Supplementary materials

August 12, 2017

1 Detailed model description of VASC

Suppose the input expression vector for every cell is denoted as $\boldsymbol{x} \in \mathcal{R}^d$, and we suppose this vector has been $\log_2$ transformed (with an addition of 1 to avoiding log of zeroes) and scaled to $[0, 1]$ by dividing the maximum values. Then the whole network is composed of:

- A dropout layer with a high drop-ratio: 0.5.
- Encoder network:

$$\boldsymbol{h}_1 = \boldsymbol{W_1 x} + \boldsymbol{b_1},\ \boldsymbol{h_1} \in \mathcal{R}^{512} \tag{1}$$

$$\boldsymbol{h_2} = ReLU(\boldsymbol{W_2 h_1} + \boldsymbol{b_2}),\ \boldsymbol{h_2} \in \mathcal{R}^{128} \tag{2}$$

$$\boldsymbol{h_3} = ReLU(\boldsymbol{W_3 h_2} + \boldsymbol{b_3}),\ \boldsymbol{h_3} \in \mathcal{R}^{32} \tag{3}$$

  Note we add a L_1 regularization of the first layer's weights $\boldsymbol{W_1}$. $ReLU(x) = \max(0, x)$ is applied elemenwise.

- Latent sample:

$$\boldsymbol{\mu} = \boldsymbol{W_\mu h_3} + \boldsymbol{b_\mu},\ \boldsymbol{\mu} \in \mathcal{R}^2 \tag{4}$$

$$\log \boldsymbol{\Sigma} = SoftPlus(\boldsymbol{W_\sigma h_3} + \boldsymbol{n_\sigma}),\ \log \boldsymbol{\Sigma} \in \mathcal{R}^2 \tag{5}$$

  $Softplus(x) = \log(1 + \exp(x))$ is applied elementwise. Note we set $\log \boldsymbol{\Sigma} = \boldsymbol{I}$ for datasets with small number of cells. The sampling procedure is then done by:

```
def sampling(args):
    epsilon_std = 1.0
    z_mean, z_log_var = args

    ## draw samples from standard normal distribution
    epsilon = K.random_normal(shape=K.shape(z_mean),mean=0.,stddev=epsilon_std)

    ## re-parameterization
    return z_mean + K.exp( z_log_var/2 ) * epsilon
```

- Decoder network:

$$\boldsymbol{h}_4 = ReLU(\boldsymbol{W_4 z} + \boldsymbol{b_4}),\ \boldsymbol{h_4} \in \mathcal{R}^{32} \tag{6}$$

$$\boldsymbol{h_5} = ReLU(\boldsymbol{W_5 h_4} + \boldsymbol{b_5}),\ \boldsymbol{h_5} \in \mathcal{R}^{128} \tag{7}$$

$$\boldsymbol{h_6} = ReLU(\boldsymbol{W_6 h_5} + \boldsymbol{b_6}),\ \boldsymbol{h_6} \in \mathcal{R}^{512} \tag{8}$$

$$\boldsymbol{h_7} = Sigmoid(\boldsymbol{W_7 h_6} + \boldsymbol{b_7}),\ \boldsymbol{h_7} \in \mathcal{R}^{d} \tag{9}$$

  $Sigmoid(x) = \frac{1}{1+\exp(-x)}$ is applied elementwise.

- Zero-Inflated layer: We use $\boldsymbol{p} = \exp(-\boldsymbol{h_7^2})$ as the drop probability for every element. Because the binary distribution couldn't be dealed with by back-propagation algorithm, we use Gumbel-softmax distribution to approximate it and a re-parametration trick is also applied. See details from the papaer. The concrete implrmantation is as follows:

```
def sampling_gumbel(shape,eps=1e-8):
        u = K.random_uniform( shape )
        return -K.log( -K.log(u+eps)+eps )

def compute_softmax(logits,tau):
        z = logits + sampling_gumbel( K.shape(logits) )
        return K.softmax( z / tau )

def gumbel_softmax(args):
        logits,tau = args
        return compute_softmax(logits,tau)
```

  Note for datasets with large number of cells, we use an annealing strategy for the temperature τ, which is computed (i is the number of epoches, and τ is updated every 100 epoches):

$$\tau = \min(\tau_0 \exp(-\gamma i), \tau_{min}),\ \gamma = 0.0003 \tag{10}$$

- Loss function: the whole loss function is composed of two parts and could be computed as:

```
xent_loss = d * metrics.binary_crossentropy(x, x_decoded_mean)
kl_loss = - 0.5 * K.sum(1 + z_log_var - K.square(z_mean) - K.exp(z_log_var), axis=-1)
loss = K.mean(xent_loss + kl_loss)
```

  The first term is the estimated reconstruction error (we used the binary crossentropy between two distributions defined on [0,1]). Suppose $\boldsymbol{x}, \boldsymbol{y}$ is the true and predicted values, then binary cross entropy is computed by:

$$-\boldsymbol{x}\log(sigmoid(\boldsymbol{y})) - (1-\boldsymbol{x})\log(1-sigmoid(\boldsymbol{y}))$$

The second term is the K-L divergence between $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and $\mathcal{N}(\mathbf{0}, \boldsymbol{I})$ with dimension k:

$$\mathcal{D}(\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})||\mathcal{N}(\mathbf{0}, \boldsymbol{I})) = \frac{1}{2}\left(\mathbf{tr}(\boldsymbol{\Sigma}) + \boldsymbol{\mu}^{\boldsymbol{T}}\boldsymbol{\mu} - k - \log\det\boldsymbol{\Sigma}\right)$$

2 Benchmarking

We used built-in fucntions of skelarn package from python for PCA and t-SNE. Before t-SNE, we first applied PCA to reduction the dimensions to 500 dims if datasets contain over 500 cells. The key parameter 'perplexity' of t-SNE was set as 0.2 times the number of cells. The code was shown below:

```
import numpy as np
from sklearn.decomposition import PCA
from sklearn.manifold import TSNE

## expr is a 2-D array with shape (n_cells,n_features)
pca = PCA(n_components=2).fit_transform(expr)

if expr.shape[1] > 500:
        expr_tsne = PCA(n_components=500).fit_transform(expr)
else:
        expr_tsne = np.copy(expr)
tsne = TSNE(  perplexity=0.2*n_cells  ).fit_transform(expr_tsne)
```

ZIFA package was donwloaded from `https://github.com/epierson9/ZIFA`, and we used their block algorithm:

```
from ZIFA import block_ZIFA

## expr is a 2-D array with shape (n_cells,n_features)
Z,_ = block_ZIFA.fitModel(expr,2)
```

SIMLR was installed under the construction of `https://github.com/BatzoglouLabSU/SIMLR`. We used the following R code to execute it:

```
library(SIMLR)

## data is a matrix with shape (n_features,n_cells)
## k is the true number of cell types
y <- SIMLR( data,c=k,cores.ratio = 0 )
```

```

## for larger datasets which cannot be dealed with SIMLR, we used SIMLR_Large_Scale
#y <- SIMLR_Large_Scale( data,c=k )

## Obtain the two-dimension results
ydata <- y[[4]]
```

We furtherly used built-in kmeans functions of python sklearn package for clustering analysis, and used metrics from sklearn to meature the clustering quality:

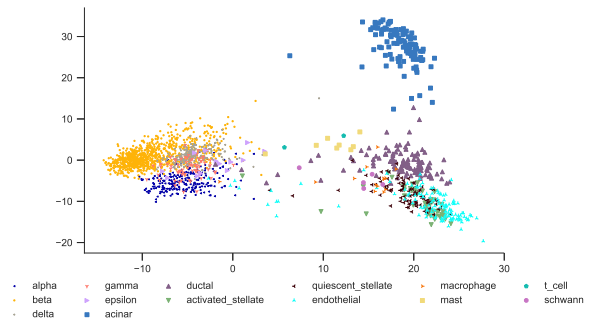
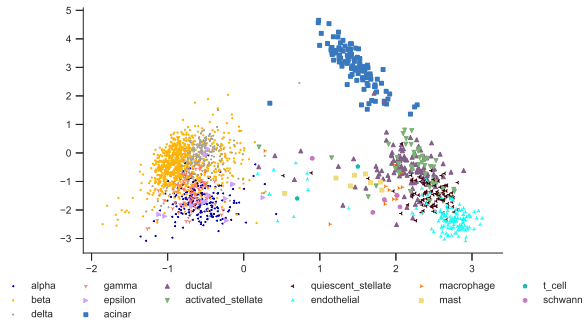
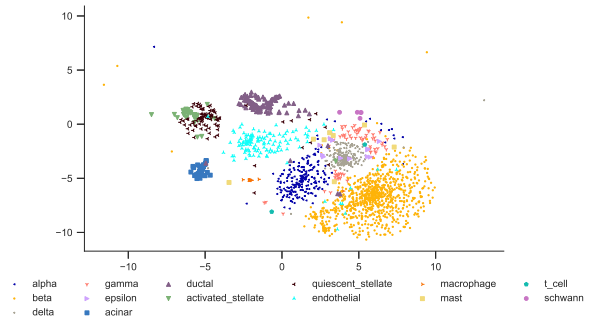
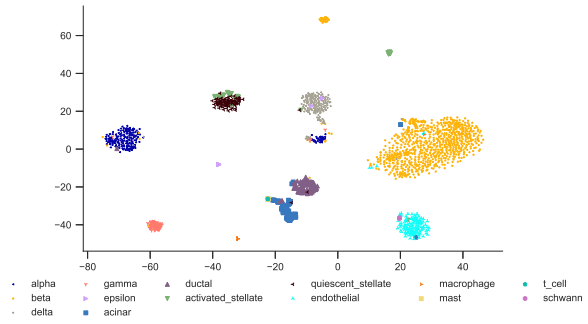
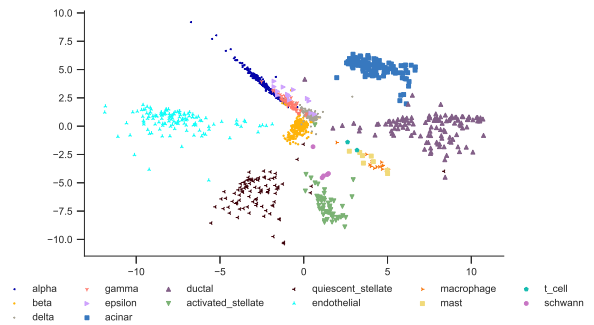
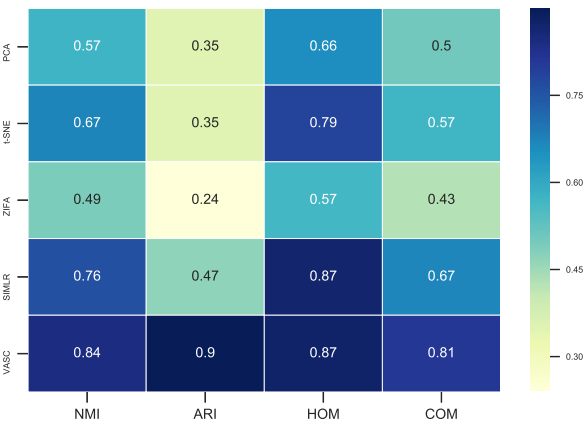
```
from sklearn.metrics import normalized_mutual_info_score,adjusted_rand_score
from sklearn.metrics import homogeneity_score,completeness_score,silhouette_score
from sklearn.cluster import KMeans
def measure( predicted,true ):
        NMI = normalized_mutual_info_score( true,predicted )
        RAND = adjusted_rand_score( true,predicted )
        HOMO = homogeneity_score( true,predicted )
        COMPLETENESS = completeness_score( true,predicted )
        return {'NMI':NMI,'RAND':RAND,'HOMOGENEITY':HOMO,'COMPLETENESS':COMPLETENESS}

## points is a 2-D array with shape (n_cells,2)
## the initialization function 'kmeans++' may raise exceptions occasionally
kmeans = KMeans( n_clusters=k,n_init=100 ).fit(points)
```

3 Visualization of scRNA-seq datasets

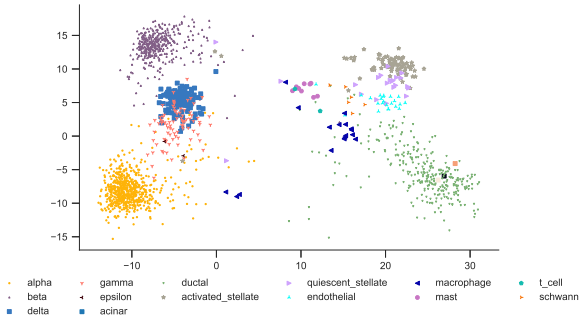
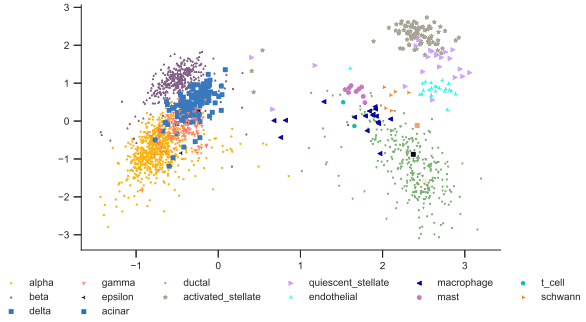
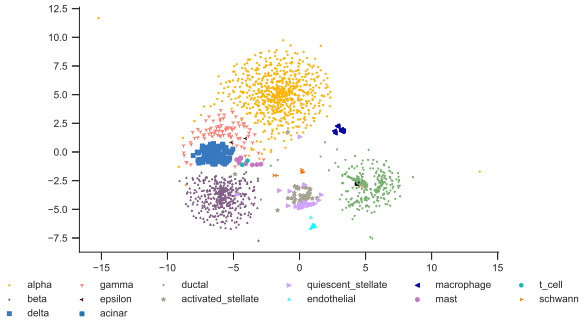
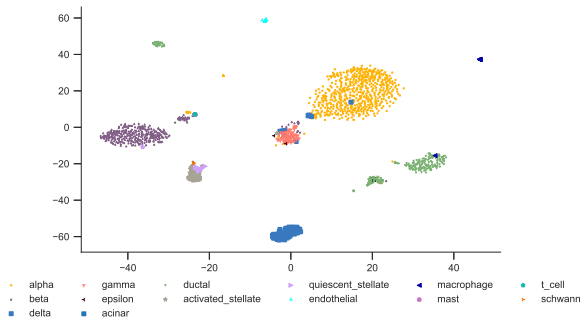
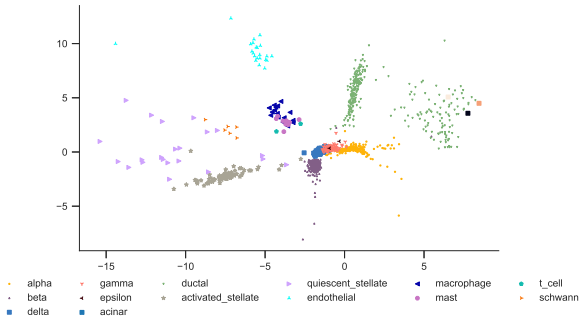
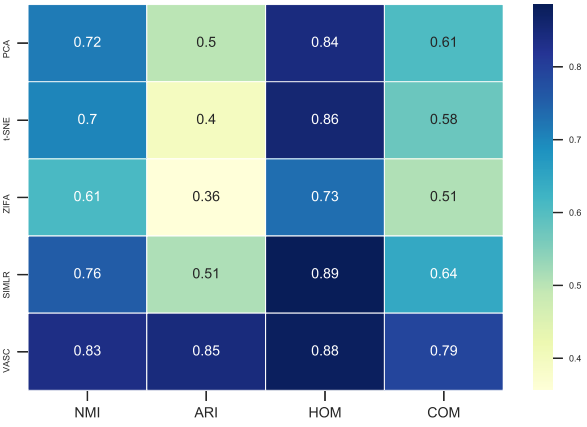
3.1 Baron datasets[1]

This dataset contains large number of cells from human and mosue pancreas. Totally, there are 4 human donors with 1937, 1724, 3605 and 1303 cells, and 2 mouses with 822 and 1,064 cells respectively.

