Deep learning enables accurate clustering with batch effect removal in single-cell RNA-seq analysis

Xiangjie Li, Mingyao Li et.al

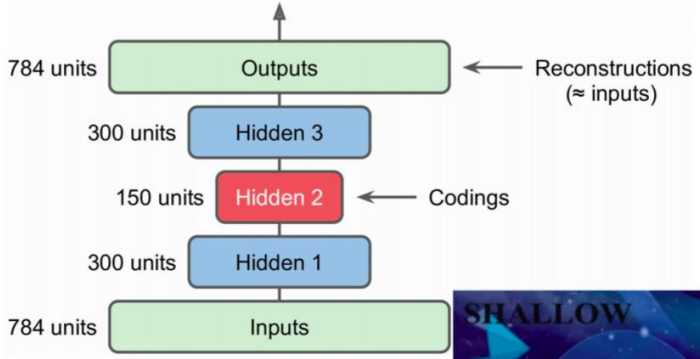
Unsupervised Deep Embedding for Clustering Analysis

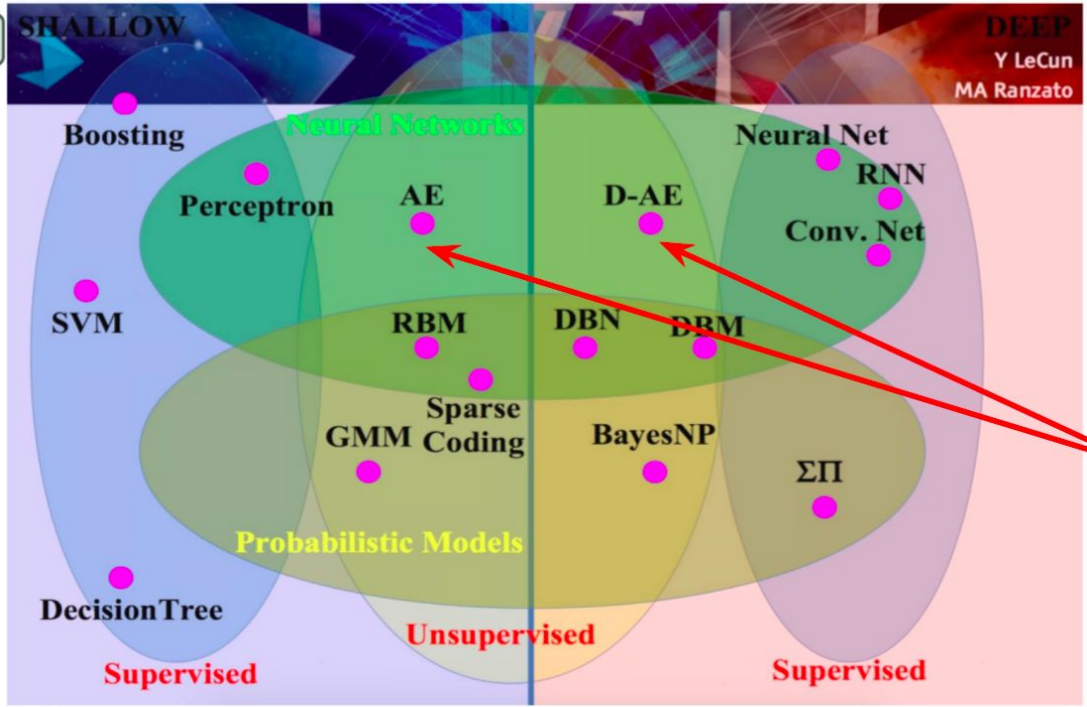
PCA vs Autoencoder

— autoencoders are much more ***flexible*** *than PCA.*

*— NN activation functions introduce "****non-linearities****" in encoding, but PCA* ***only*** *does linear transformation.*

