanwhile, we know each cell pair because of joint profiling. So, we also measure Euclidean distance of paired cells in the 2D UMAP space, before and after training, to show if paired cells become closer in the integrated representation.

Evaluation of label transferring

Once we have an integrated representation for multiple datasets, we can transfer labels from known data to other unknown data. To evaluate label transferring, we use weighted F1 and macro F1 scores, implemented in sklearn. In integration of multi-source single-cell transcriptome data, we select one source as the training set to train a Support Vector Machine (SVM) classifier, and test the accuracy measured through weighted and macro F1 scores in other sources. In integration of multimodal single-cell data, we have taken advantage of RNA-seq and Methyl for cell-type identification. Thus, we use RNA-seq or Methyl part data to train the SVM classifier, and we test label transferring with the ATAC-seq or Hi-C part.

Clustering of mouse skin data based on gene expression

To cluster cells based on gene expression, we use a standard pipeline implemented in Scanpy.[21] Raw gene expression count data was normalized through a 'LogNormalize' method, which normalizes the raw count for each cell by its total count, then multiplies this by a scale factor (10,000 in our analysis), and log-transforms the result. Then, we use "highly_variable_genes" function in Scanpy to find most variable genes for building a k-Nearest Neighborhood (kNN) graph.

Based on the kNN graph, we use Leiden clustering to identify cell-types with a resolution that can approximately reproduce cell clusters reported in the original study. Meanwhile, the normalized gene expression matrix with identified variable genes would be the input for SMILE.

Processing of ATAC-seq, CG methylation and Hi-C data

For ATAC-seq data at peak level, we perform TF-IDF transformation and select top 90-percentile peaks as the input for SMILE. For ATAC-seq data at gene level, we first use "CreateGeneActivityMatrix" function in Seurat_v3 to sum up all peaks that fall within a gene body and its 2,000bp upstream, and we use this new quantification to represent gene activity.[20] Then, we apply LogNormalize to gene activity matrix and find most variable genes as input for SMILE. For CG methylation data, we first calculate CG methylation level for all non-overlapping autosomal 100 kb bins across entire human genome. Then, we apply LogNormalize to the binned CG methylation data. For Hi-C data, we use scHiCluster with default setting to generate an imputed Hi-C matrix at 1MB resolution for each cell.[22] Due to the size of Hi-C matrix, we are unable to concatenate all chromosomes to get a genome-wide Hi-C matrix. Therefore, we use a dimension-reduced Hi-C data of whole genome, which is implemented in scHiCluster.

Data integration through SMILE/pSMILE/mpSMILE

We used "StandardScaler" in sklearn to scale all input data before training SMILE, except that the dimension-reduced Hi-C data has been scaled. We trained SMILE with batch size as 512 for all multi-source single-cell transcriptome data in this study, and the SMILE model can converge within 5 epochs in all cases, indicated by the total loss. In integration of SHARE-seq mouse skin data, we randomly selected half of cells as training set, and we trained mpSMILE for 20 epochs with batch size as 512. After training, we projected all cells into the integrated representation. For sn-m3C-seq data, we trained p/mpSMILE on whole data for 10 epochs with batch size as 512.

Identification of key and non-key differential genes

We performed a t-test to identify key differential genes and their ranking in both mouse skin RNA-seq and ATAC-seq data, using cell clusters that we identified through RNA-seq clustering. In both tests, we identified most genes are significantly differentially expressed (RNA-seq) or active (ATAC-seq), since these genes are also highly variable. Thus, we only selected top 10 genes in RNA-seq and top 80 genes in ATAC-seq of each cluster as key differential genes. This resulted 167 key differential genes against 419 non-key genes in RNA-seq and 927 key genes against 2356 non-key genes in ATAC-seq. For testing which features contribute to the representation learned by SMILE, key genes for each cluster were sequentially assigned values of zero and then the dataset was fed back through the encoder to determine the effect on the representation. This is conceptually similar to the motif screening used to probe a deep learning representation in Fudenberg et al. 2020.[23]

Batch removal through Harmony

We use Harmony with default setting to integrate multi-source single-cell transcriptome data, and the code can be found at https://pypi.org/project/harmony-pytorch/. Though we did not observe much difference between SMILE and Harmony, in terms of batch-effect correction, we noticed that SMILE ran much faster than Harmony, because of use of GPU computation in training SMILE (Supplementary Table 2).

Data integration through Seurat_v3

We follow a default workflow of Seurat_v3 to do the integration of SHARE-seq mouse skin data. The workflow code can be found at https://satijalab.org/seurat/v3.2/atacseq_integration_vignette.html. For a fair comparison, we did not use "FindTransferAnchors" function to find anchors. Instead, we selected random pairs of half size of the data, as we did in finding pairs for SMILE. We used these paired cells as anchors for Seurat to learn the co-embedding of RNA-seq and ATAC-seq and then projected all cells into the integrated space.

