Sparse Autoencoder for Unsupervised Clustering, Imputation, and Embedding (SAUCIE)

Exploring **Single-cell Data** with Deep Multitasking Neural Networks

Data Science Journal Club,

July 1, 2020

Yunhua Zhu

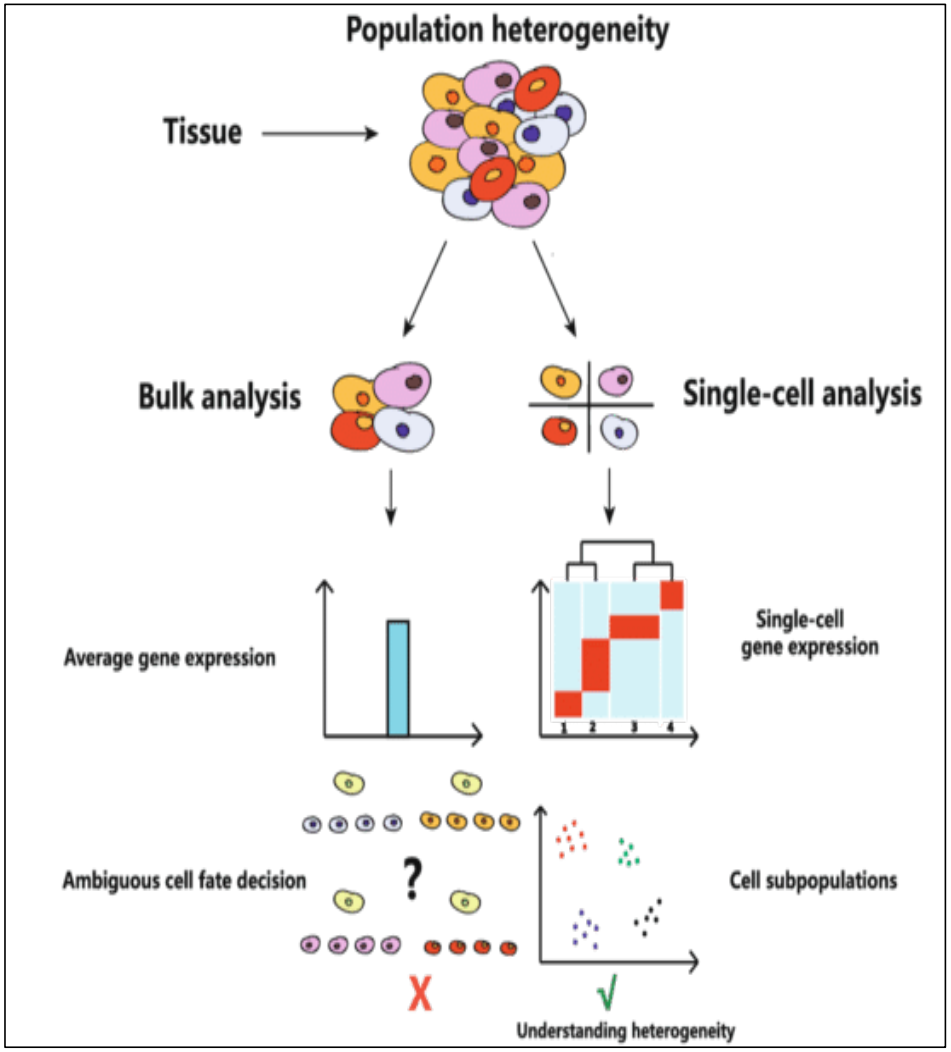
National Institute of Allergy and Infectious Diseases

Outline of today's talk:

- Introduce problems in single cell analysis
  - single cell RNA-seq and
  - single cell mass Cytometry measured by Time of Flight (CyTOF)
- Introduce neural network and autoencoder
- Introduce the how the SAUCIE package were designed
- Discuss the results
  - Batch correction
  - Imputation
  - Clustering
  - Dimension reduction
  - Scale up with problematic real patient data
- Summary

Why study single cells?