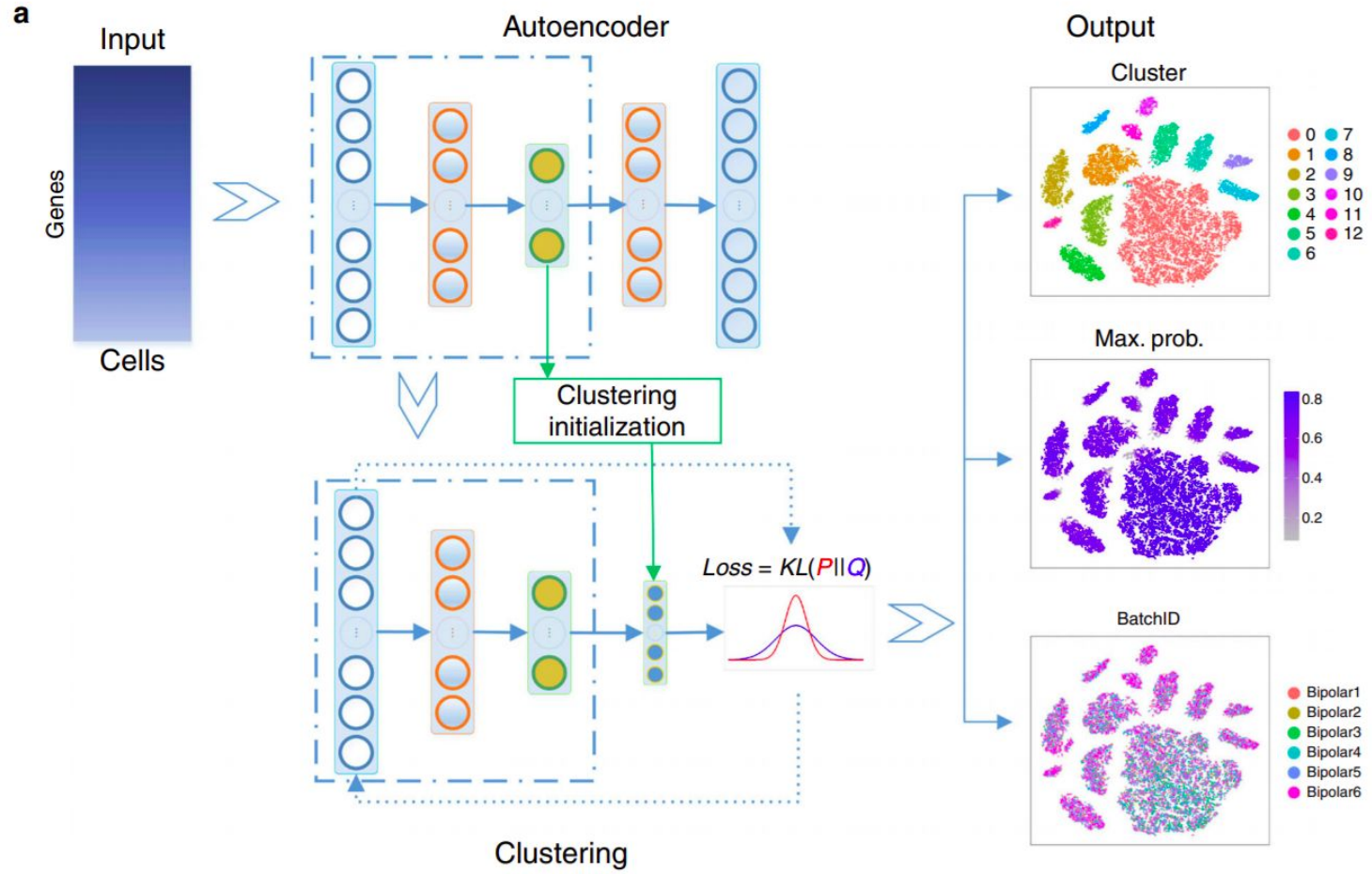
— we can stack autoencoders to form a ***deep autoencoder network***