Development of SMILE

a. **Comparison of SMILE with existing deep clustering methods.** In this part of developing SMILE, we would like to address how different SMILE is from other deep learning models that have been proposed to learn representation for single-cell data. Variants of Variational Autoencoder (VAE) models, differential in the sampling approaches, were proposed to learn representation for single-cell data with batch-effect correction.[24-27] The core component of VAE is the use of reconstruction loss, but VAE also encodes a sample in the representation where this sample is drawn from a certain distribution, for example Gaussian distribution (Supplementary Fig. 10a). The use of reconstruction loss does not necessarily push representation to be discriminative, and it is also unlikely that the sampling approach based on one underlying distribution in VAE will enforce samples from different classes to have different means and variances. This led to a drawback that representation learned by these VAE-based methods may not be discriminative enough for clustering or clustering-friendly at least. Another deep clustering approach is by directly modeling clustering loss and two studies have shown the ability to learn more discriminative representation for scRNA-seq data.[28, 29] Both scDeepCluster and DESC require initiation of neural network through training an autoencoder (AE) model. However, a deep clustering tool CoSTA that we recently developed demonstrated that clustering-loss-based deep clustering can directly learn discriminative representation without an initiation through AE (Supplementary Fig. 10b).[30] The common but core concept in these methods is modeling clustering loss of K-means, which usually would assume that samples are split uniformly among the K classes. However, most single cell data are not well balanced, and the number of cells belonging to dominant cell-types may far exceed cells belonging to rare but distinct cell-types. In such a case, clustering-loss-based deep clustering is likely to fail in the identification of non-dominant cell-types. To show that the learning approach of SMILE can overcome issues above, we tested SMILE, VAE and CoSTA in a single-source human pancreas data. The number of each cell-type in this dataset is not equal. 4 major cell-types (alpha, beta, gamma, and delta) range from 266 to 2507, and 4 minor cell-types (macrophage, mast cell, epsilon cell and Schwann cell) are as low as 13. We used an encoder with the same structure in all of SMILE, VAE and CoSTA and trained these 3 models in the same manner. Doing this rules out unfavorable settings for comparison. We evaluated how clustering-friendly representations learned by these three models are by ARI and NMI, and we trained SVM classifiers to evaluate how well minor cell-types are recovered by these representations. Indeed, representation by VAE is not clustering-friendly like CoSTA (Supplementary Fig. 10c). This supports our argument that VAE is less able to learn discriminative features if the underlying data layout is not inherently class discriminative. In contrast, CoSTA lacks the ability to recover non-dominant cell-types, as shown by micro F1 score of SVM classifier (Supplementary Fig. 10d). SMILE outperforms both VAE and CoSTA in aspects of clustering-friendly

representation and recovery of minor cell types. Moreover, three methods do not have much difference on the 4 major cell-types, but SMILE shows superiority for minor cell-types.

b. **Use of SMILE in single-source single-cell omics data**. In the second part of developing SMILE, we would like show that SMILE can handle most types of single-cell data. Besides RNA-seq data shown above, we tested SMILE in ATAC-seq data from Mouse ATAC Atlas and a Hi-C data from mouse embryo cells.[31, 32] SMILE can distinguish tissue types within Mouse ATAC Atlas, and it also recovers most cell-types in the brain tissue (Supplementary Fig. 11a). For clustering single cell Hi-C data, SMILE has a slight advantage of distinguishing different cell stages than PCA (the baseline) (Supplementary Fig. 11b). However, we want to point out that, for a single source data, SMILE does not show substantial difference from a standard PCA approach. Therefore, we would not recommend users to use SMILE for single-source data.

c. **Reframing SMILE into semi-supervised learning.** When we designed SMILE, we introduced two independent one-layer MLPs, and one of them was designed to produce cell type probabilities as output. Therefore, SMILE can also be reframed into semi-supervised learning. In the third part of developing SMILE, we combined cross entropy (CE) (Eq. 3) with the other two loss functions mentioned before and examined performance of a semi-supervised SMILE. Without changing the architecture of SMILE, CE served as the loss function to supervise the RNA-seq or Methyl part, and in this study, we kept labels unknown for the ATAC-seq or Hi-C part. Because CE has no mechanism to force cells from two different sources to be confined in the same latent space, using CE can only drive the model to learn discriminative representation for the RNA-seq or Methyl part but cannot help to learn an integrated representation for joint RNA-seq/ATAC-seq and Methyl/Hi-C data. In contrast, MSE has no mechanism of learning discriminative representation but can force representation of RNA-seq/Methyl to be closer to representation of ATAC-seq/Hi-C. So, combining CE and MSE should be enough to learning discriminative and integrated representation. However, in 4 multimodal single-cell data we tested, using CE and MSE is not a guarantee of integrated representation learning, and the learned representation is also not useful for transferring RNA-seq/Methyl labels to ATAC-seq/Hi-C. However, using CE and NCE can achieve integrated representation learning, which is also useful for label transferring, supporting that NCE is the key to integration. Meanwhile, we also found combining all three losses can guide SMILE to learn discriminative and integrated representation for all 4 multimodal single-cell data, with better performance than CE+MSE and CE+NCE (Supplementary Fig. 12).