- Advantages
  - High resolution for novel details
    - Reveal minor populations
    - Reveal gradual transitions
    - Niche interactions
  - High throughput
    - Completeness for an atlas study of a target tissue or an entire organism -- ecosystem
    - High statistic power to infer relationships between genes

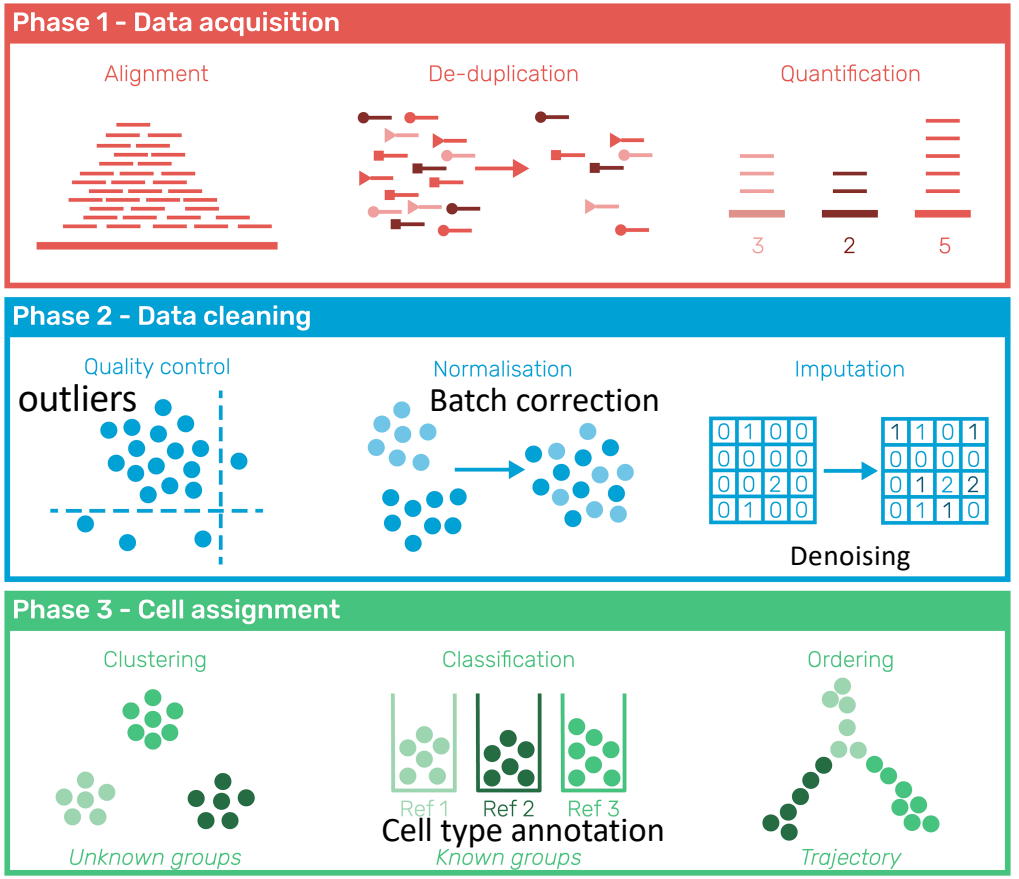


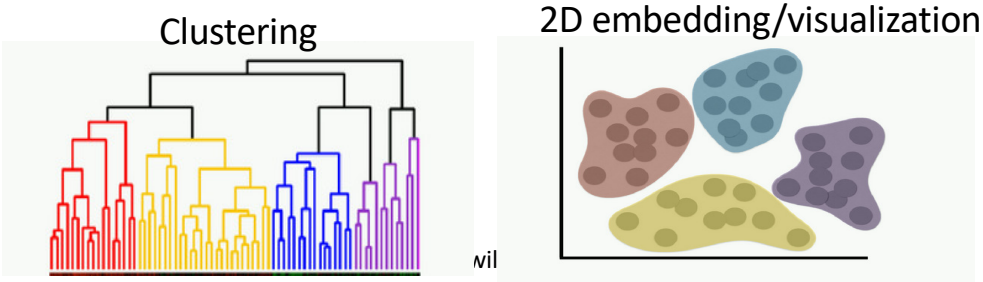
National Institute of Allergy and Infectious Diseases

https://www.researchgate.net/figure/Single-cell-analysis-reveals-heterogeneity-Traditional-experiments-on-bulk-samples-mask_fig1_312664044