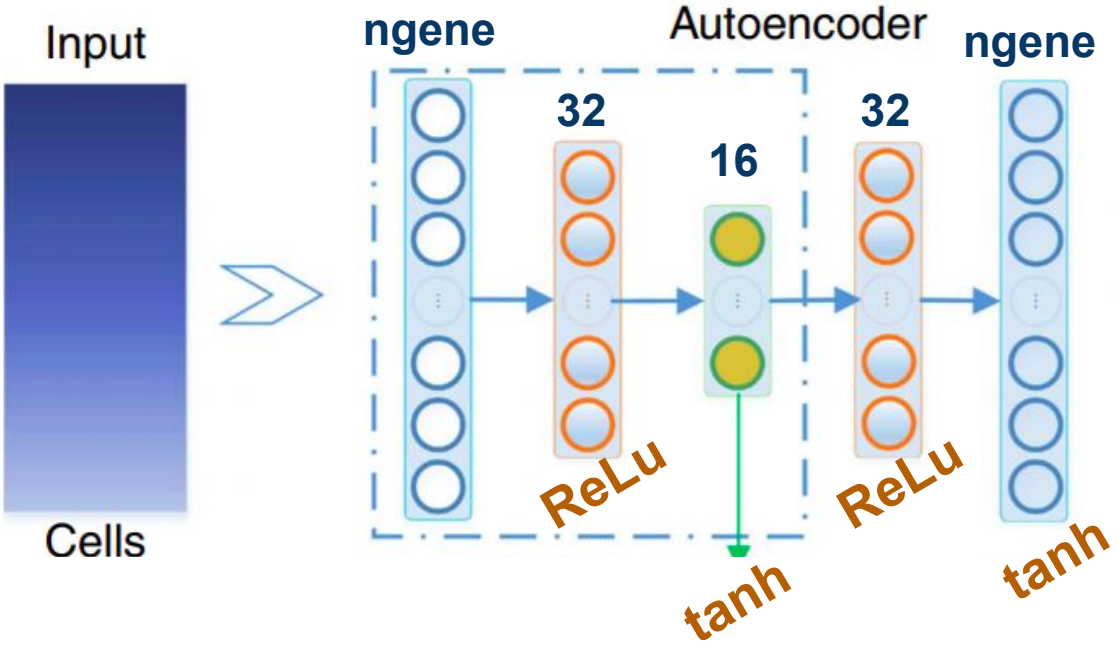


1 Method

1- Overall Schematic

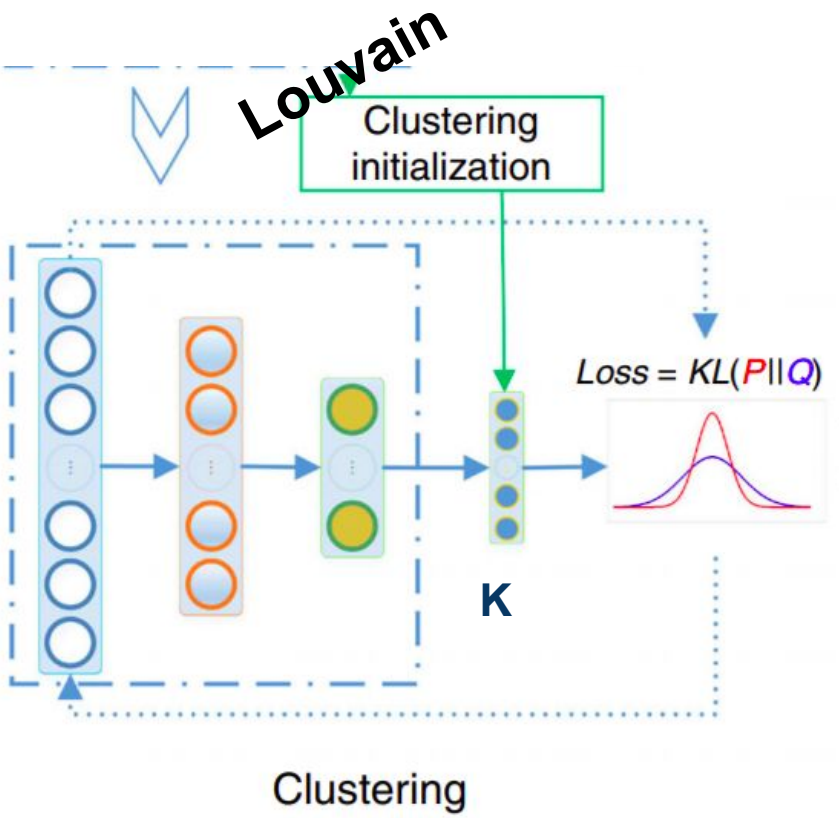


1-1 Autoencoder




```python
def getdims(x=(10000,200)):
    """
    This function will give the s
    return the dims for network
    """
    assert len(x)==2
    n_sample=x[0]
    if n_sample>20000:# may be ne
        dims=[x[-1],128,32]
    elif n_sample>10000:#10000
        dims=[x[-1],64,32]
    elif n_sample>5000: #5000
        dims=[x[-1],32,16] #16
    elif n_sample>2000:
        dims=[x[-1],128]
    elif n_sample>500:
        dims=[x[-1],64]
    else:
        dims=[x[-1],16]
    return dims
```