$Eq.3: loss_{CE}(x) = -\log(\frac{\exp(p_i)}{\sum_{j=1}^{K}\exp(p_j)})$, where i indicates the correct class for sample x, and p_i is the predicated probability of sample x in class i. K is the total number of classes.

Data availability

All data used in this study are publicly available with GEO accession (Supplementary Table 2). Some studies have processed data that are hosted by institution data portals (Supplementary Table 2).

Code availability

The source code of SMILE including analyses of key results in the study can be found at: https://github.com/rpmccordlab/SMILE.

Author Contributions

Y.X. conceived and developed the method with guidance from R.P.M. and produced all the figures. P.D. computationally processed raw sequencing data for input to SMILE. Y.X. and R.P.M. wrote the manuscript with input from P.D.

Acknowledgements

We thank Joseph Ecker's lab for sharing information about processing their sn-m3C-seq data. This research was supported in part by NIH NIGMS grant R35GM133557 to R.P.M.

Declaration of Interests

All authors declare no competing interests.

Figure legends

Fig. 1

Integration of multi-source single-cell transcriptome data using SMILE. a, Architecture of SMILE. Original input *X* represents a gene expression matrix where each row represents a cell, and each column stands for a gene. Random Gaussian noise is added to differentiate the input *X* into two *X*s, which are the same except for the added noise. Encoder (green) consists of three fully connected layers that projects *X* into 128-dimension representation *z*. Two independent fully connected layers (blue and grey) are stacked onto the encoder to further reduce *z* into 32-dimension output and *K* pseudo cell-type probabilities, respectively. Computing NCE using the 32-dimension output and *K* pseudo cell-type probabilities will generate two matrices. One is in the shape of *N* by *N* (*N* is the number of cells in one batch), and the other is the shape of *K* by *K*. Then gradient descending guides the model to increase values on the diagonal of the two matrices while decreasing values in the off diagonal. b-d, UMAP visualization of integrated representation of b) human pancreas data, c) human PBMC data, and d) human heart data, using raw data, or representations learned by Harmony and SMILE. Cells in upper row are colored according to their sources or batch ID, and cells in the lower row are colored by putative cell-types reported in original studies. e, Evaluation of recovery of cell-types through clustering. ARI shows how well the learned representation can recover cell-types. ARI closer to 1 indicates that the clustering labels better match original cell-type labels in that study. f, Evaluation of label transferring. SVM classifiers are trained with single source data, and then macro F1s are calculated by assigning cell types to the rest of the data sources using that classifier. g, Integration of 6 tissues of Human Cell Landscape data. Expression of two endothelial markers (PECAM1 and ICAM1) are visualized in the integrated 2D UMAP (left column). Cells are colored either by cell-type or tissue-type and visualized in the integrated 2D UMAP (right column).

Fig. 2

Integration of multimodal single-cell omics data through p/mpSMILE. a, Architecture of mpSMILE. scRNA-seq or scMethyl data are forwarded through Encoder A to produce representation z^a, and scATAC-seq or scHi-C are forwarded through Encoder B to produce representation z^b. mpSMILE has duplicated Encoder As, and cells in scRNA-seq/scMethyl part would be duplicated by adding gaussian noises and become self-pairs, as we did in SMILE for multi-source single cell

transcriptome data. Two one-layer MLPs in mpSMILE are the same as those in SMILE. b, UMAP visualization of integrated representation of mouse skin data, by Seurat_v3 and mpSMILE. Cells are colored by clusters in the left panels and colored by data types in the right panels. c, Evaluation of integration by weighted F1, macro F1 and silhouette score. For all three metrics, 1 indicates the best performance, and higher is better. d, boxplot of Euclidean distances between paired cells. Blue box: original data was forwarded through trained mpSMILE and Euclidean distances between cells in RNA-seq and their corresponding cells in ATAC-seq were measured in the integrated 2D UMAP. Orange box or green box: either key differential genes or non-key genes were suppressed to zeros, then the suppressed data was forwarded through trained mpSMILE, and Euclidean distances between cells in RNA-seq and their corresponding cells in ATAC-seq were measured in the integrated 2D UMAP. Upper panel is suppression of key gene expression, and lower panel is suppression of key gene activity. e and f, UMAP visualization of integrated representation of e) mESC and NMuMG cells and f) human PFC data, by p/mpSMILE. Cells are colored by cell-types reported by the author (left panel), or data types (right panel).