Key challenges

- Noisy, dropouts (>90%)
  - Statistically estimate the missing expression values -- imputation
- Batch effects due to processing
  - Batch correction
- High dimension (20-30,000 genes)
  - Dimension reduction/visualization
  - PCA, tSNE, UMAP, etc
- Group cells into same type/state
- Scale, few cells → millions
  - Huge amount of data representing computational challenge

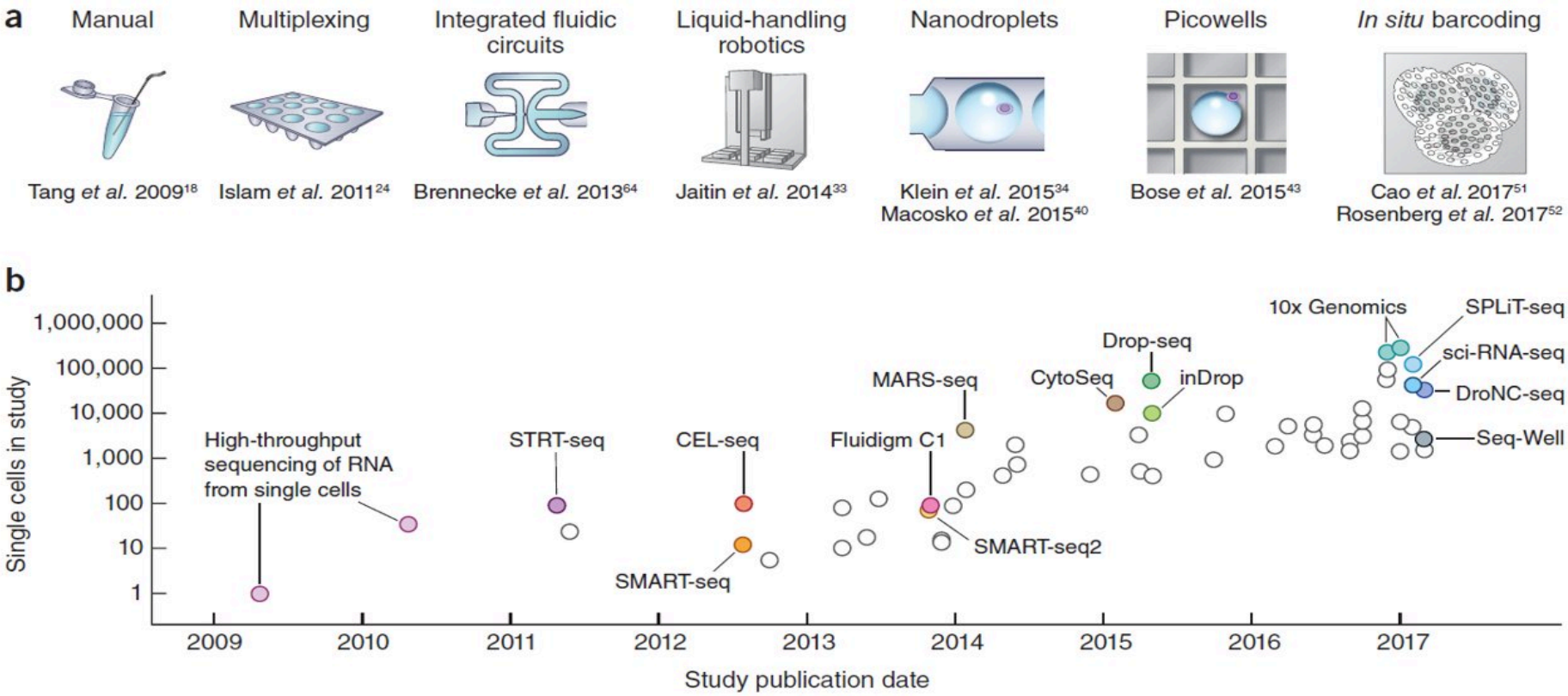