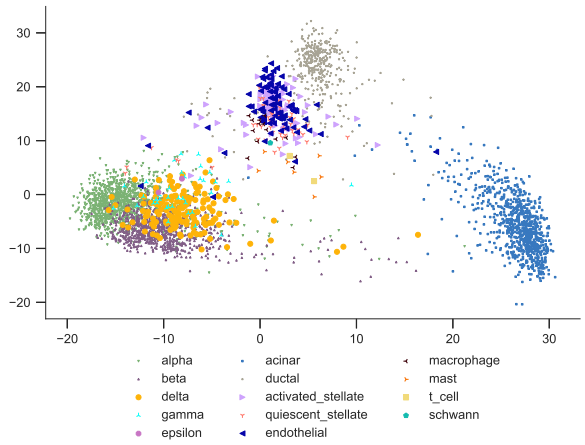
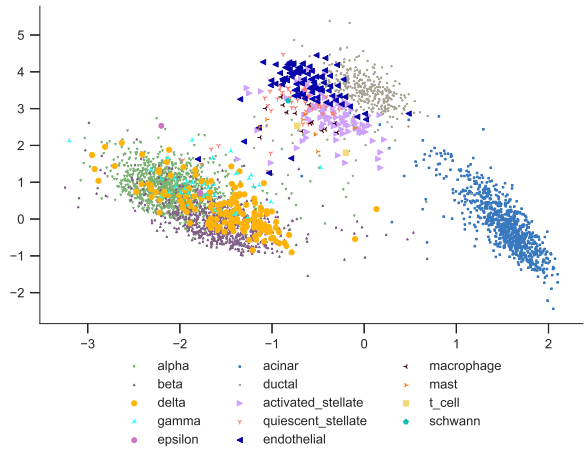
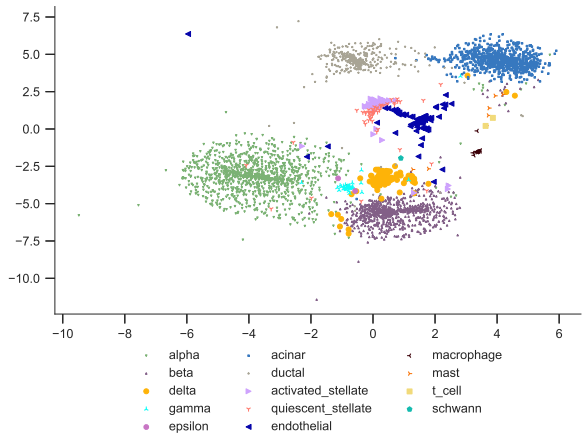
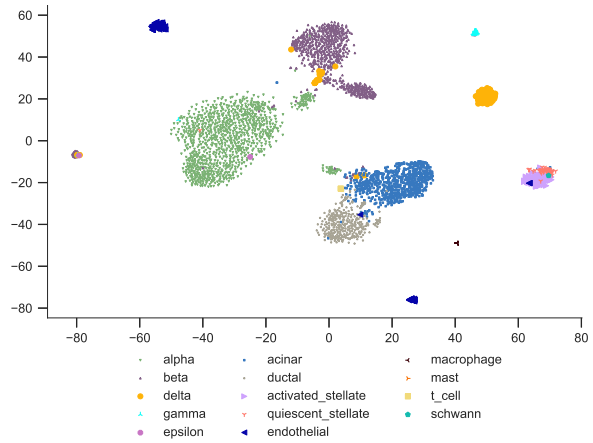
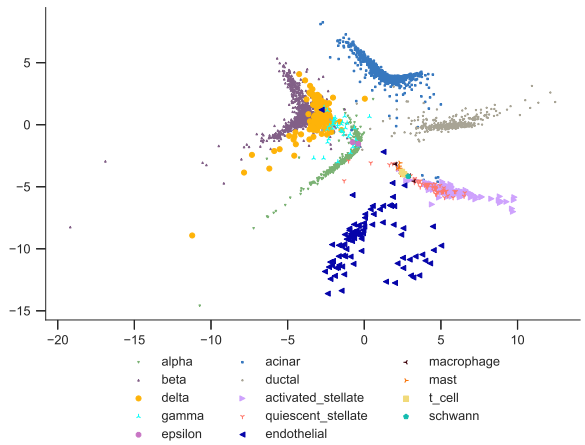
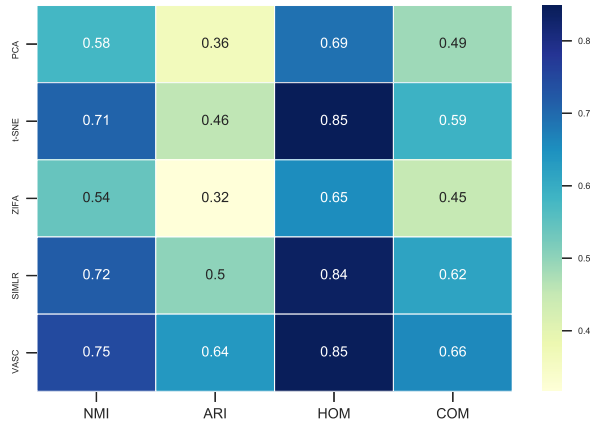
3.1.1 Baron-human-1

PCA	ZIFA
	
t-SNE	**SIMLR**
	
VASC	**clustering**
	

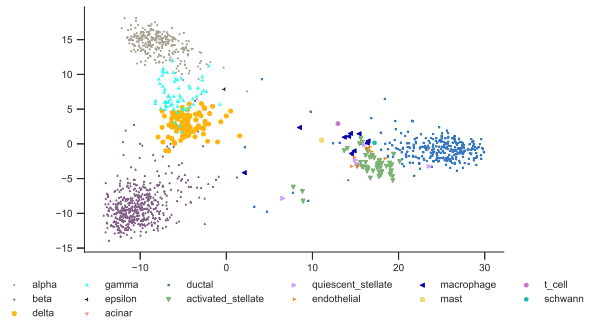
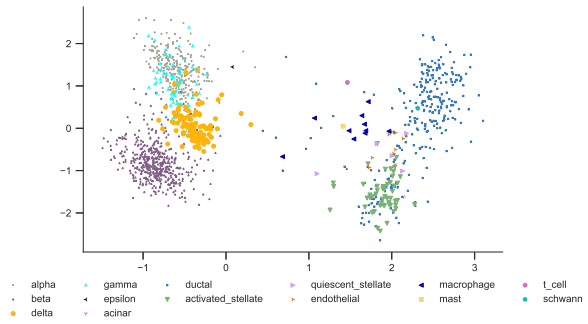
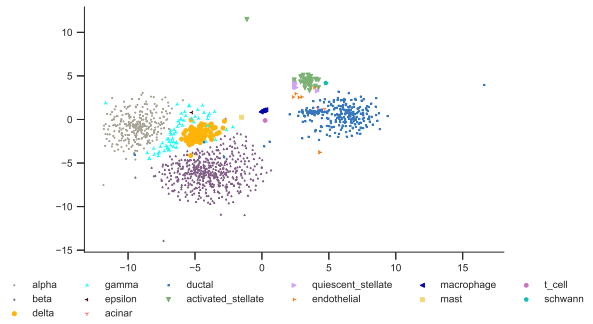
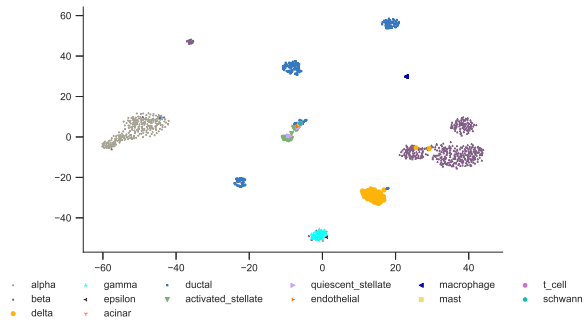
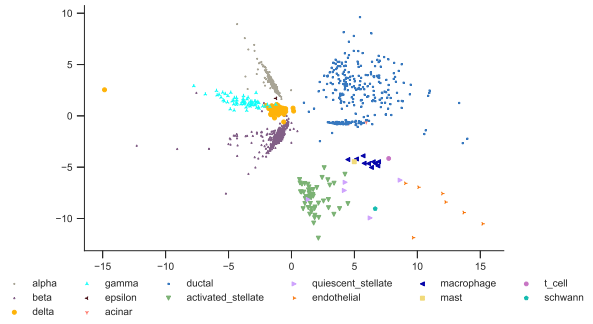
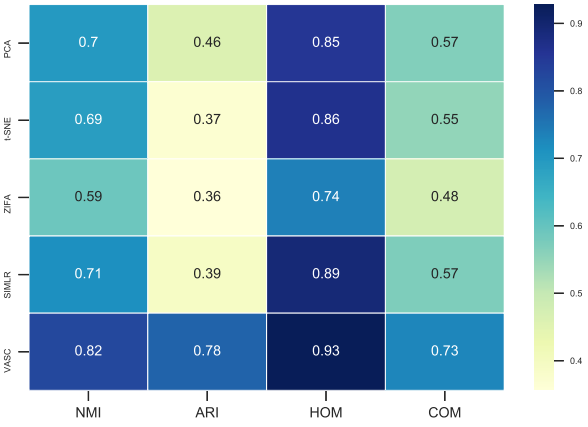
3.1.2 Baron-human-2

PCA	ZIFA
	
t-SNE	**SIMLR**
	
VASC	**clustering**
	

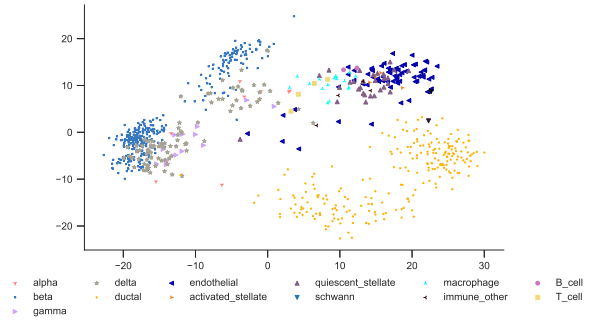
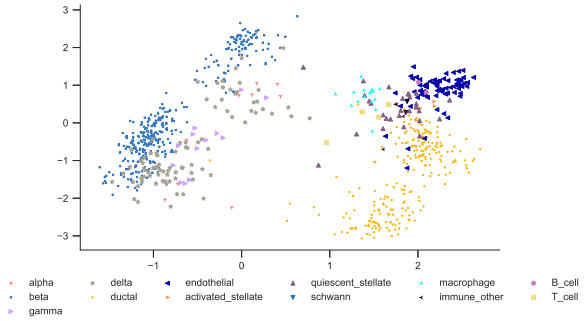
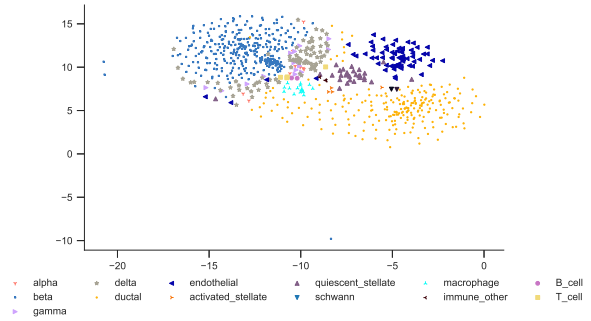
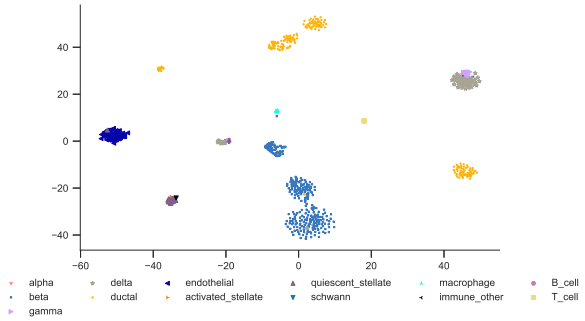
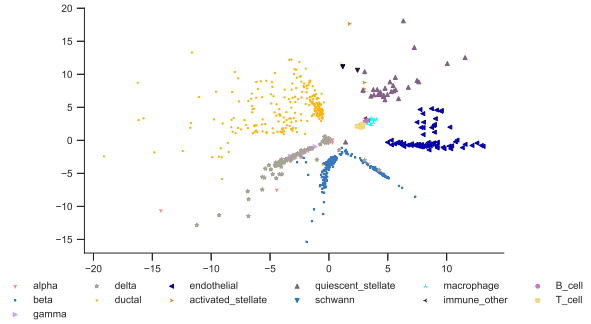
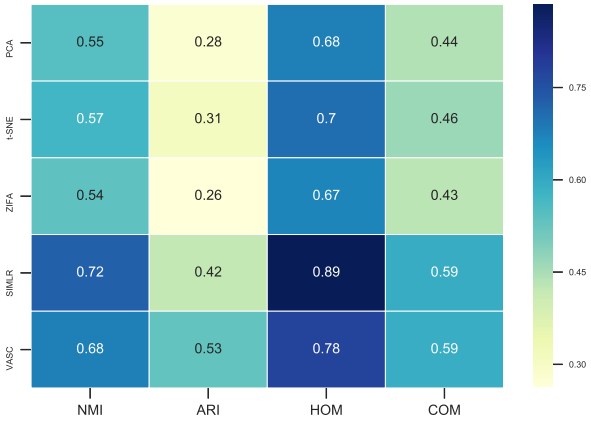
3.1.3 Baron-human-3