Supplementary Fig. 1

Visualization of 6 tissues of Human Cell Landscape data without integration: inputting raw data into UMAP. Expression of two endothelial markers (PECAM1 and ICAM1) are visualized in the 2D UMAP (left column). Cells are colored either by cell-types or tissue-types and visualized in the 2D UMAP (right column).

Supplementary Fig. 2

Integration of synthetic multimodal single-cell omics data through pSMILE. a, Architecture of pSMILE. scRNA-seq/scMethly would be forwarded through Encoder A to produce representation z^a, and scATAC-seq/scHi-C would be forwarded through Encoder B to produce representation z^b. Two one-layer MLPs in pSMILE are the same as those in SMILE. b, Construction of synthetic multimodal single-cell data. The synthetic multimodal data is based on a real single-cell RNA-seq data from mouse cortex.[33] Data 1 and data 2 have the same cells, and each cell in data 1 is paired with its corresponding cell in data 2. Data 1 and data 2 are generated from the original data through splitting genes into two halves. Therefore, data1 and data 2 do not share any common features. c, UMAP visualization of integrated representation of mouse cortex by pSMILE. Cells are colored by cell-types (left) and data types (right).

Supplementary Fig. 3

Integration of SNARE-seq (left) and sci-CAR (right) data through pSMILE. The first two rows show independent visualization of RNA-seq and ATAC-seq at peak level. Cells are colored by cell-types reported in original studies. The 3rd and 4th rows show the UMAP visualization of integrated representation of mixed cell line data (left) and mouse kidney data (right) by pSMILE. Cells are colored by cell-types (3rd row) and data types (4th row).

Supplementary Fig. 4

Integration of SNARE-seq and sci-CAR data through mpSMILE. a, UMAP visualization of integrated representation of mixed cell line data (upper) and mouse kidney data (lower) by mpSMILE. Cells are colored by cell-types (left) and data types (right). b, Histogram of Euclidean distance between paired cells, before and after training, based on the SMILE integrated 2D UMAP.

Supplementary Fig. 5

Integration of SHARE-seq mouse skin data through p/mpSMILE. a, Independent UMAP visualization of RNA-seq, ATAC-seq at peak level and ATAC-seq at gene level. Cells are colored by cell clusters identified through RNA-seq clustering. b, Differential gene expression and activity across 21 clusters. T-tests were performed to identify significant differential genes, independently in RNA-seq data and ATAC-seq data at gene level. Only the top-3 genes for each cluster were plotted. c, UMAP visualization of integrated representation of mouse skin data by pSMILE (upper) and mpSMILE (lower). Cells are colored by clusters (left) and data types (right). d, Histogram of Euclidean distance between paired cells, before and after training, based on the SMILE integrated 2D UMAP.

Supplementary Fig. 6

Explanation of co-embedding of RNA-seq and ATAC-seq by cluster-specific differential genes. a and b, boxplot of Euclidean distances of paired cells in 2D UMAP. Blue box: original data was forwarded through trained mpSMILE and Euclidean distances between cells in RNA-seq and their corresponding cells in ATAC-seq were measured in the integrated 2D UMAP. boxes of other colors: key differential genes that are specific to each cluster were suppressed to zeros, then the suppressed data were forwarded through trained mpSMILE, and Euclidean distances between cells in RNA-seq and their corresponding cells in ATAC-seq were measured in the integrated 2D UMAP. a) suppression of gene expression, and b) suppression of gene activity.

Supplementary Fig. 7

Integration of mESC and NMuMG data and human PFC data through pSMILE. a, Projection of Hi-C onto tSNE space of CG methylation using SMILE. We used the tSNE space of CG methylation as input for Encoder A instead of 100kb bins of CG methylation. Training SMILE in this case becomes training Encoder B to project Hi-C data into the tSNE space of CG methylation, though we visualized the integrated representation through UMAP. Top row: Hi-C and methylation data on the same graph. Second row: Same representation as above, but with CG methylation (left) and Hi-C (right) shown separately. Red circle highlights region of indistinct neuronal cell types. b, Evaluation of pSMILE with different loss functions. pSMILE was trained using MSE, NCE and NCE+MSE, respectively. Weighted F1 and macro F1 scores quantify label transferring from Methyl data to Hi-C data. Silhouette scores are based on author's cell-type label.