https://febs.onlinelibrary.wiley.com/doi/full/10.1002/1873-3468.12684
https://journals.plos.org/ploscompbiol/article/figure?id=10.1371/journal.pcbi.1006245.g002
https://www.pnas.org/content/116/52/27151
https://public.ornl.gov/site/gallery/detail.cfm?id=323

National Institute of Allergy and Infectious Diseases

The increasing throughput makes analysis more challenging…



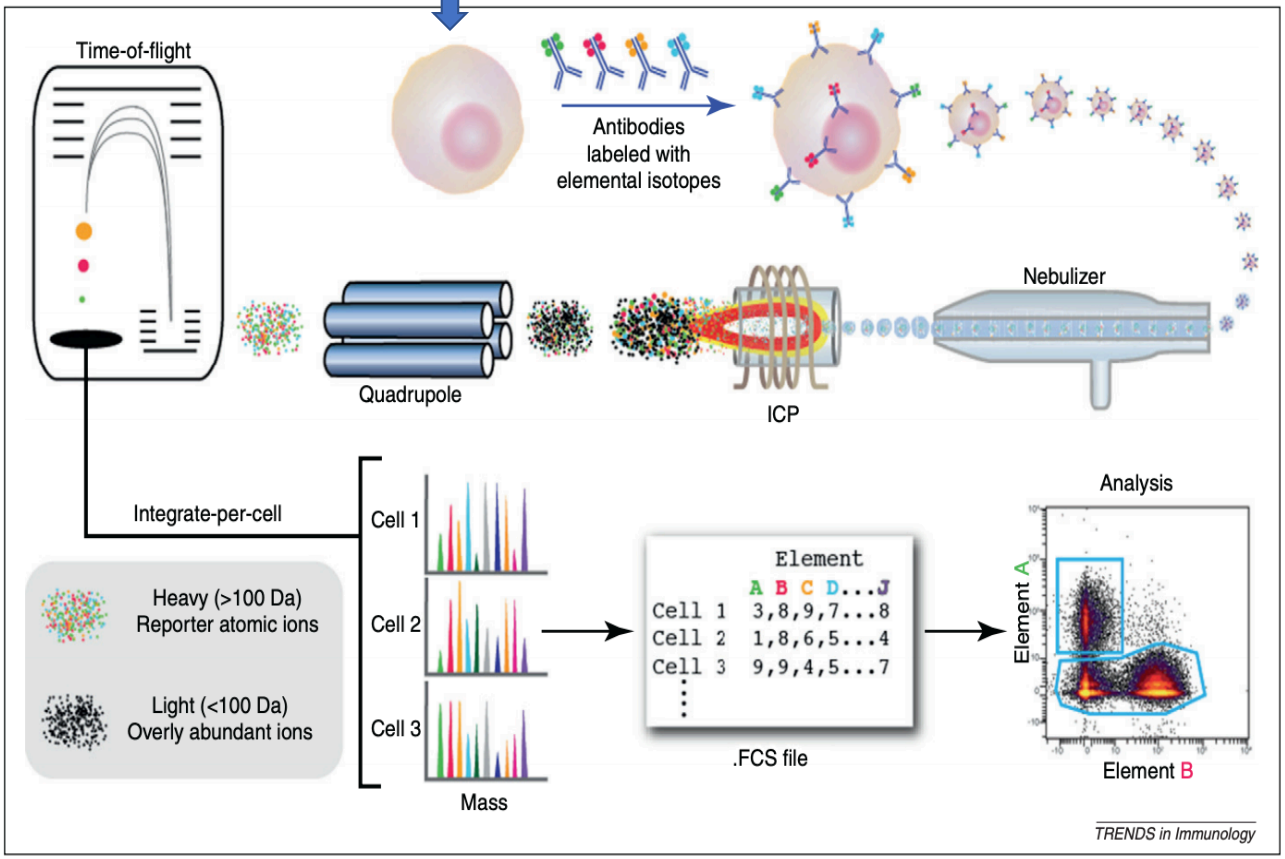
National Institute of Allergy and Infectious Diseases