1-2 Clustering with KL divergence



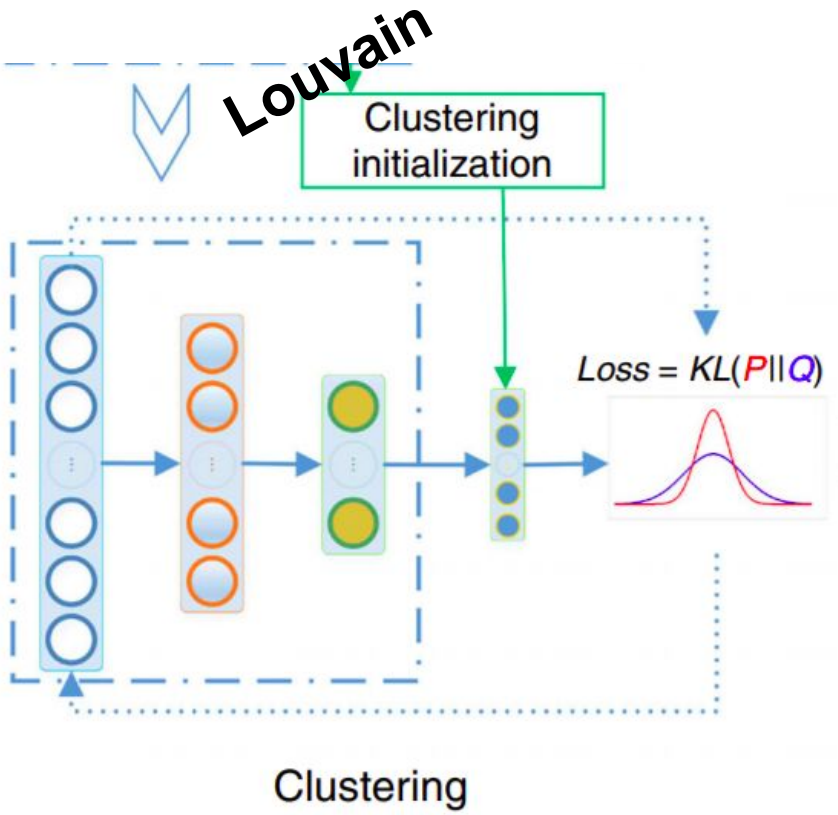
$$q_{ij} = \frac{\left(1+ \|z_i - \mu_j\|^2 / \alpha\right)^{-1}}{\sum_{j'} \left(1+ \|z_i - \mu_{j'}\|^2 / \alpha\right)^{-1}},$$

$$L = KL\left(\mathbf{P} \parallel \mathbf{Q}\right) = \sum_{i=1}^{n} \sum_{j=1}^{K} p_{ij} \log \frac{p_{ij}}{q_{ij}},$$

$$p_{ij} = \frac{q_{ij}^2 / \sum_{i=1}^{n} q_{ij}}{\sum_{j=1}^{K} \left(q_{ij}^2 / \sum_{i=1}^{n} q_{ij} \right)}.$$

K is the number of clusters estimated by louvain

1-3 Back Propagation



$$\frac{\partial L}{\partial z_i} = \frac{\alpha+1}{\alpha}\sum_{j=1}^{K}\left(1+\frac{z_i-\mu_j^2}{\alpha}\right)^{-1} \times (p_{ij}-q_{ij})(z_i-\mu_j),$$

$$\frac{\partial L}{\partial \mu_j} = \frac{-(\alpha+1)}{\alpha}\sum_{i=1}^{n}\left(1+\frac{z_i-\mu_j^2}{\alpha}\right)^{-1} \times (p_{ij}-q_{ij})(z_i-\mu_j).$$

2 Metrics

2-1 Evaluation metric for clustering

$X \backslash Y$	Y_1	Y_2	$\cdots$	Y_s	sums
X_1	n_{11}	n_{12}	$\cdots$	n_{1s}	a_1
X_2	n_{21}	n_{22}	$\cdots$	n_{2s}	a_2
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
X_r	n_{r1}	n_{r2}	$\cdots$	n_{rs}	a_r
sums	b_1	b_2	$\cdots$	b_s	