Supplementary Fig. 8

Integration of multimodal PBMC data through Seurat and pSMILE. a, UMAP visualization of integrated representation of PBMC data, by Seurat_v3 and pSMILE. Cells are colored by cell-types predicted by Seurat_v3 (left panel) and colored by data types (right panel). b, Comparison of Seurat_v3 and pSMILE in label transferring and data integration. Calculations of weighted F1, macro F1 and silhouette score are based on Seurat_v3 predicted cell-types as the ground truth.

Supplementary Fig. 9

Integration and evaluation of mouse skin data through mpSMILE. mpSMILE was trained using MSE, NCE and NCE+MSE, respectively. Cells are colored by cell-types and data types. Silhouette scores are based on cell-type identified through RNA-seq clustering.

Supplementary Fig. 10

Comparison of SMILE, VAE and CoSTA. a, Architectures of VAE and CoSTA. Encoders in VAE and CoSTA have the same structure as the encoder in SMILE. b, UMAP visualization of representation of a single-source human pancreas data and numerical comparison in terms of recovering major and minor cell-types.

Supplementary Fig. 11

Application of SMILE in single-cell ATAC-seq and Hi-C. a, UMAP visualization of SMILE representation of Mouse ATAC Atlas. Left panel is the visualization of whole dataset and cells are colored by tissue types. Right panel is subset from mouse brain. Cells are colored by cell-types reported by the author. b, UMAP visualization of SMILE representation of mouse embryo cell Hi-C data, and comparison of SMILE and PCA (baseline) for each different chromosome separately. Cells are colored by developmental stages. Calculation of ARI and macro F1 is based on the developmental stage of a cell as the ground truth.

Supplementary Fig. 12

Evaluation of semi-supervised learning. We evaluated semi-supervised models using three combinations of loss functions (CE+MSE, CE+NCE, and CE+NCE+MSE), in a 4 multimodal single-cell data (human PBMC, mouse kidney, mouse skin, and human PFC) and in aspects of data integration and label transferring. Semi-supervised SMILE uses CE+NCE or CE+NCE+MSE as loss function.

Supplementary Table 1

Running time of SMILE and Harmony on human pancreas, human PBMC, human heart and 6 tissues of Human Cell Landscape data.

Supplementary Table 2

Summary of data sources used in this study, including GEO accession, brief description, and link to the processed data if applicable.

REFERENCES

1. Cao, J. et al. Joint profiling of chromatin accessibility and gene expression in thousands of single cells. *Science (American Association for the Advancement of Science)* **361**, 1380-1385 (2018).
2. Ma, S. et al. Chromatin Potential Identified by Shared Single-Cell Profiling of RNA and Chromatin. *Cell (Cambridge)* **183**, 1103-1116.e1120 (2020).
3. Li, G. et al. Joint profiling of DNA methylation and chromatin architecture in single cells. *Nature methods* **16**, 991-993 (2019).
4. Lee, D.-S. et al. Simultaneous profiling of 3D genome structure and DNA methylation in single human cells. *Nature methods* **16**, 999-1006 (2019).
5. Chen, T., Kornblith, S., Norouzi, M. & Hinton, G. arXiv:2002.05709 (2020).
6. Wu, Z., Xiong, Y., Yu, S. & Lin, D. arXiv:1805.01978 (2018).
7. Grün, D. et al. De Novo Prediction of Stem Cell Identity using Single-Cell Transcriptome Data. *Cell stem cell* **19**, 266-277 (2016).
8. Muraro, Mauro J. et al. A Single-Cell Transcriptome Atlas of the Human Pancreas. *Cell systems* **3**, 385-394.e383 (2016).