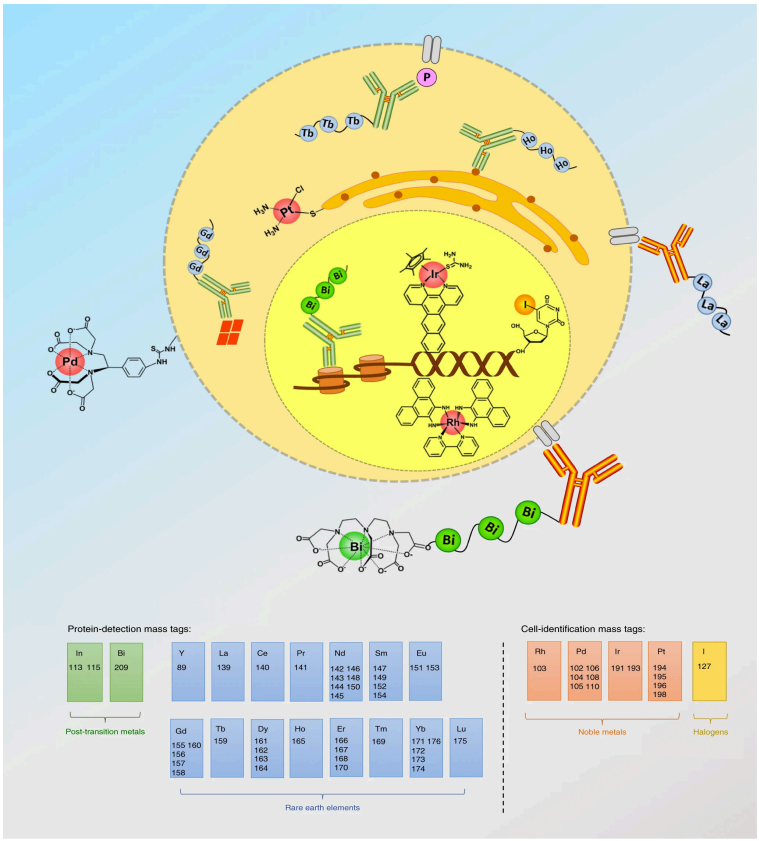
”The Human Cell Atlas[101], which aims to map **35 trillion cells** from the human body, has already started a few pilot studies.”

Hwang et al., 2018

single-cell CyTOF (mass Cytometry -Time Of Flight) is similar, demand higher computational power

180 samples and **11 million cells**, processed on different dates and sequenced on different machines. Acutely infected dengue patients, recovered and healthy controls.