PCA	ZIFA
	
t-SNE	**SIMLR**
	
VASC	**clustering**
	

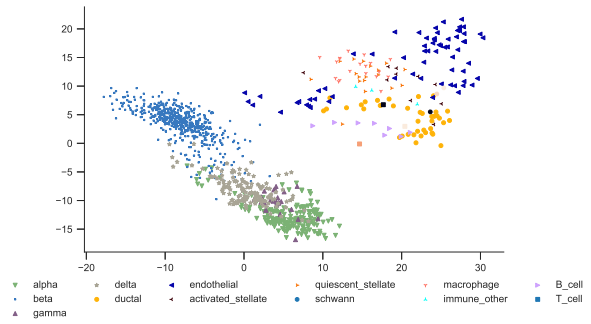
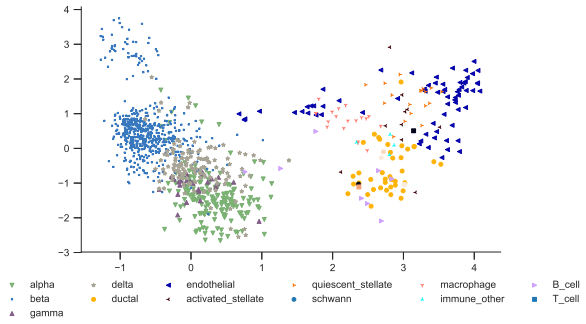
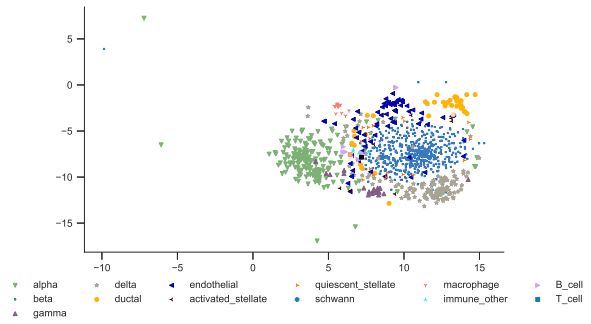
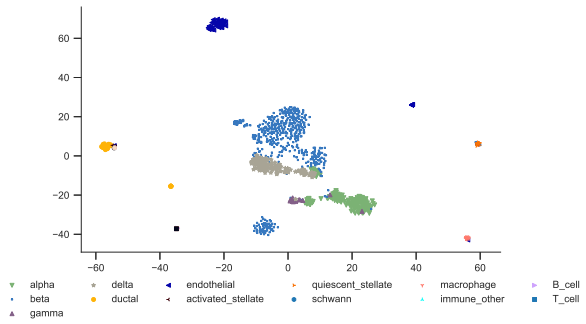
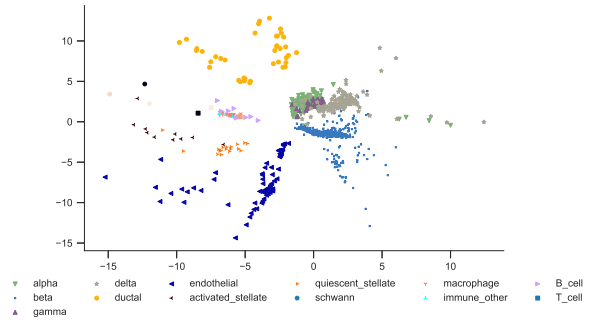
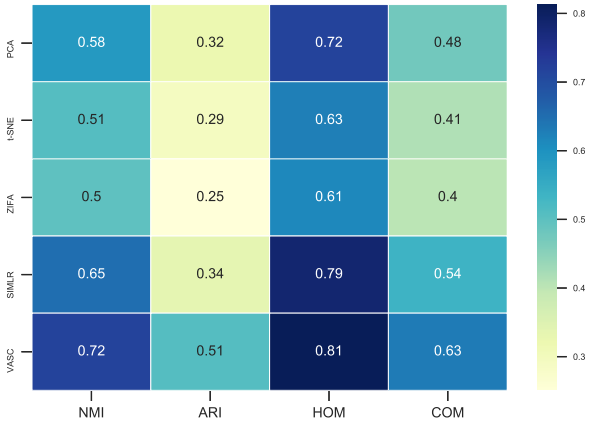
3.1.4 Baron-human-4

PCA	ZIFA
	
t-SNE	**SIMLR**
	
VASC	**clustering**
	

3.1.5 Baron-mouse-1

PCA	ZIFA
	
t-SNE	**SIMLR**
	
VASC	**clustering**
	

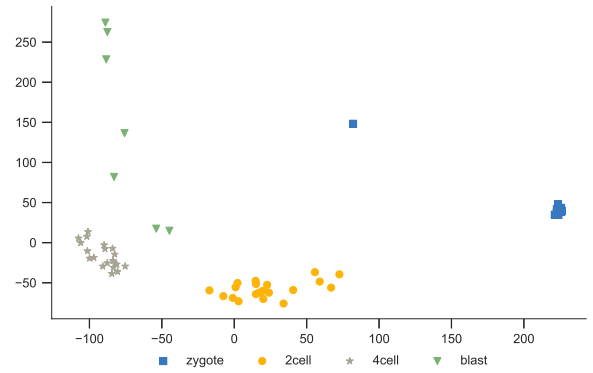
3.1.6 Baron-mouse-2

PCA	ZIFA
	
t-SNE	**SIMLR**
	
VASC	**clustering**
	

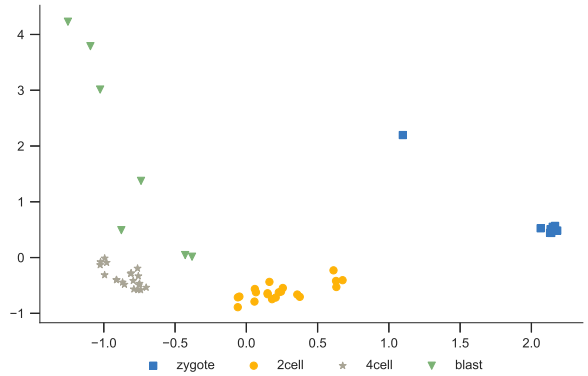
3.2 Biase dataset[2]

This dataset contains mouse embryos cells, including cell stage zygote, 2cell, 4cell and blast, with overall 56 cells.