9. Lawlor, N. et al. Single-cell transcriptomes identify human islet cell signatures and reveal cell-type–specific expression changes in type 2 diabetes. *Genome research* **27**, 208-222 (2017).
10. Segerstolpe, Å. et al. Single-Cell Transcriptome Profiling of Human Pancreatic Islets in Health and Type 2 Diabetes. *Cell metabolism* **24**, 593-607 (2016).
11. Baron, M. et al. A Single-Cell Transcriptomic Map of the Human and Mouse Pancreas Reveals Inter- and Intra-cell Population Structure. *Cell systems* **3**, 346-360.e344 (2016).
12. Zheng, G.X.Y. et al. Massively parallel digital transcriptional profiling of single cells. *Nature communications* **8**, 14049-14049 (2017).
13. Tucker, N.R. et al. Transcriptional and Cellular Diversity of the Human Heart. *Circulation (New York, N.Y.)* **142**, 466-482 (2020).
14. Litviňuková, M. et al. Cells of the adult human heart. *Nature* (2020).
15. Tran, H.T.N. et al. A benchmark of batch-effect correction methods for single-cell RNA sequencing data. *Genome biology* **21**, 12-12 (2020).
16. Korsunsky, I. et al. Fast, sensitive and accurate integration of single-cell data with Harmony. *Nature methods* **16**, 1289-1296 (2019).
17. Han, X. et al. Construction of a human cell landscape at single-cell level. *Nature* **581**, 303-309 (2020).
18. Brbić, M. et al. MARS: discovering novel cell types across heterogeneous single-cell experiments. *Nature methods* **17**, 1200-1206 (2020).
19. 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).
20. Stuart, T. et al. Comprehensive Integration of Single-Cell Data. *Cell* **177**, 1888-1902.e1821 (2019).
21. Wolf, F.A., Angerer, P. & Theis, F.J. SCANPY: large-scale single-cell gene expression data analysis. *Genome biology* **19**, 15-15 (2018).
22. Zhou, J. et al. Robust single-cell Hi-C clustering by convolution- and random-walk-based imputation. *Proceedings of the National Academy of Sciences - PNAS* **116**, 14011-14018 (2019).
23. Fudenberg, G., Kelley, D.R. & Pollard, K.S. Predicting 3D genome folding from DNA sequence with Akita. *Nature Methods* **17**, 1111-1117 (2020).
24. Wang, T. et al. BERMUDA: a novel deep transfer learning method for single-cell RNA sequencing batch correction reveals hidden high-resolution cellular subtypes. *Genome biology* **20**, 165-165 (2019).
25. Lopez, R., Regier, J., Cole, M.B., Jordan, M.I. & Yosef, N. Deep generative modeling for single-cell transcriptomics. *Nature methods* **15**, 1053-1058 (2018).
26. Lotfollahi, M., Wolf, F.A. & Theis, F.J. (2018).
27. Bahrami, M. et al. Deep feature extraction of single-cell transcriptomes by generative adversarial network. *Bioinformatics (Oxford, England)* (2020).
28. Tian, T., Wan, J., Song, Q. & Wei, Z. Clustering single-cell RNA-seq data with a model-based deep learning approach. *Nature Machine Intelligence* **1**, 191-198 (2019).
29. Li, X. et al. Deep learning enables accurate clustering with batch effect removal in single-cell RNA-seq analysis. *Nature communications* **11**, 2338-2338 (2020).
30. Xu, Y. & McCord, R.P. CoSTA: Unsupervised Convolutional Neural Network Learning for Spatial Transcriptomics Analysis. *bioRxiv*, 2021.2001.2012.426400 (2021).
31. Cusanovich, D.A. et al. A Single-Cell Atlas of In Vivo Mammalian Chromatin Accessibility. *Cell (Cambridge)* **174**, 1309-1324.e1318 (2018).
32. Collombet, S. et al. Parental-to-embryo switch of chromosome organization in early embryogenesis. *Nature (London)* **580**, 142-146 (2020).
33. Zeisel, A. et al. Cell types in the mouse cortex and hippocampus revealed by single-cell RNA-seq. *Science (American Association for the Advancement of Science)* **347**, 1138-1142 (2015).