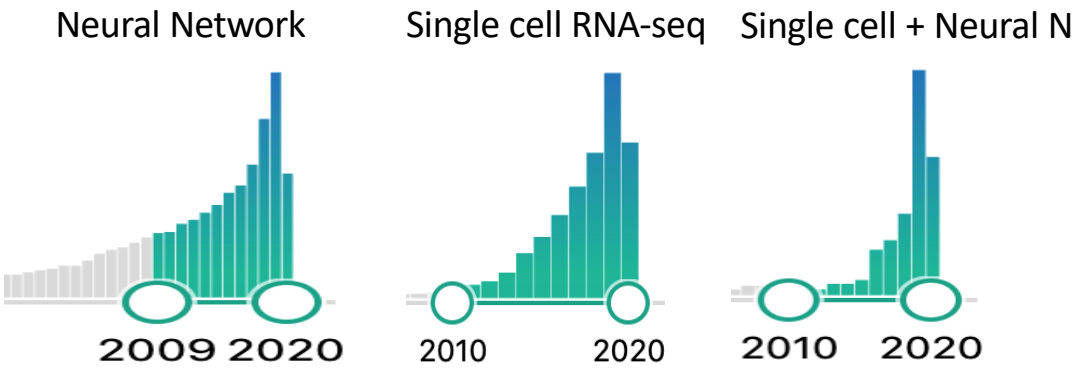




Deep learning getting more momentum in single cell analysis

- Neural network papers is increasing
- Single cell analysis is a recent development
- Single cell analysis with neural network is picking up fast

No. of publications in Pubmed search



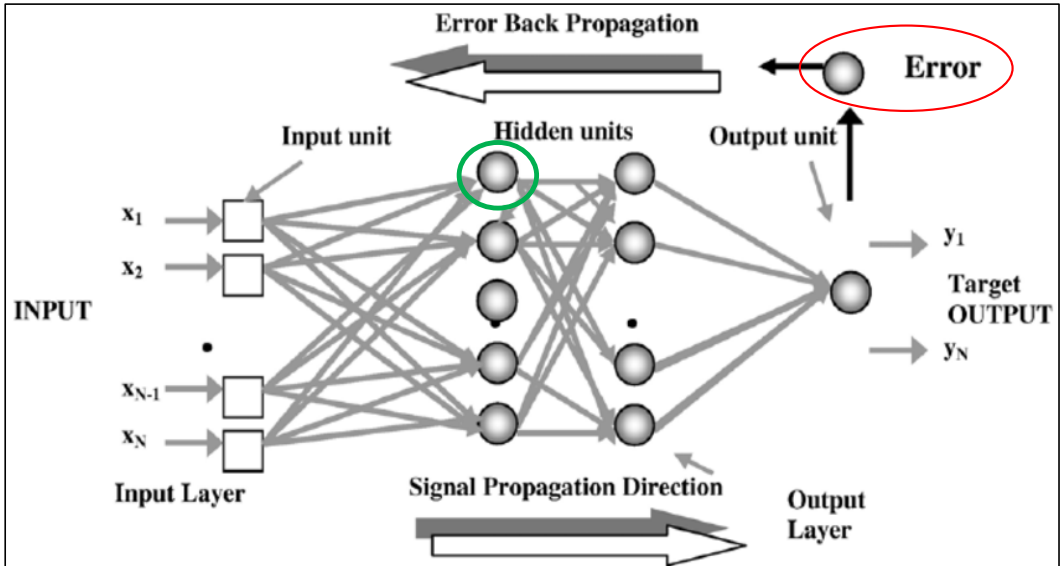
National Institute of Allergy and Infectious Diseases

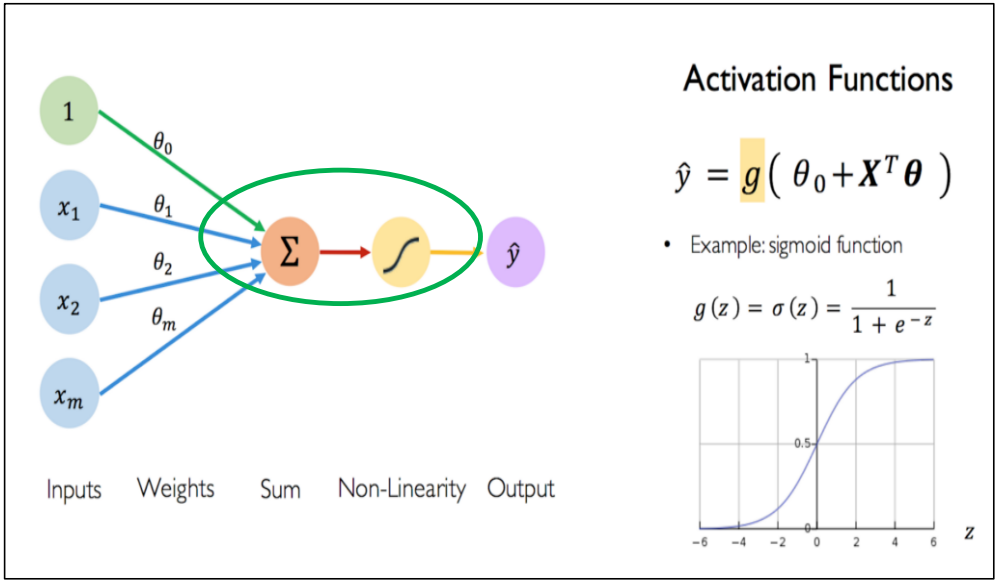
Deep learning and Artificial Neural Networks (ANN)