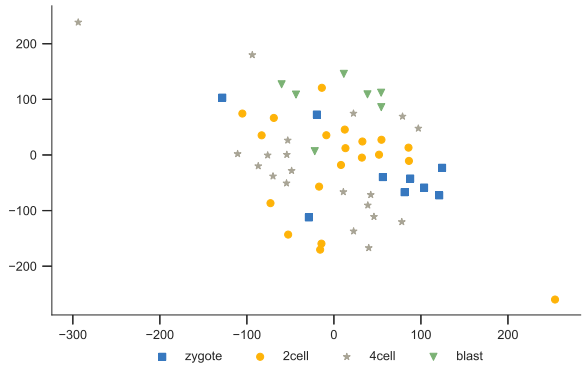
PCA



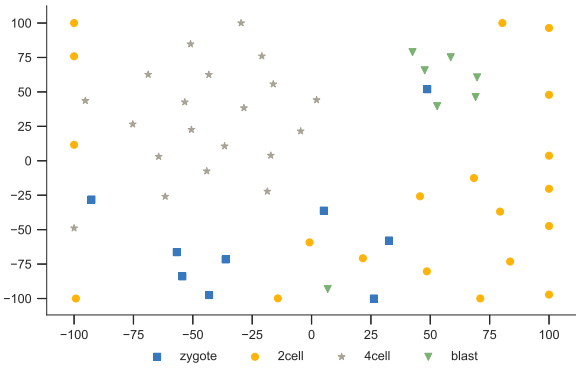
ZIFA



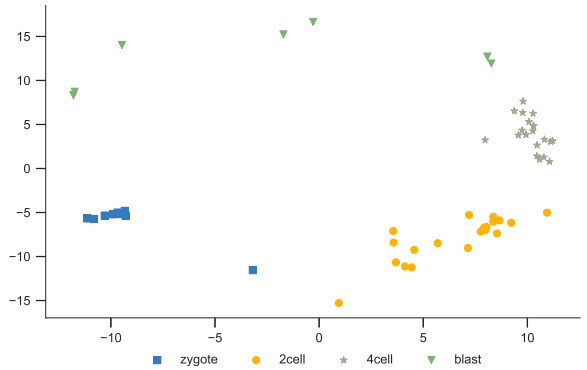
t-SNE



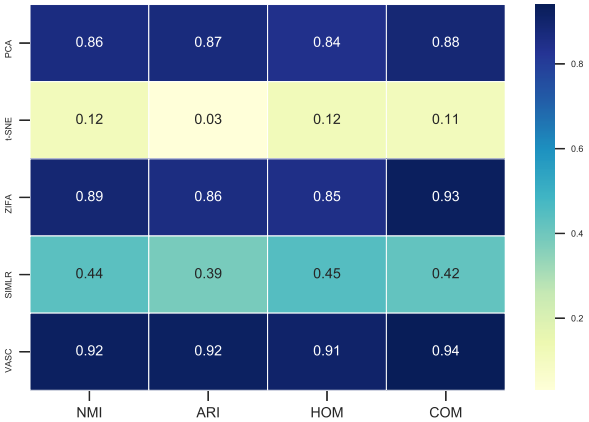
SIMLR



VASC

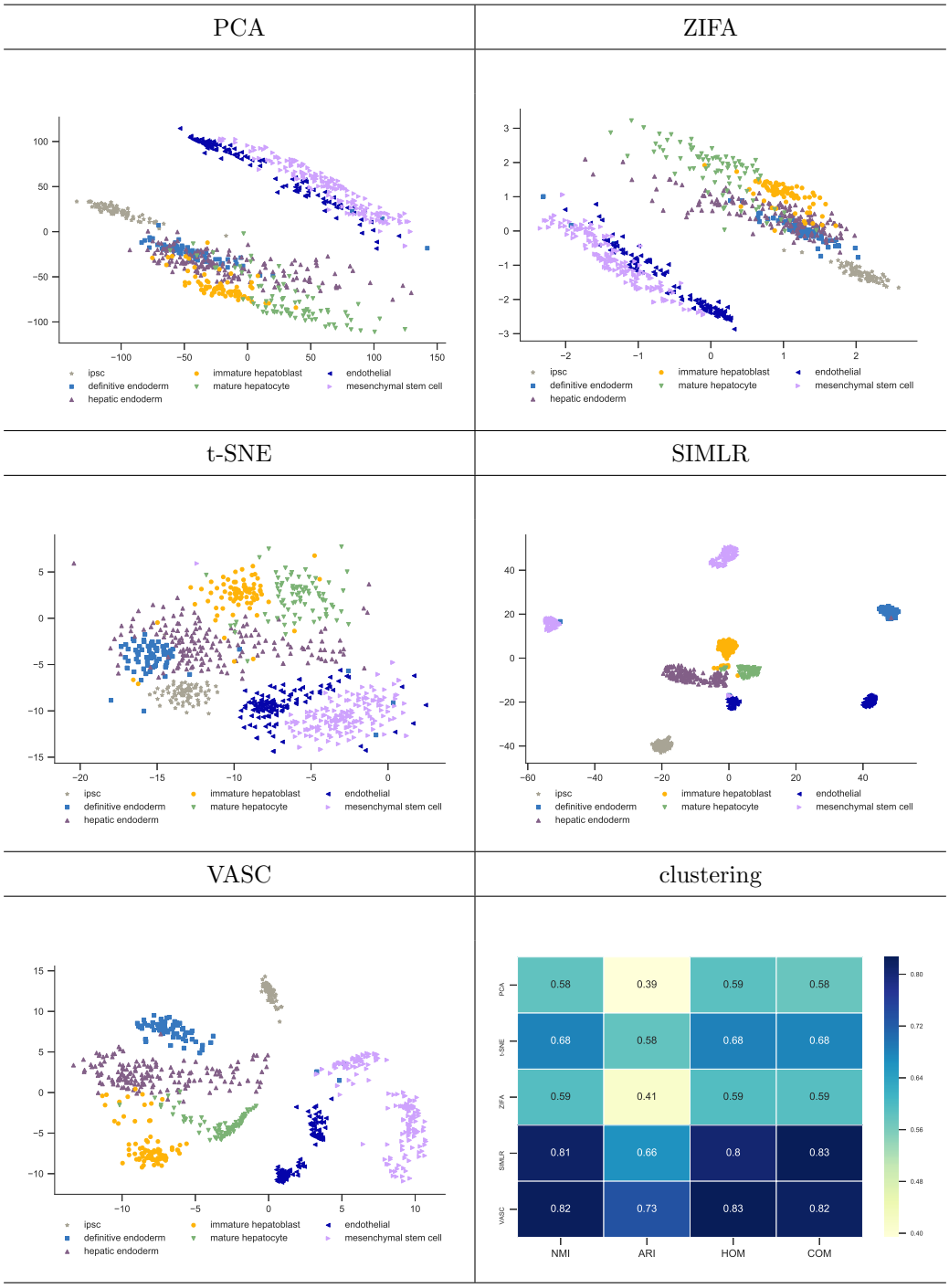


clustering



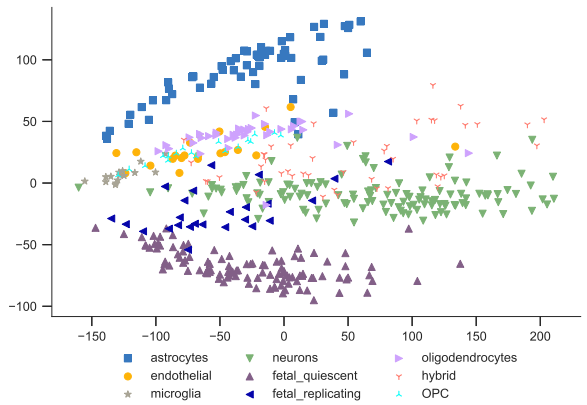
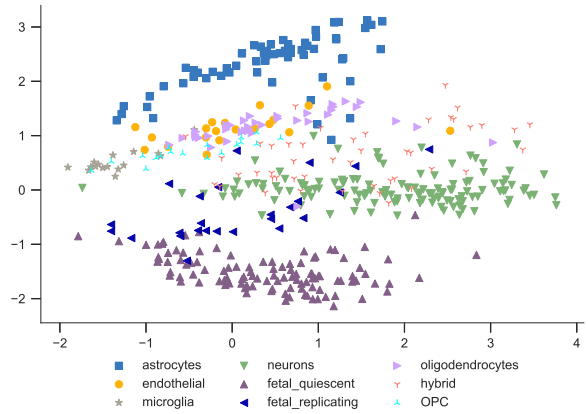
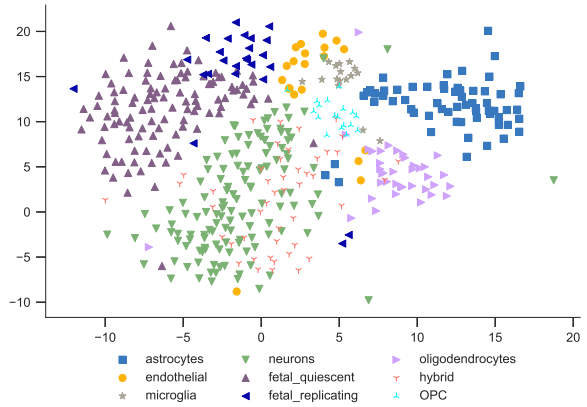
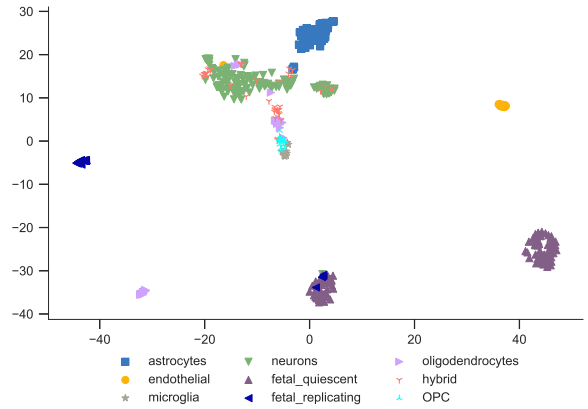
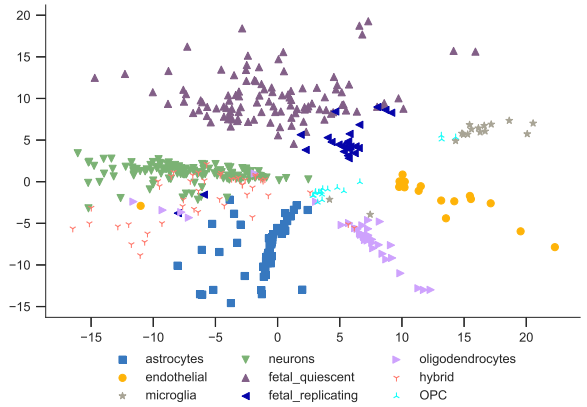
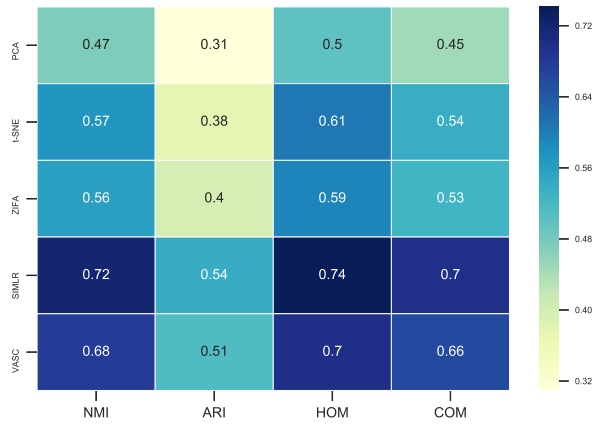
3.3 Camp dataset[3]

This dataset contains 777 cells about human liver hud development from pluripotency. These cells were sampled from different time points during hepatic cell differentiation: iPS , Definitive endoderm, Hepatic enoderm, Immature hepatoblast, Mature hepatocyte. (VASC made the same order.) Another two cell types: endothelial and mesenchymal are supportive cells.(VASC and SIMLR made two sub-populations of these cells respectively.)



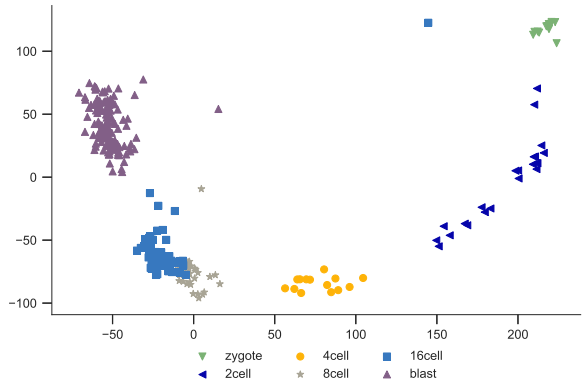
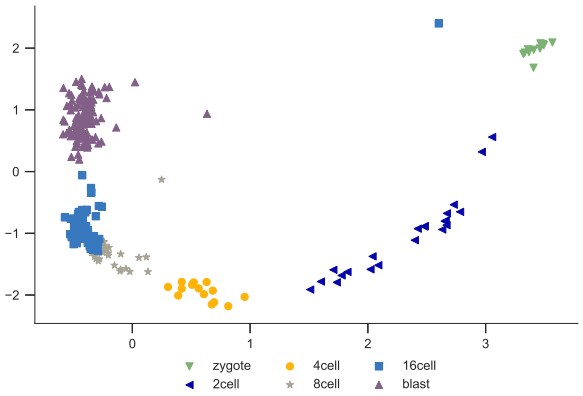
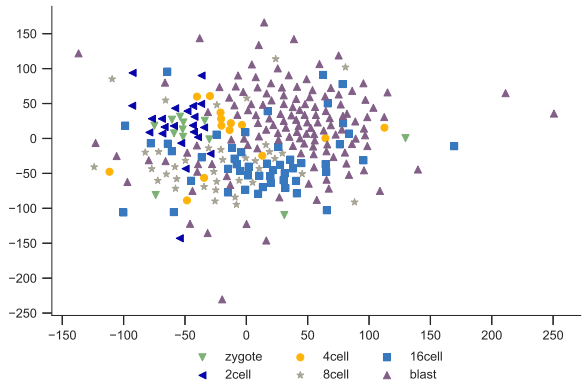
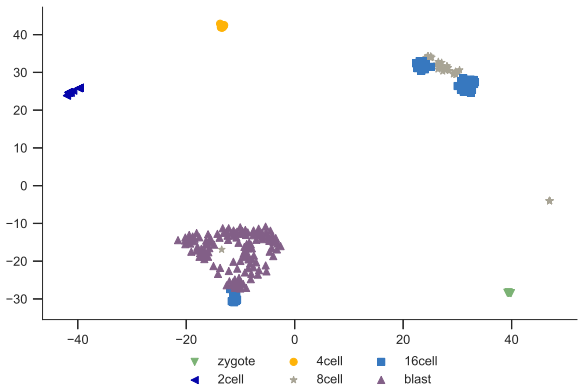
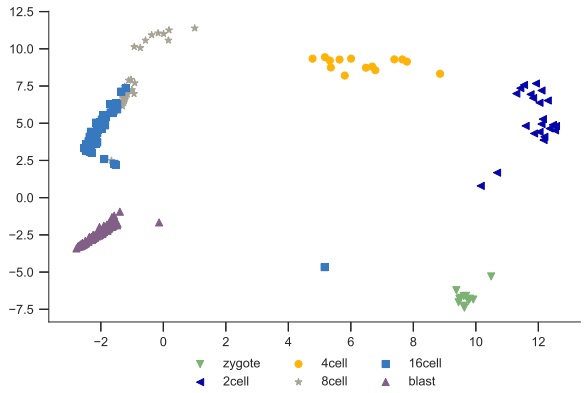
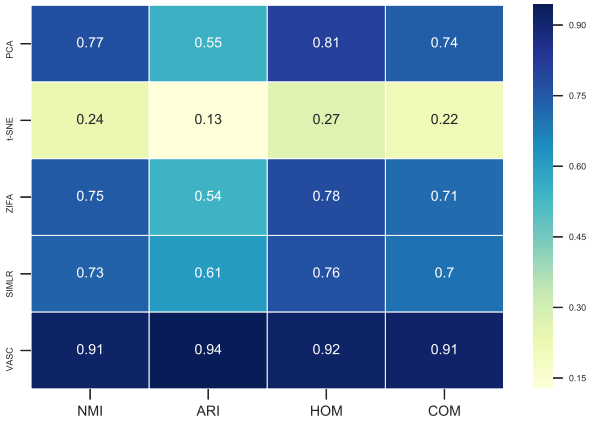
3.4 Darmanis dataset[4]

This dataset contains 466 cells from human cerebral cortex.

PCA	ZIFA
	
t-SNE	**SIMLR**
	
VASC	**clustering**
	

3.5 Deng dataset[5]

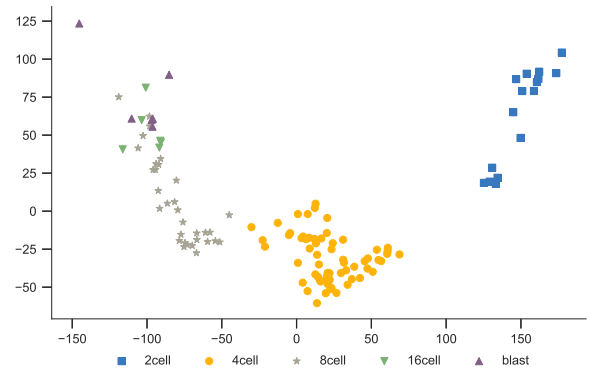
This dataset also contains mouse embryos cells, with overall 268 cells from six stages: zygote, 2cell, 4cell, 8cell, 16cell and blast.

PCA	ZIFA
	
t-SNE	**SIMLR**
	
VASC	**clustering**
	

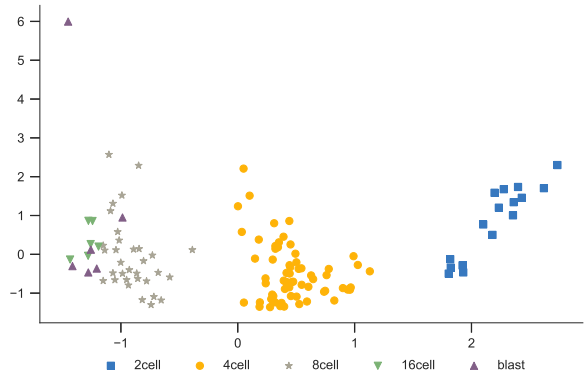
3.6 Goolam dataset[6]

This dataset also contains mouse embryos cells, including cell stage 2cell, 4cell, 16cell and blast, with overall 124 cells.