Adjusted Rand index

$$ARI = \frac{\sum_{ij}\binom{n_{ij}}{2} - \left[\sum_i \binom{a_i}{2}\sum_j \binom{b_j}{2}\right] / \binom{n}{2}}{\frac{1}{2}\left[\sum_i \binom{a_i}{2} + \sum_j \binom{b_j}{2}\right] - \left[\sum_i \binom{a_i}{2}\sum_j \binom{b_j}{2}\right] / \binom{n}{2}}$$

$$AdjustedIndex = \frac{Index - ExpectedIndex}{MaxIndex - ExpectedIndex}$$

2-2 Evaluation metric for batch effect removal

$$KL = \sum_{b=1}^{B} p_b \log \frac{p_b}{q_b},$$

q_b is the proportion of cells from batch b among all cells.

p_b is the the proportion of cells from batch b in a given region based on results from a clustering algorithm.

Region is determined by k-nearest neighbor

Smaller final KL divergence indicates **better** batch mixing.

3 Results

3-1 Complex batch effect

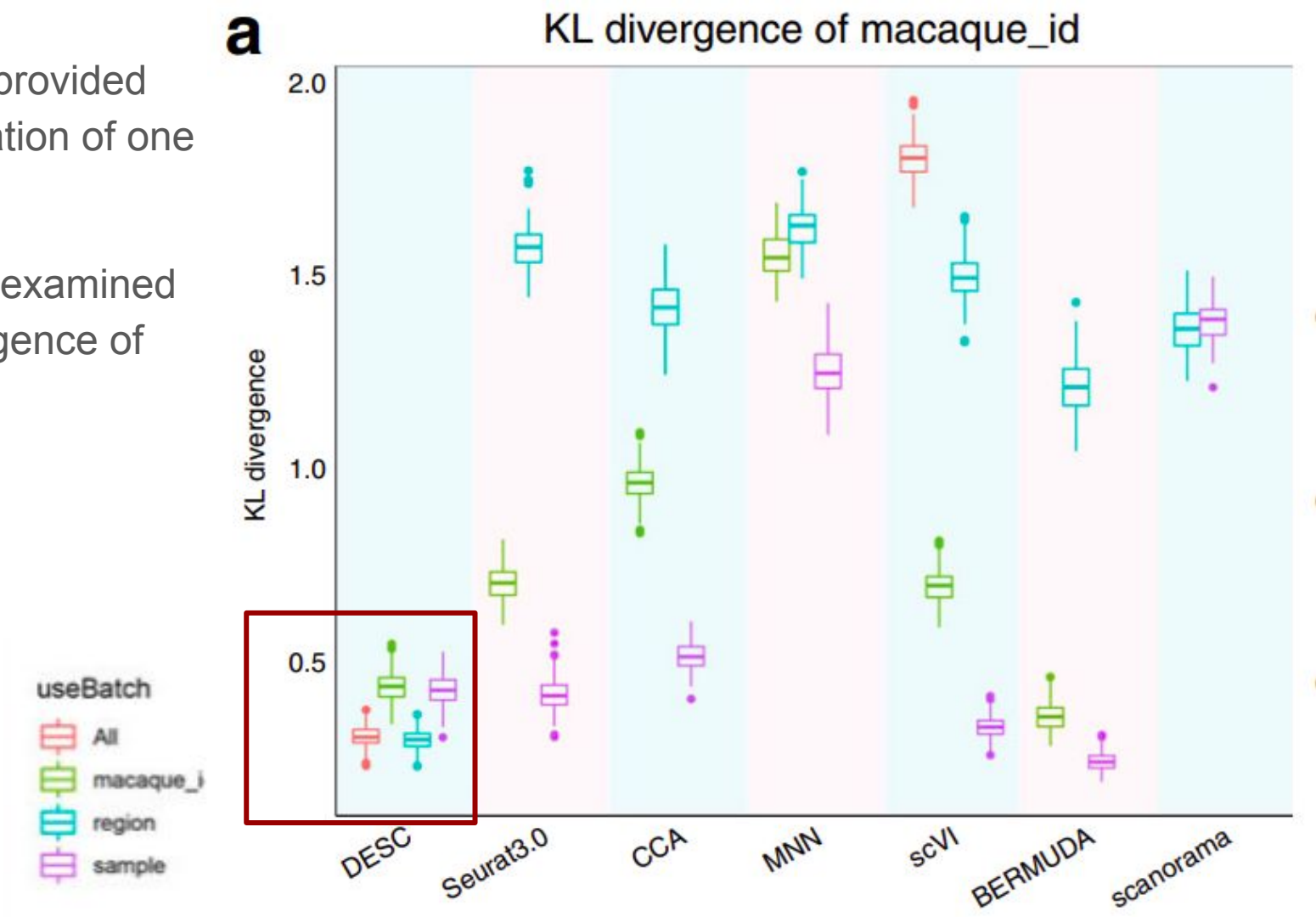
Analyzed a scRNA-seq dataset that includes 21,017 foveal and 9285 peripheral bipolar cells from retina in four macaques.