Fig. 1

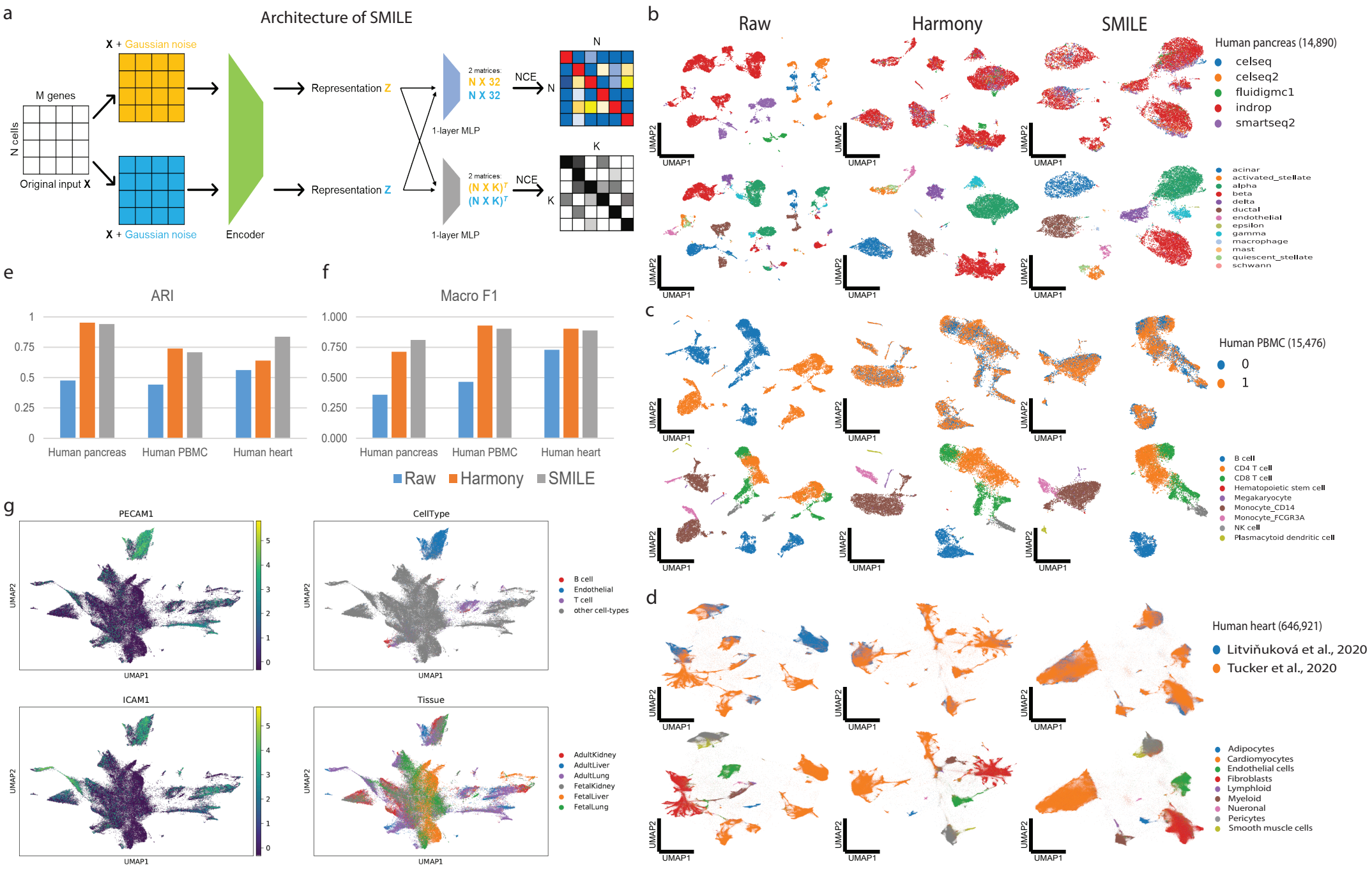
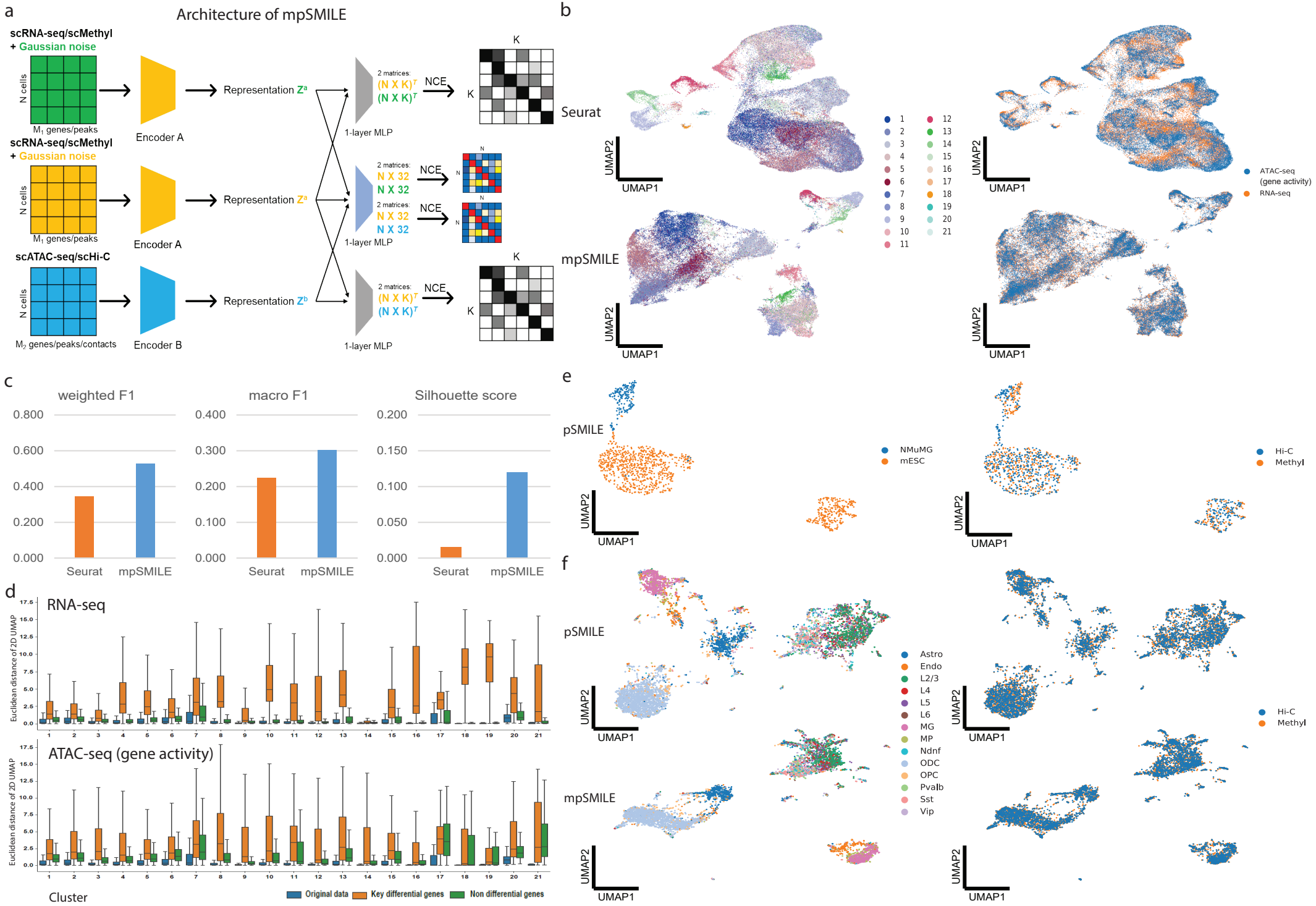