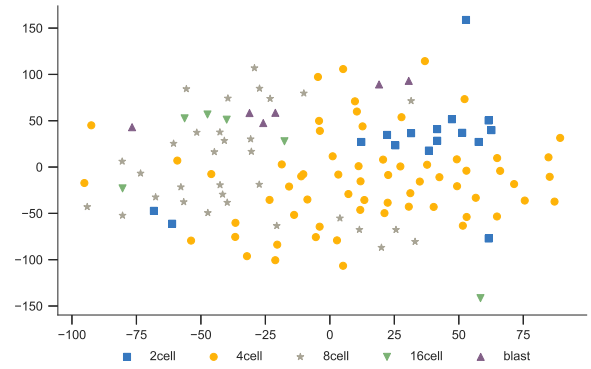
PCA



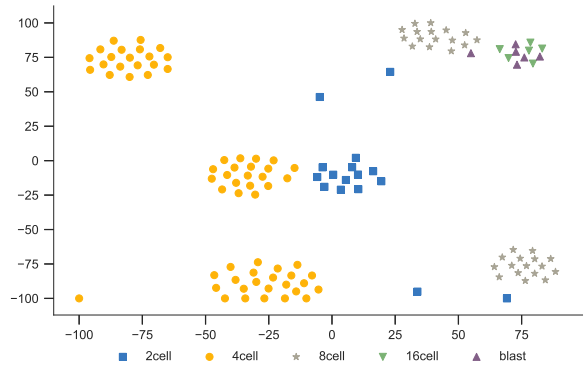
ZIFA



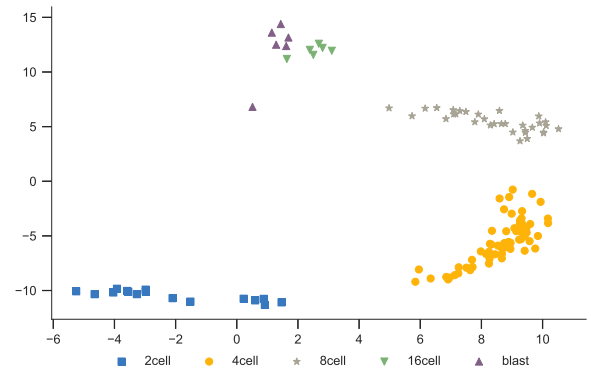
t-SNE



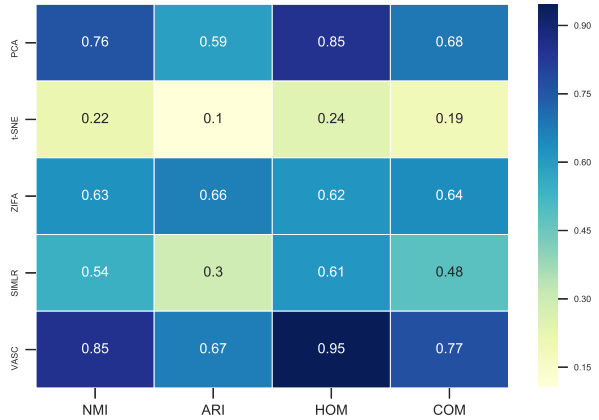
SIMLR



VASC

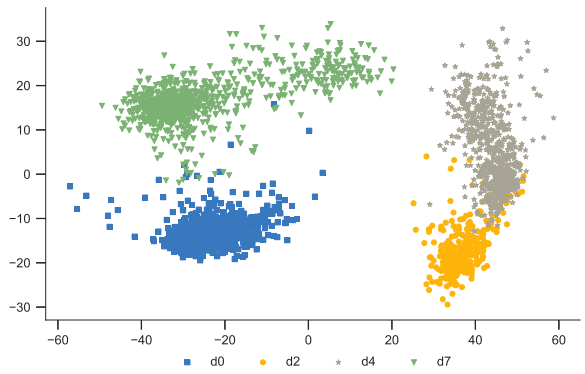
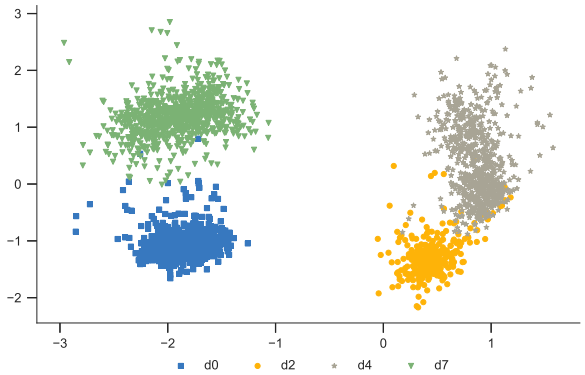
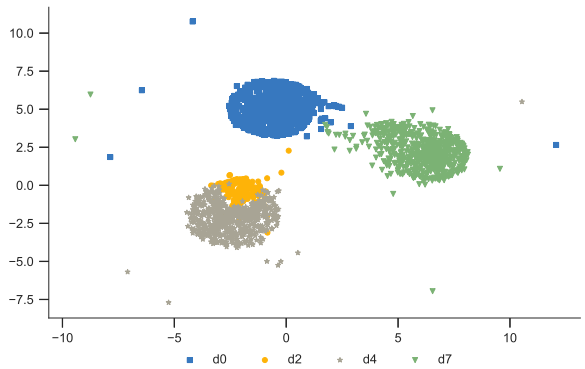
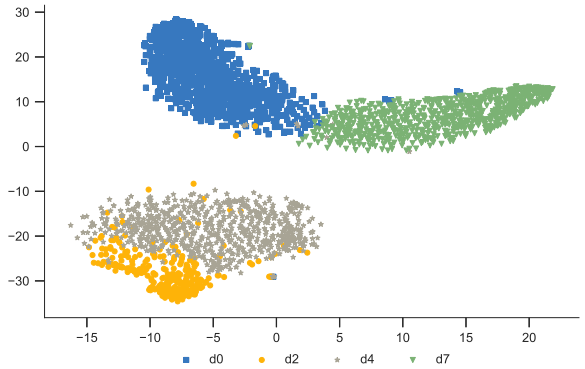
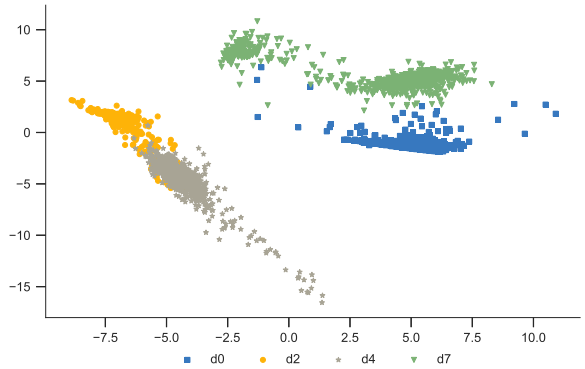
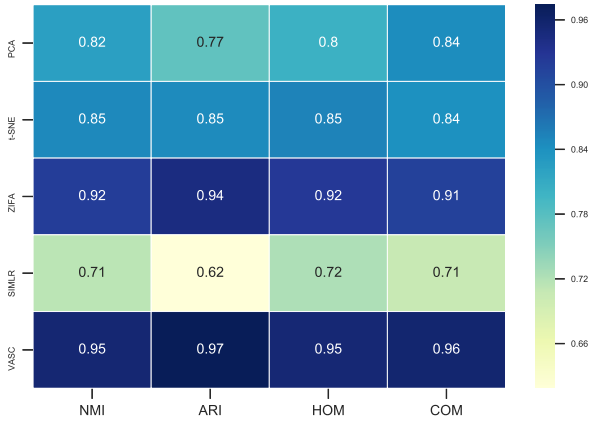


clustering



3.7 Klein dataset[7]

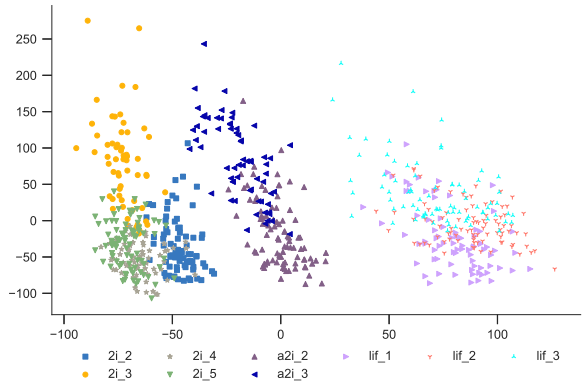
This dataset contains 2717 cells, which was sequenced by droplet barocoding, and are mouse embryonic stem cells. Our method and PCA split d7 into two parts.

PCA	ZIFA
	
t-SNE	**SIMLR**
	
VASC	**clustering**
	

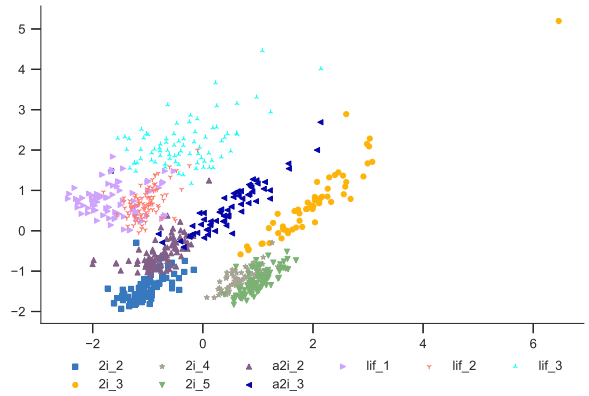
3.8 Kolodziejczyk dataset[8]

This dataset contains cells from three different embryonic stem cell culture conditions: serum, 2i and alternative groudn state 2i, However, every cell type also contains results from different chips.

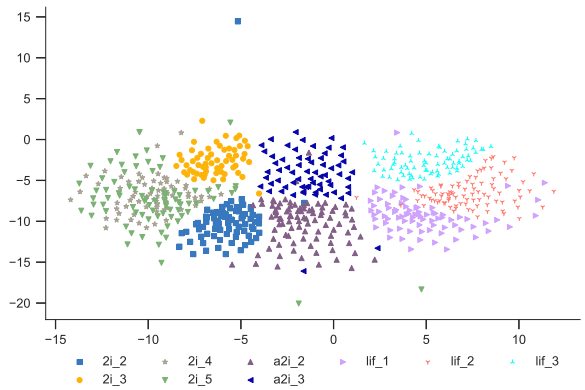
PCA



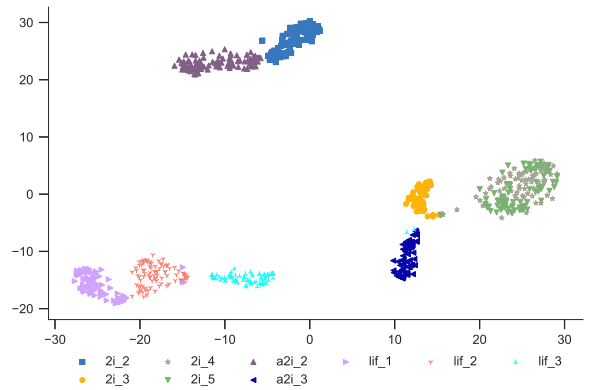
ZIFA



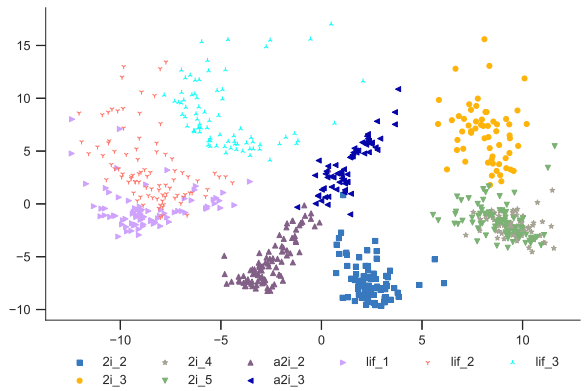
t-SNE



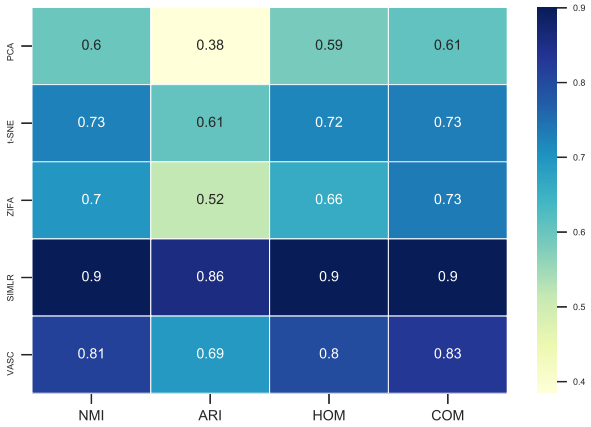
SIMLR



VASC



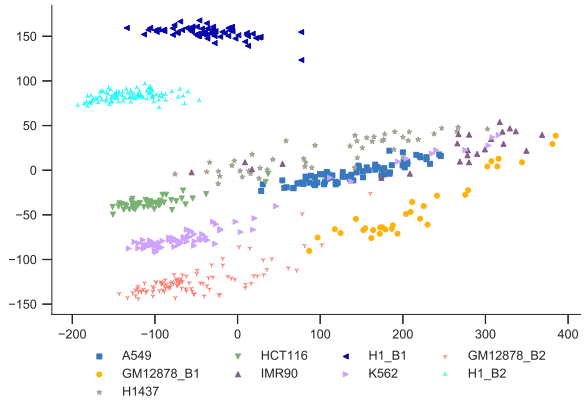
clustering



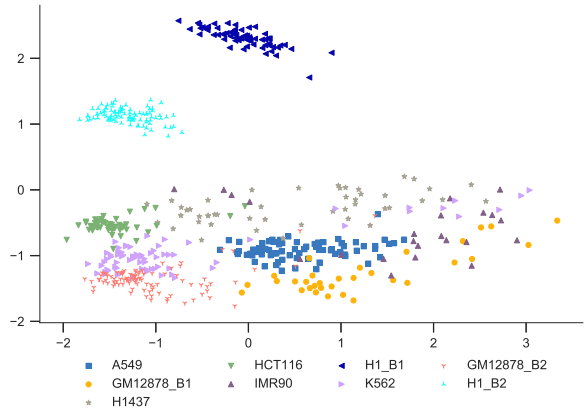
3.9 Li dataset[9]

This dataset contains 561 cells from human colorectal tumors. We found SIMLR made almost perfect split of these cells, while VASC also made comparable results.

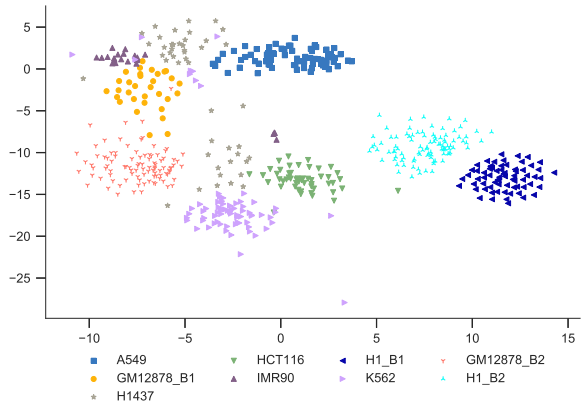
PCA



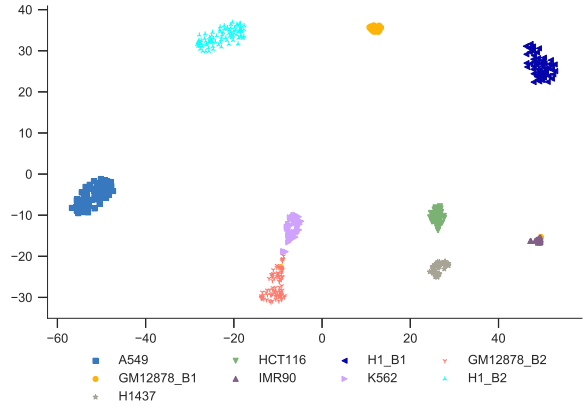
ZIFA



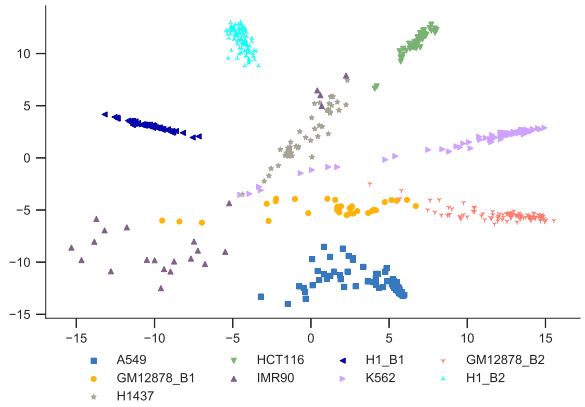
t-SNE



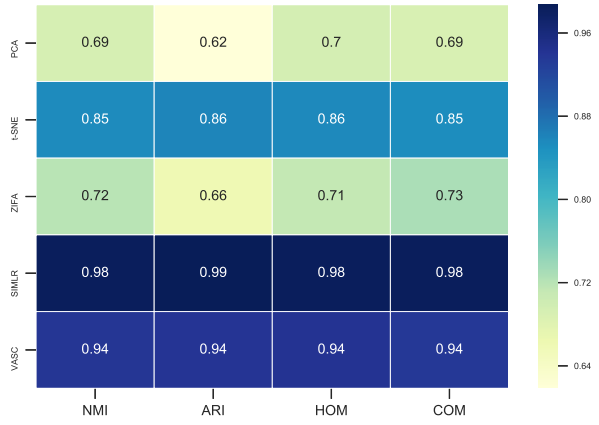
SIMLR



VASC



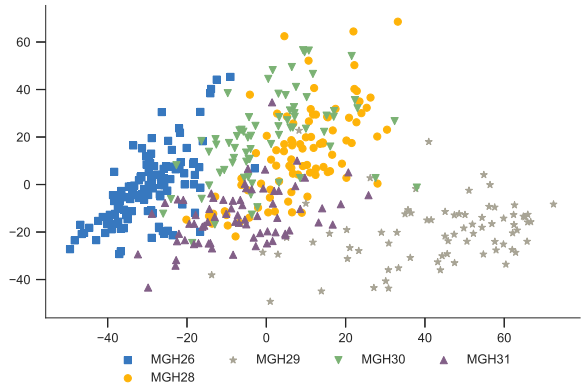
clustering



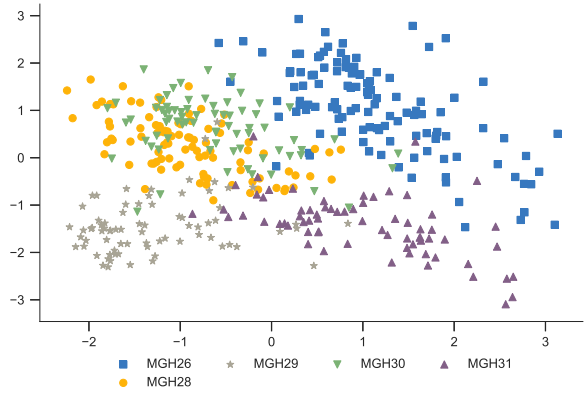
3.10 Patel dataset([10])

This dataset contains 430 cells from five primary glioblastomas. Only VASC and SIMLR could split these cells apart.