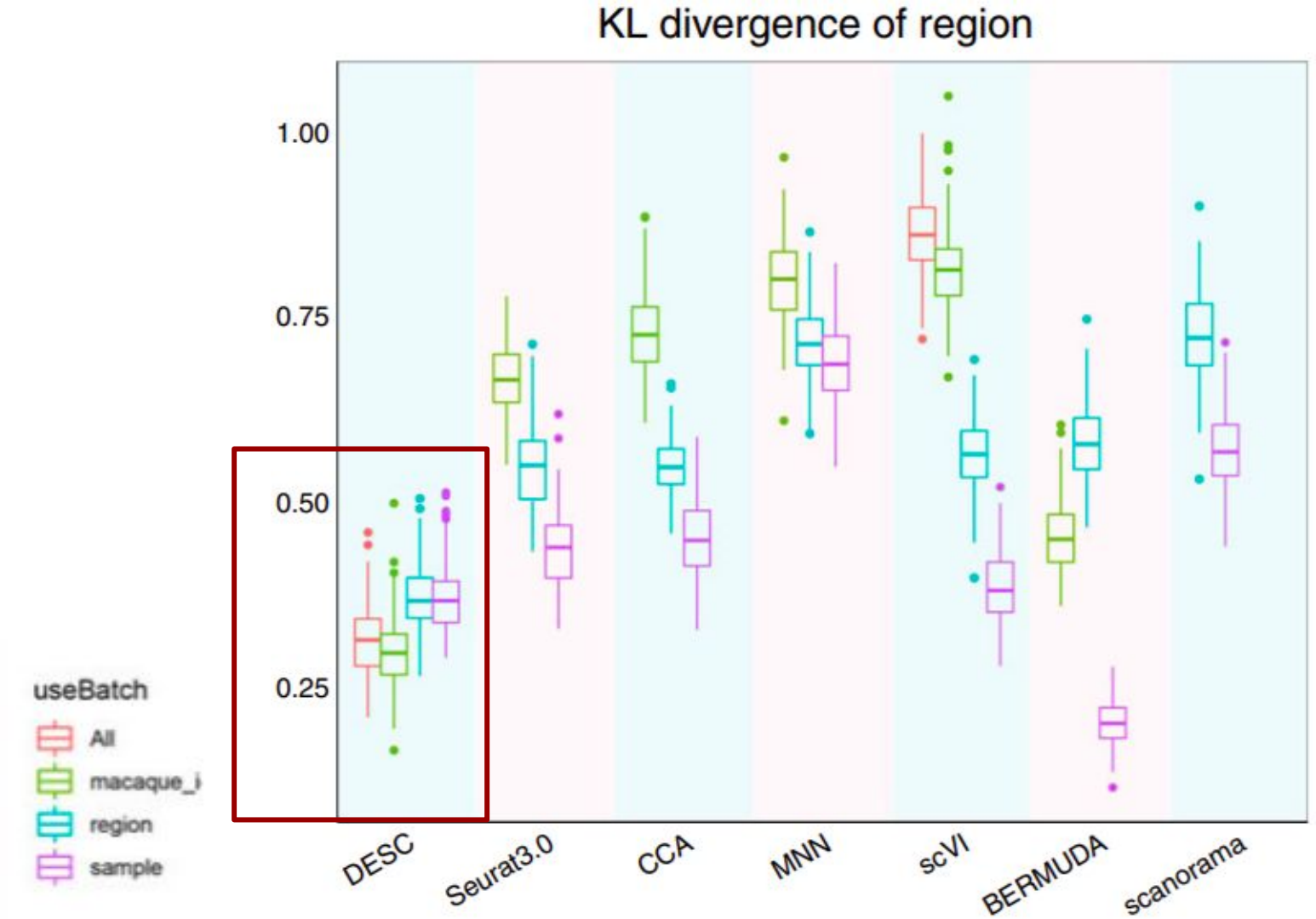
This dataset is relatively complex because it contains three different levels of batch:

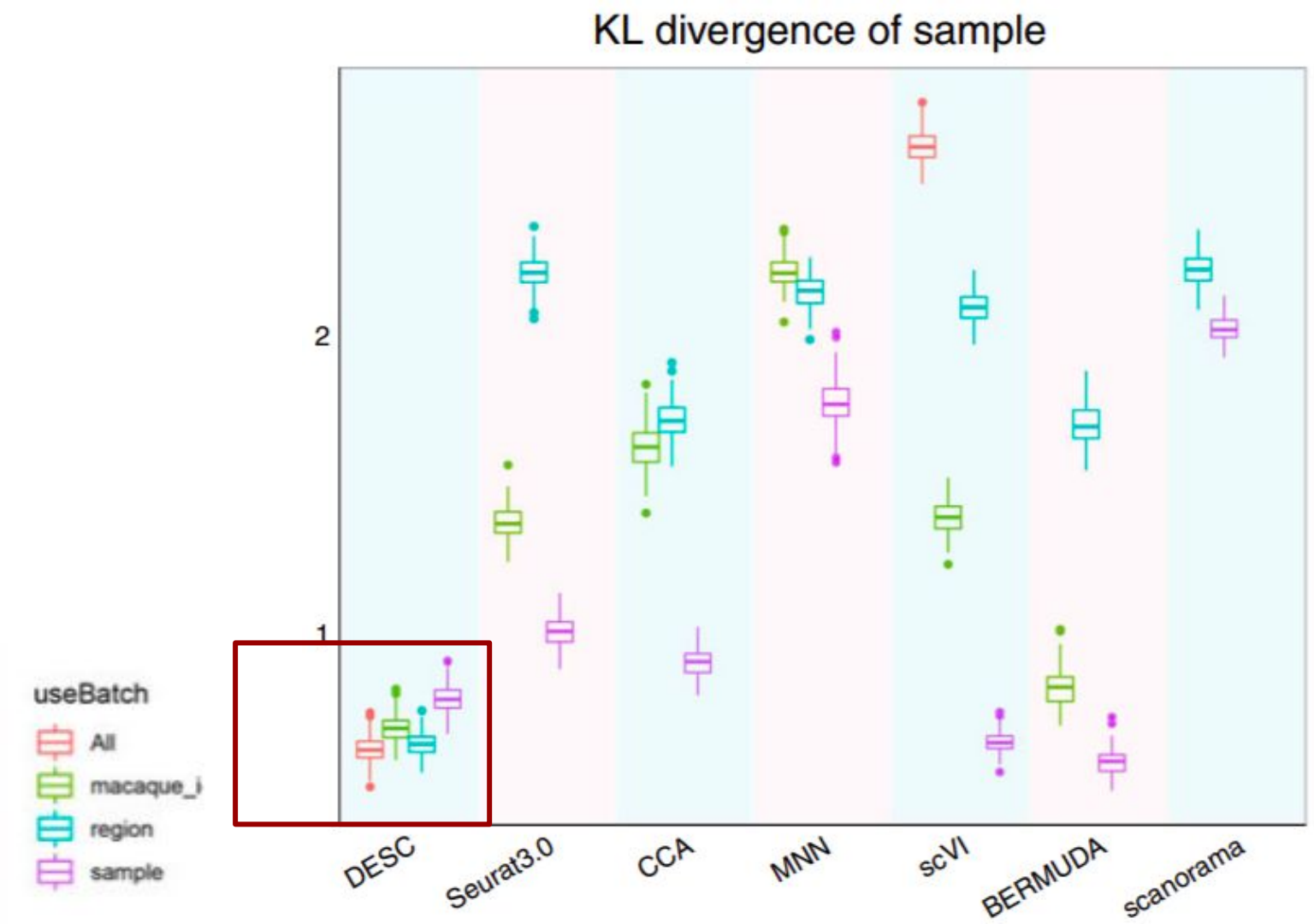
macaque ids, **sample ids**, and **region ids**.