a
Architecture of SMILE
X + Gaussian noise
M genes
N cells
Original input X
X + Gaussian noise
Encoder
Representation Z
Representation Z
2 matrices:
N X 32
N X 32
1-layer MLP
2 matrices:
(N X K)^T
(N X K)^T
1-layer MLP
NCE
NCE
N
N
K
K
b
Raw
Harmony
SMILE
Human pancreas (14,890)
celseq
celseq2
fluidigmc1
indrop
smartseq2
acinar
activated_stellate
alpha
beta
delta
ductal
endothelial
epsilon
gamma
macrophage
mast
quiescent_stellate
schwann
UMAP1
UMAP2
c
Human PBMC (15,476)
0
1
B cell
CD4 T cell
CD8 T cell
Hematopoietic stem cell
Megakaryocyte
Monocyte_CD14
Monocyte_FCGR3A
NK cell
Plasmacytoid dendritic cell
d
Human heart (646,921)
Litviňuková et al., 2020
Tucker et al., 2020
Adipocytes
Cardiomyocytes
Endothelial cells
Fibroblasts
Lymphloid
Myeloid
Nueronal
Pericytes
Smooth muscle cells
e
ARI
1
0.75
0.5
0.25
0
Human pancreas
Human PBMC
Human heart
f
Macro F1
1.000
0.750
0.500
0.250
0.000
Human pancreas
Human PBMC
Human heart
Raw
Harmony
SMILE
g
PECAM1
CellType
B cell
Endothelial
T cell
other cell-types
ICAM1
Tissue
AdultKidney
AdultLiver
AdultLung
FetalKidney
FetalLiver
FetalLung
UMAP1
UMAP2

Fig. 2


a
Architecture of mpSMILE
scRNA-seq/scMethyl
+ Gaussian noise
N cells
M1 genes/peaks
Encoder A
Representation Za
scRNA-seq/scMethyl
+ Gaussian noise
N cells
M1 genes/peaks
Encoder A
Representation Za
scATAC-seq/scHi-C
N cells
M2 genes/peaks/contacts
Encoder B
Representation Zb
2 matrices:
(N X K)T
(N X K)T
NCE
1-layer MLP
2 matrices:
N X 32
N X 32
NCE
2 matrices:
N X 32
N X 32
NCE
1-layer MLP
2 matrices:
(N X K)T
(N X K)T
NCE
1-layer MLP
K
K
N
N
b
Seurat
mpSMILE
UMAP2
UMAP1
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
ATAC-seq
(gene activity)
RNA-seq
c
weighted F1
macro F1
Silhouette score
0.800
0.600
0.400
0.200
0.000
0.400
0.300
0.200
0.100
0.000
0.200
0.150
0.100
0.050
0.000
Seurat
mpSMILE
d
RNA-seq
ATAC-seq (gene activity)
Euclidean distance of 2D UMAP
Cluster
Original data
Key differential genes
Non differential genes
e
pSMILE
NMuMG
mESC
Hi-C
Methyl
f
pSMILE
mpSMILE
Astro
Endo
L2/3
L4
L5
L6
MG
MP
Ndnf
ODC
OPC
Pvalb
Sst
Vip
Hi-C
Methyl