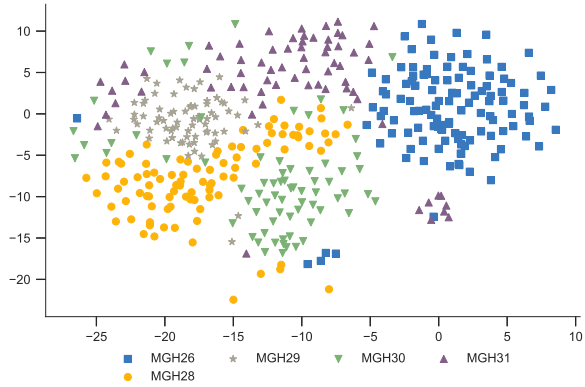
PCA



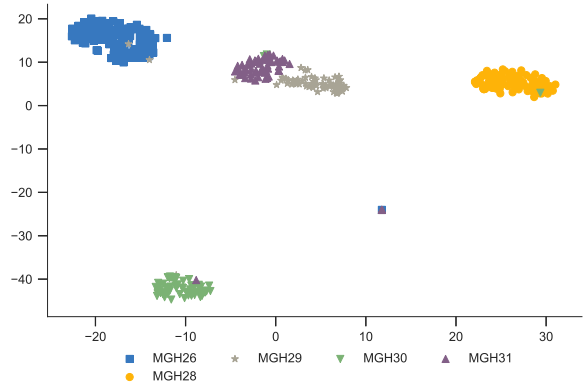
ZIFA



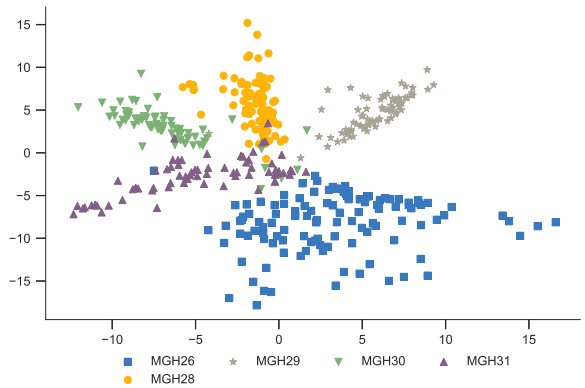
t-SNE



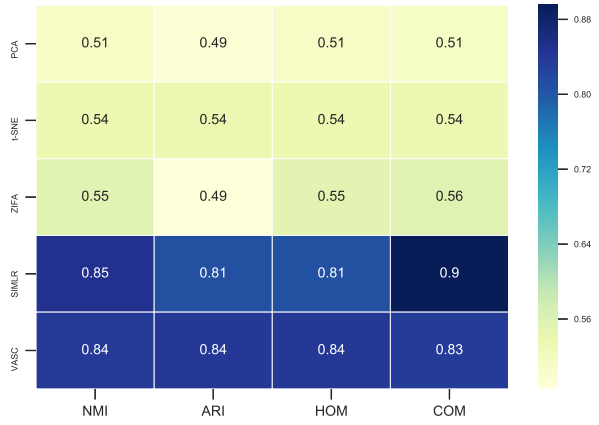
SIMLR



VASC



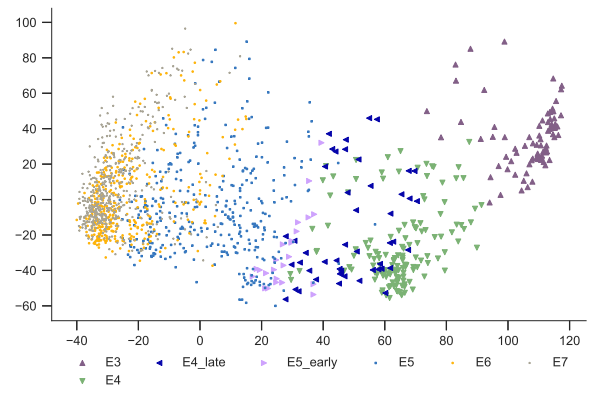
clustering



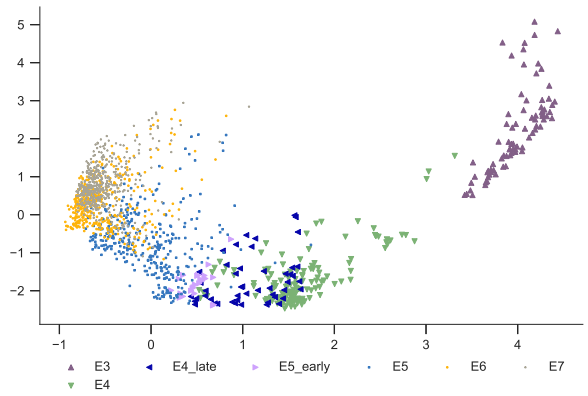
3.11 Petropoulos dataset[11]

This dataset contains human embryos cell with different development time, lineage stages. There're overall 1529 cells. **This dataset is hard to cluster because cells may vary in a continuous space.**

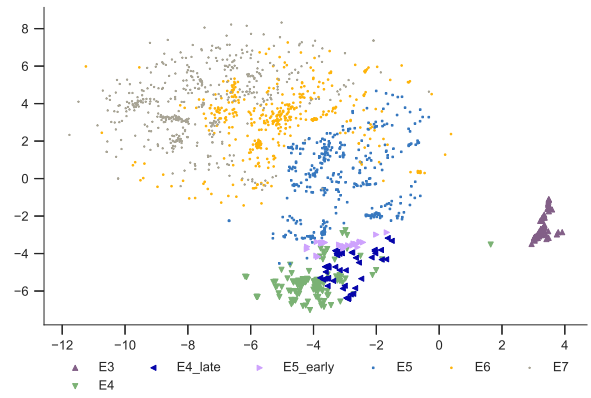
PCA



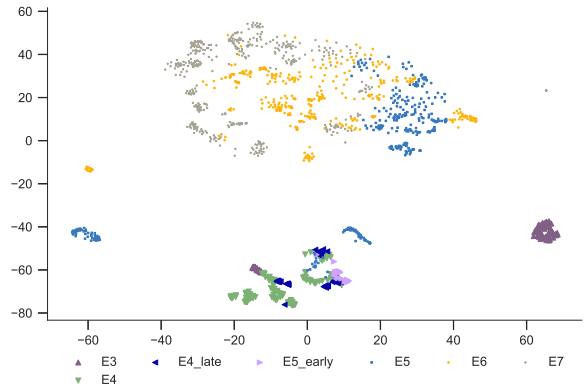
ZIFA



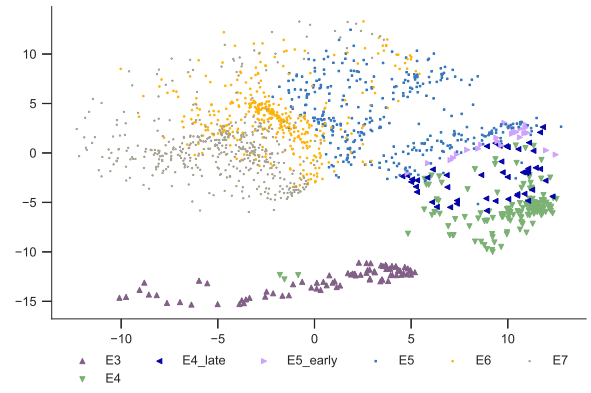
t-SNE



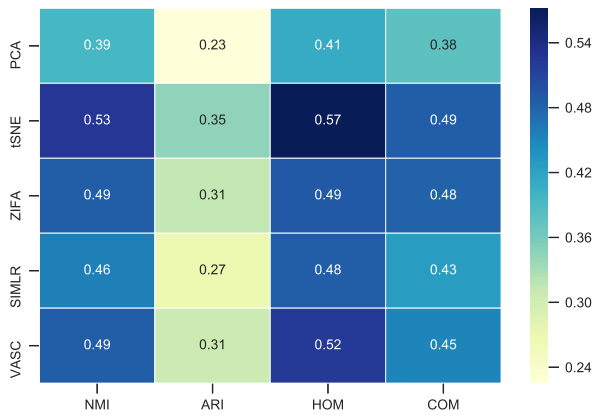
SIMLR



VASC



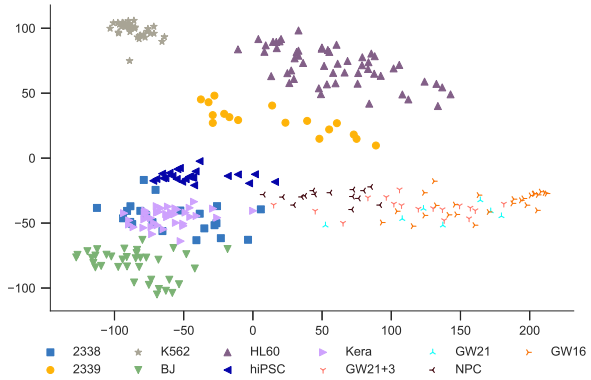
clustering



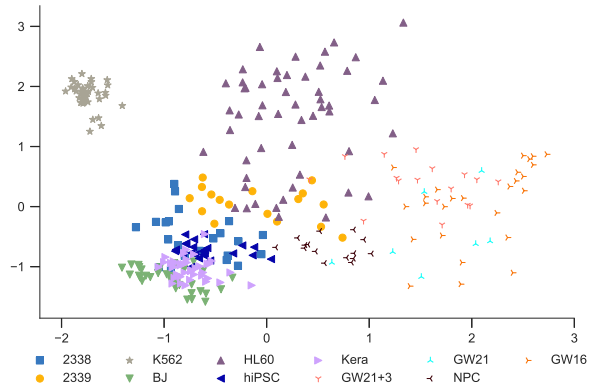
3.12 Pollen dataset[12]

This dataset contains diverse cell types, including skin cells, blood cells, pluripotent stem cells and neural cells. Overall there are 301 cells from 11 different cell types. There are 704 cells.