Algorithms are provided with the information of one batch.

The results are examined by the KL divergence of three batches,




KL divergence of region
1.00
0.75
0.50
0.25
useBatch
All
macaque_i
region
sample
DESC
Seurat3.0
CCA
MNN
scVI
BERMUDA
scanorama


KL divergence of sample
2
1
useBatch
All
macaque_i
region
sample
DESC
Seurat3.0
CCA
MNN
scVI
BERMUDA
scanorama

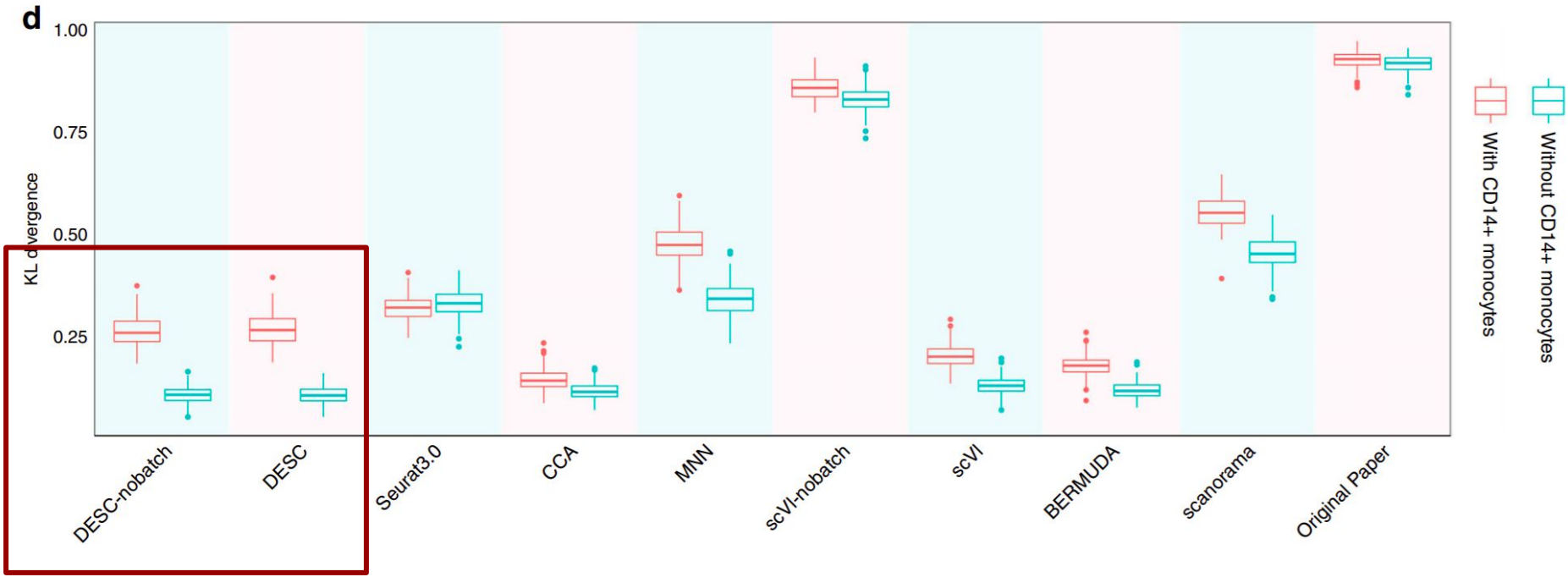
3-2 Batch Confounds Biological Variation

Dataset that includes 24,679 human PBMCs from eight patients with lupus.

The cells were split into a control group and a matched group stimulated with interferon-beta (INF-β), which leads to a drastic but highly cell-type-specific response.

This dataset is extremely challenging because removal of technical batch effect is complicated by the presence of biological differences, both between cell types under the same condition and between different conditions for the same cell type.

KL divergence



Relative Reading

- Comprehensive Intro on Autoencoder [Link]
- Dimension Estimation Using Autoencoders [Link]
- Unsupervised Deep Embedding for Clustering Analysis [Link]