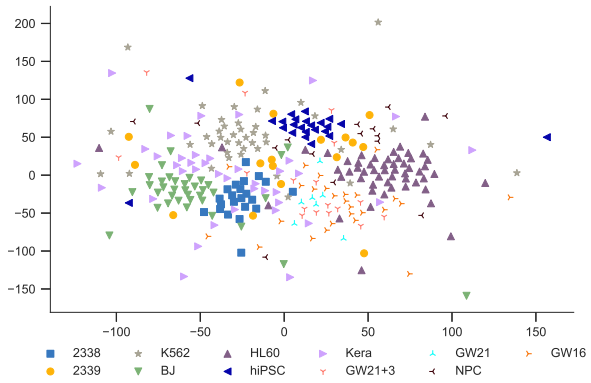
PCA



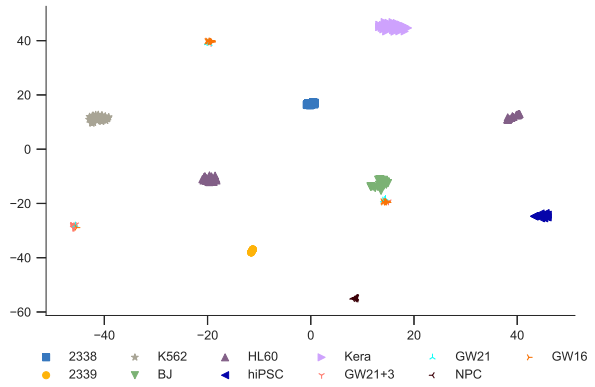
ZIFA



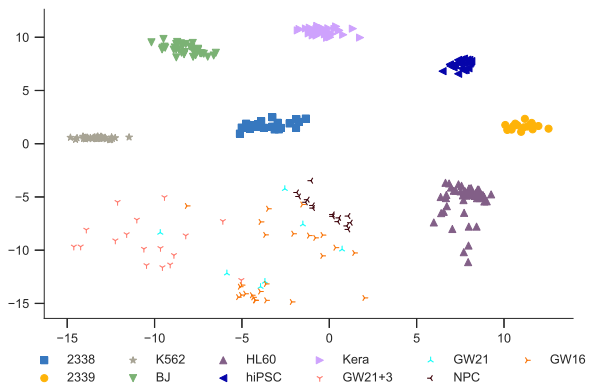
t-SNE



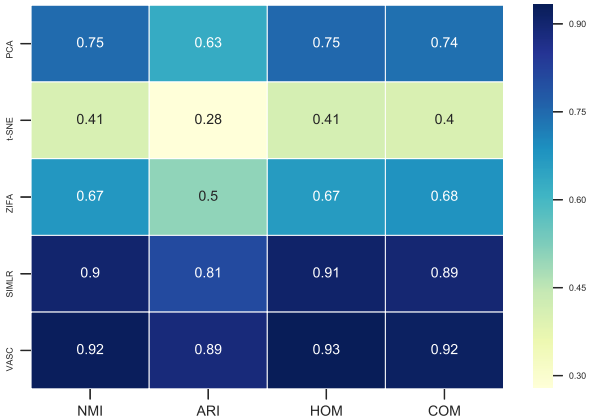
SIMLR



VASC

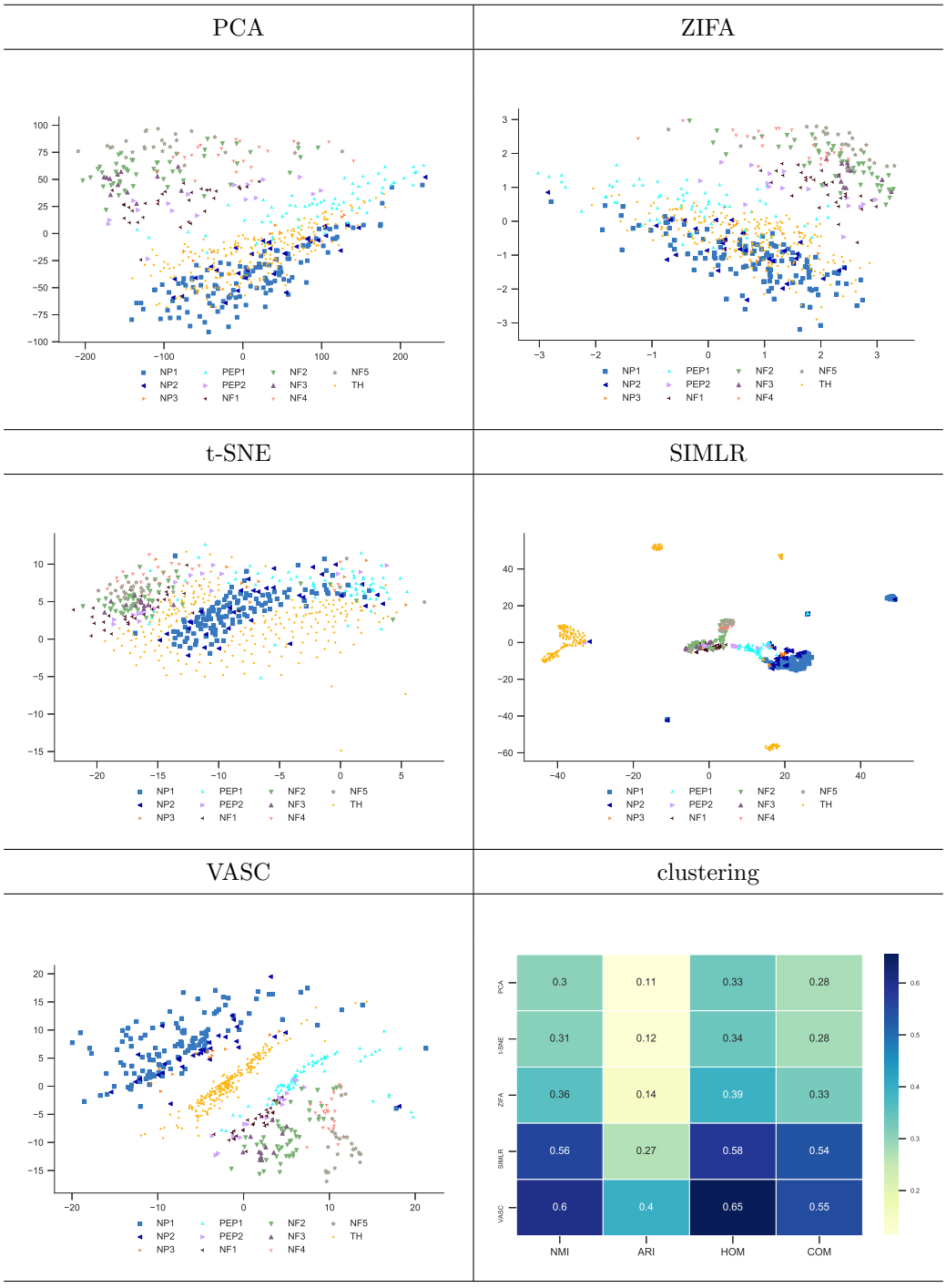


clustering



3.13 Usoskin dataset[13]

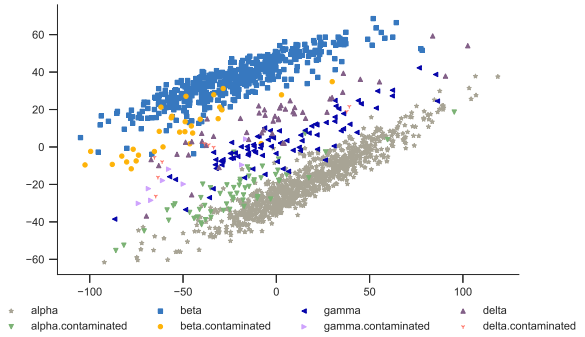
This dataset contains 622 cells of different sensory neuron types. They used computational methods to perform cell classification. There're 11 types of cells: 5 NF clusters (expressing neurofilament heavy chain), 3 Np clusters (non-peptidergic nociceptors), 2 PEP clusters (peptidergic nociceptors) and a TH cluster (tyrosine hydroxylase containing). The original paper found no heterogeneity in the TH cluster. No method could distinct these 11 clusters. VASC and SIMLR could seperate TH from others, while SIMLR made two clusters of TH.



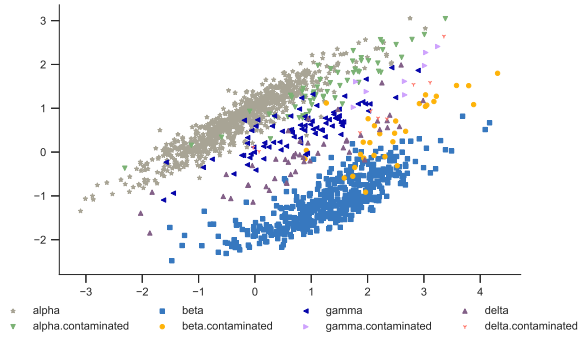
3.14 Xin dataset[14]

This dataset contains 1600 cells from human pancreatic islets, and were sampled from 12 non-diabetic donors and 6 type II diabetes donors.

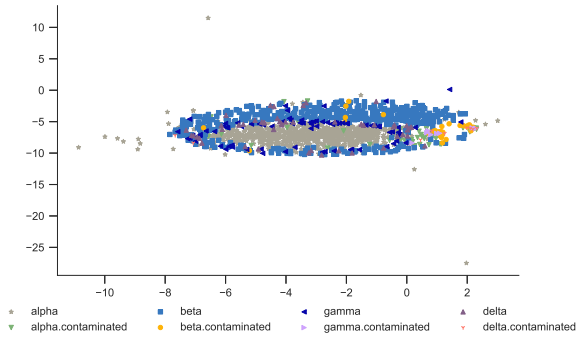
PCA



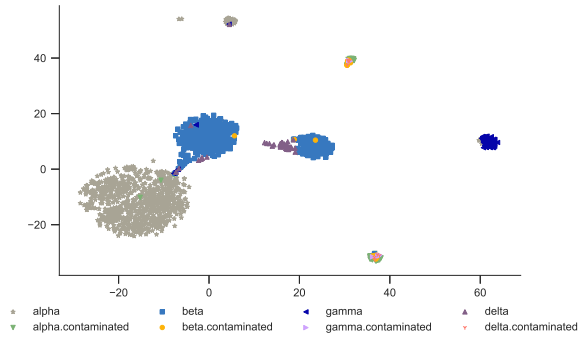
ZIFA



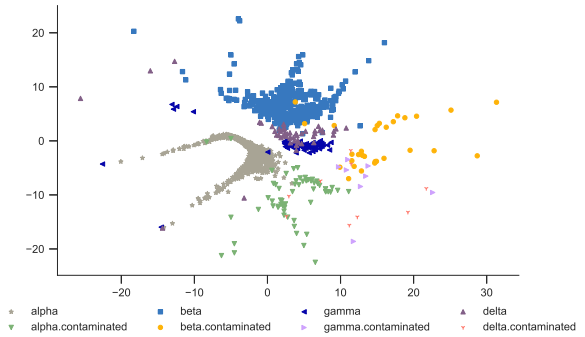
t-SNE



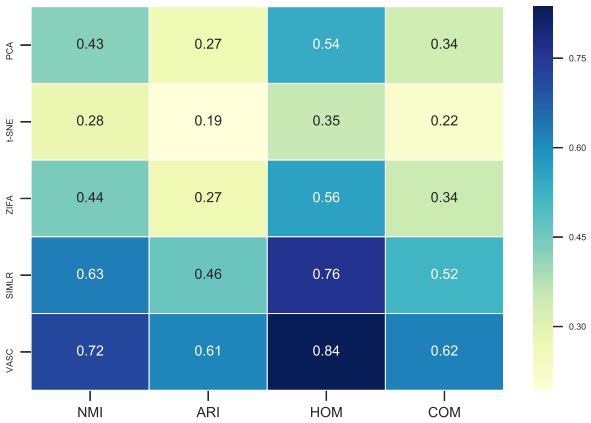
SIMLR



VASC



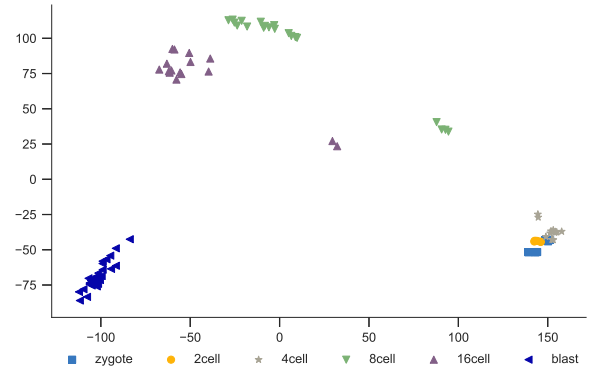
clustering



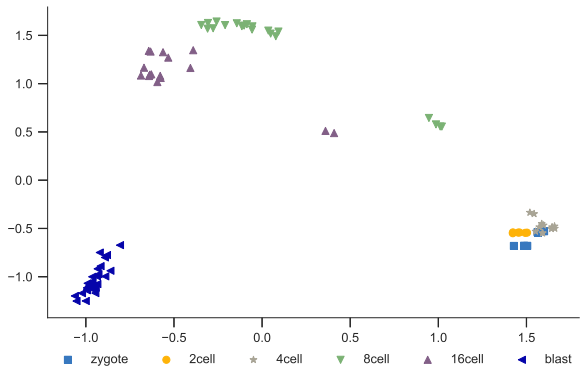
3.15 Yan dataset[15]

This is also a dataset containing mouse embryos cells with stage:zygote, 2cell, 4cell, 8cell, 16cell and blast. There're overall 90 cells.

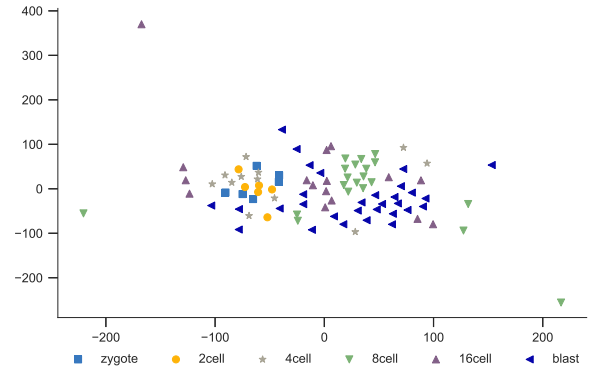
PCA



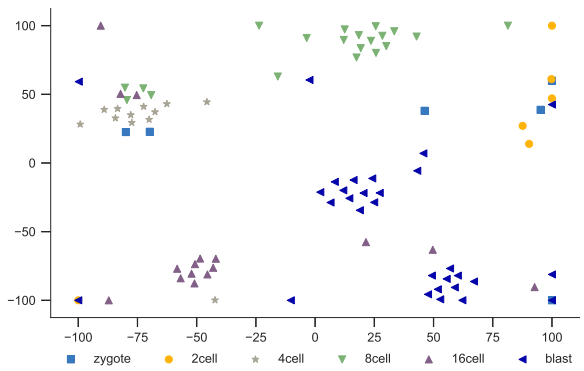
ZIFA



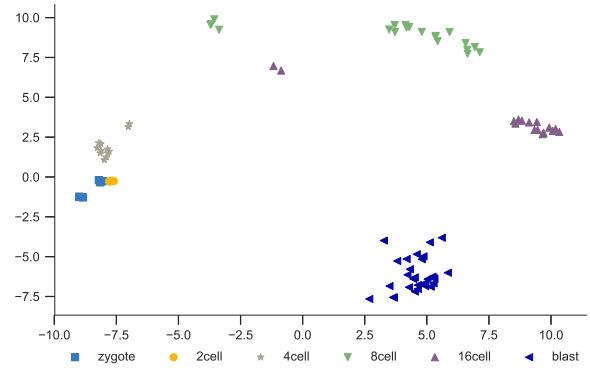
t-SNE



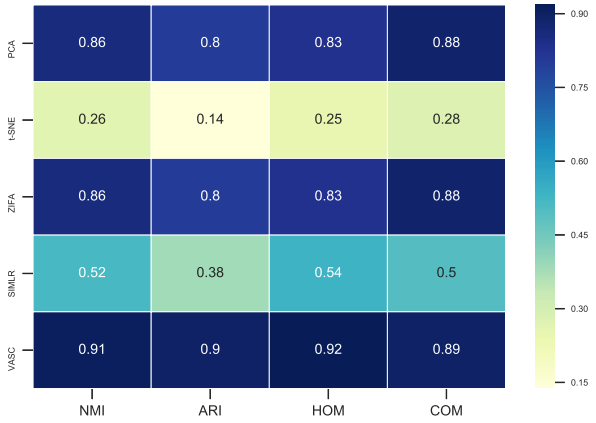
SIMLR



VASC



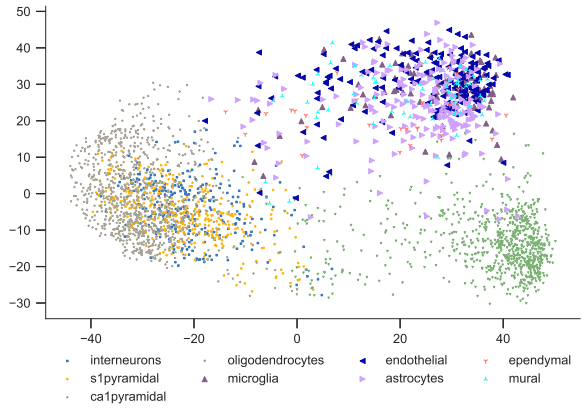
clustering



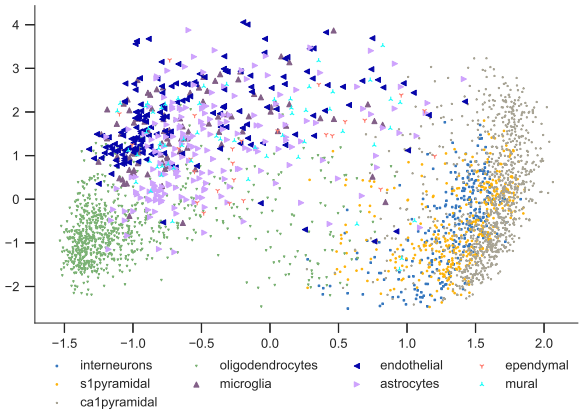
3.16 Zeisel dataset[16]

This dataset contains 3005 cells from 9 cell types of mouse brain cells.

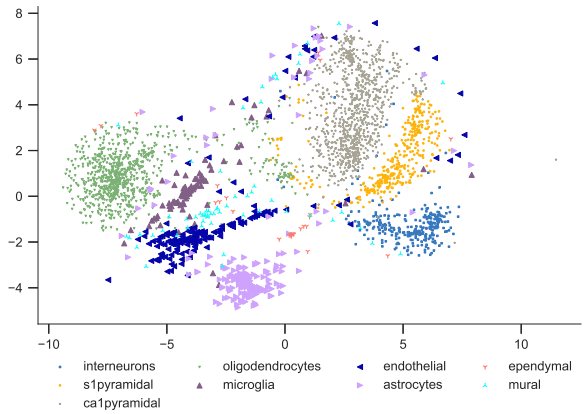
PCA



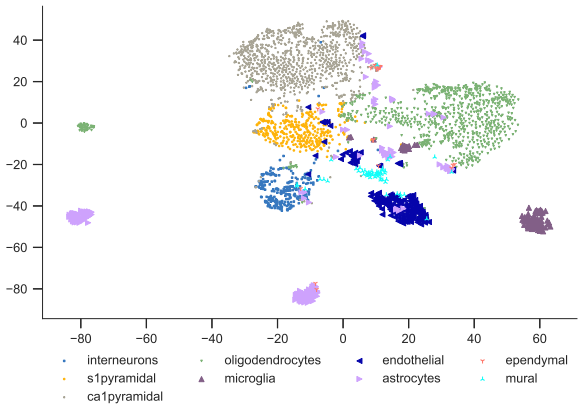
ZIFA



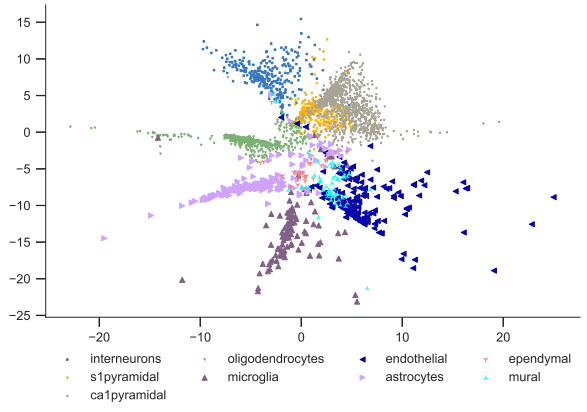
t-SNE



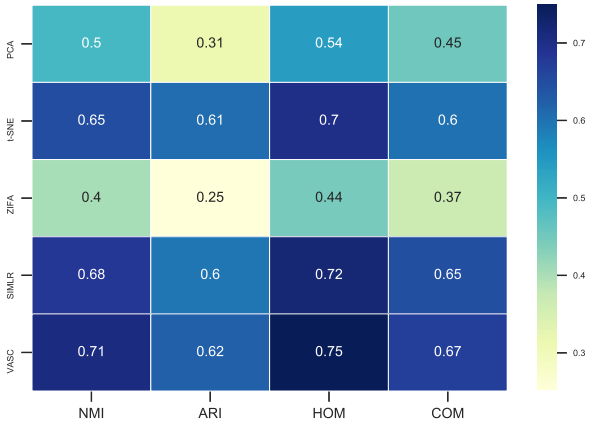
SIMLR



VASC

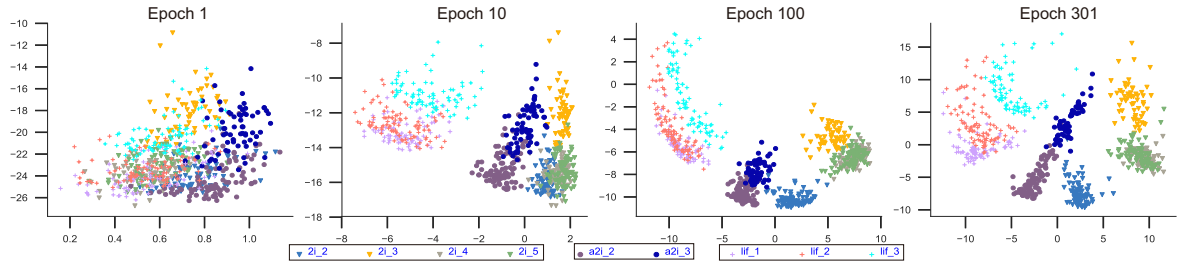


clustering



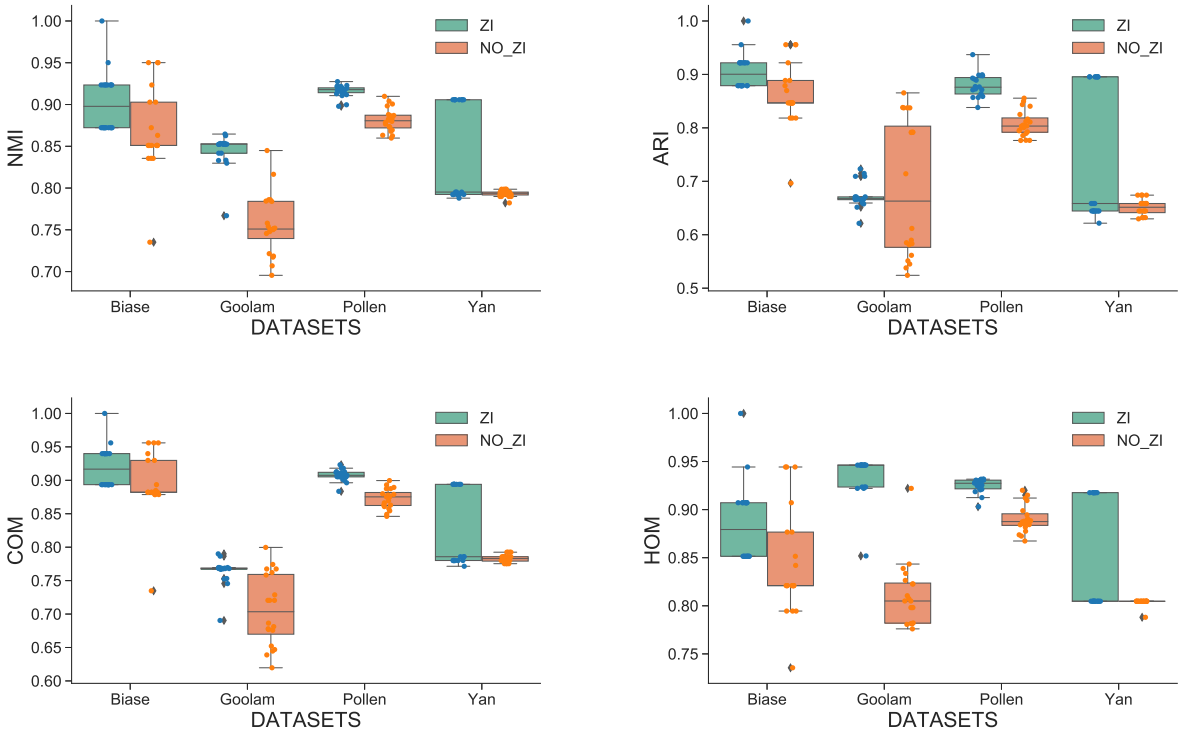
4 Iteration process

We used the Kolodziejczyk dataset to visualize our training procedure. This dataset is composed of three cell types called 2i,a2i and lif, and however, every cell types are obtained by different batches. We showed the dim-2 results of this datasets at epoch 0,10,100 and the convergent results in the following figure. We marked the cell types using different shapes and batches using different colors. As we expected, initially, all kinds of cells tend to clump. Just after 10 epochs, different kinds of cells begin to split, but the batches are still mixed. After 100 epochs, these batches begin to split, too. And finally, all batches almost distribute in different regions, but we still see the same types of cells tend to be more closed. This might because during the initial stages, cell types difference dominated the reconstruction errors, and should be regarded as the primary information of original space. We see the shape of epoch 10 looks like the result of PCA. And then, as the iteration keeps going, the error may mainly come from secondary variance, such as the batches.



5 Reproducibility of VASC and Zero-inflated layers

We ran our methods for 20 times in some datasets with small number of cells and golden cell type labels: Biase,Goolam,Pollen and Yan. We observed a consistent results in terms of NMI,ARI,COM and HOM values. VASC with zero-inflated layers gave better results than withput zeo-inflated layers.



References

[1] Maayan Baron, Adrian Veres, Samuel L Wolock, Aubrey L Faust, Renaud Gaujoux, Amedeo Vetere, Jennifer Hyoje Ryu, Bridget K Wagner, Shai S Shen-Orr, Allon M Klein, et al. A single-cell transcriptomic map of the human and mouse pancreas reveals inter-and intra-cell population structure. *Cell systems*, 3(4):346–360, 2016.

[2] Fernando H Biase, Xiaoyi Cao, and Sheng Zhong. Cell fate inclination within 2-cell and 4-cell mouse embryos revealed by single-cell rna sequencing. *Genome research*, 24(11):1787–1796, 2014.

[3] J Gray Camp, Keisuke Sekine, Tobias Gerber, and Henry Loeffler-Wirth. Multilineage communication regulates human liver bud development. 2017.

[4] Spyros Darmanis, Steven A Sloan, Ye Zhang, Martin Enge, Christine Caneda, Lawrence M Shuer, Melanie G Hayden Gephart, Ben A Barres, and Stephen R Quake. A survey of human brain transcriptome diversity at the single cell level. *Proceedings of the National Academy of Sciences*, 112(23):7285–7290, 2015.

[5] Qiaolin Deng, Daniel Ramsköld, Björn Reinius, and Rickard Sandberg. Single-cell rna-seq reveals dynamic, random monoallelic gene expression in mammalian cells. *Science*, 343(6167):193–196, 2014.

[6] Mubeen Goolam, Antonio Scialdone, Sarah JL Graham, Iain C Macaulay, Agnieszka Jedrusik, Anna Hupalowska, Thierry Voet, John C Marioni, and Magdalena Zernicka-Goetz. Heterogeneity in oct4 and sox2 targets biases cell fate in 4-cell mouse embryos. *Cell*, 165(1):61–74, 2016.

[7] Allon M Klein, Linas Mazutis, Ilke Akartuna, Naren Tallapragada, Adrian Veres, Victor Li, Leonid Peshkin, David A Weitz, and Marc W Kirschner. Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. *Cell*, 161(5):1187–1201, 2015.

[8] Aleksandra A Kolodziejczyk, Jong Kyoung Kim, Jason CH Tsang, Tomislav Ilicic, Johan Henriksson, Kedar N Natarajan, Alex C Tuck, Xuefei Gao, Marc Bühler, Pentao Liu, et al. Single cell rna-sequencing of pluripotent states unlocks modular transcriptional variation. *Cell Stem Cell*, 17(4):471–485, 2015.

[9] Huipeng Li, Elise T Courtois, Debarka Sengupta, Yuliana Tan, Kok Hao Chen, Jolene Jie Lin Goh, Say Li Kong, Clarinda Chua, Lim Kiat Hon, Wah Siew Tan, et al. Reference component analysis of single-cell transcriptomes elucidates cellular heterogeneity in human colorectal tumors. *Nature Genetics*, 49(5):708–718, 2017.

[10] Anoop P Patel, Itay Tirosh, John J Trombetta, Alex K Shalek, Shawn M Gillespie, Hiroaki Wakimoto, Daniel P Cahill, Brian V Nahed, William T Curry, Robert L Martuza, et al. Single-cell rna-seq highlights intratumoral heterogeneity in primary glioblastoma. *Science*, 344(6190):1396–1401, 2014.

[11] Sophie Petropoulos, Daniel Edsgärd, Björn Reinius, Qiaolin Deng, Sarita Pauliina Panula, Simone Codeluppi, Alvaro Plaza Reyes, Sten Linnarsson, Rickard Sandberg, and Fredrik Lanner. Single-cell rna-seq reveals lineage and x chromosome dynamics in human preimplantation embryos. *Cell*, 165(4):1012–1026, 2016.

[12] Alex A Pollen, Tomasz J Nowakowski, Joe Shuga, Xiaohui Wang, Anne A Leyrat, Jan H Lui, Nianzhen Li, Lukasz Szpankowski, Brian Fowler, Peilin Chen, et al. Low-coverage single-cell mrna sequencing reveals cellular heterogeneity and activated signaling pathways in developing cerebral cortex. *Nature biotechnology*, 32(10):1053–1058, 2014.

[13] Dmitry Usoskin, Alessandro Furlan, Saiful Islam, Hind Abdo, Peter Lönnerberg, Daohua Lou, Jens Hjerling-Leffler, Jesper Haeggström, Olga Kharchenko, Peter V Kharchenko, et al. Unbiased classification of sensory neuron types by large-scale single-cell rna sequencing. *Nature neuroscience*, 18(1):145–153, 2015.

[14] Yurong Xin, Jinrang Kim, Haruka Okamoto, Min Ni, Yi Wei, Christina Adler, Andrew J Murphy, George D Yancopoulos, Calvin Lin, and Jesper Gromada. Rna sequencing of single human islet cells reveals type 2 diabetes genes. *Cell metabolism*, 24(4):608–615, 2016.

[15] Liying Yan, Mingyu Yang, Hongshan Guo, Lu Yang, Jun Wu, Rong Li, Ping Liu, Ying Lian, Xiaoying Zheng, Jie Yan, et al. Single-cell rna-seq profiling of human preimplantation embryos and embryonic stem cells. *Nature structural & molecular biology*, 20(9):1131–1139, 2013.

[16] Amit Zeisel, Ana B Muñoz-Manchado, Simone Codeluppi, Peter Lönnerberg, Gioele La Manno, Anna Juréus, Sueli Marques, Hermany Munguba, Liqun He, Christer Betsholtz, et al. Cell types in the mouse cortex and hippocampus revealed by single-cell rna-seq. *Science*, 347(6226):1138–1142, 2015.