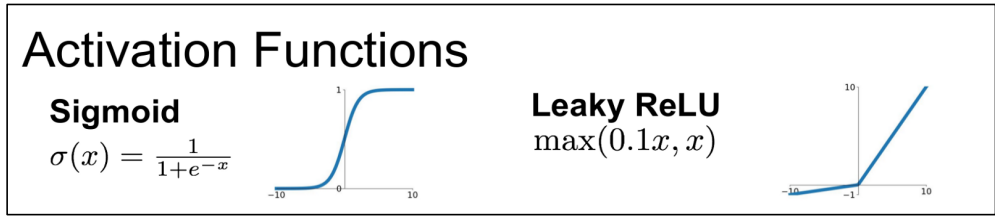
- Handles big data
- Handles variation well
- Combine Linear + Non-linear transformation

- Differentiable loss function for back propagation
- Parallel processing makes efficient computation with HPC

$$\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} - \eta \frac{\partial}{\partial \boldsymbol{\theta}} Loss$$







Training the network with backpropagation in supervised learning

$$\mathbf{w} \leftarrow \mathbf{w} - \eta \frac{\partial}{\partial \mathbf{w}} Loss$$

$$Loss(y, \hat{y}) = \sum_{i=1}^{n} (y - \hat{y})^2$$

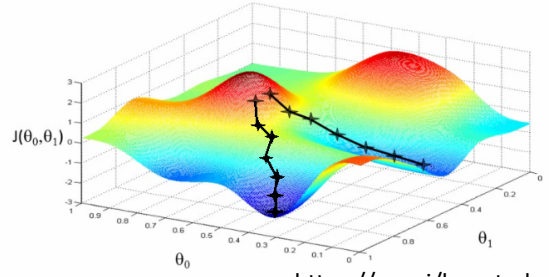
$$\frac{\partial\, Loss(y,\hat{y})}{\partial W} = \frac{\partial Loss(y,\hat{y})}{\partial \hat{y}} * \frac{\partial \hat{y}}{\partial z} * \frac{\partial z}{\partial W} \quad \text{where } z = Wx + b$$

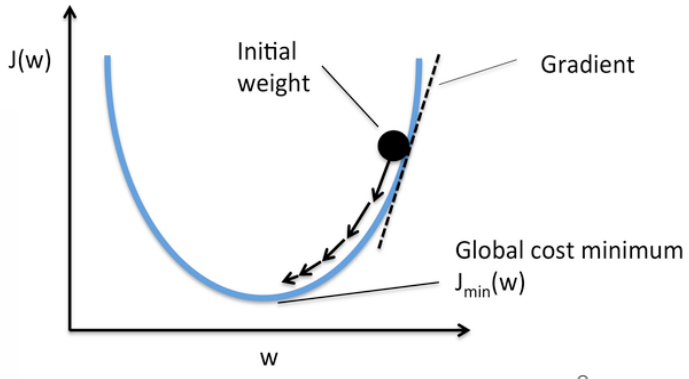
$$= 2(y - \hat{y}) * \text{derivative of sigmoid function} * x$$

$$= 2(y - \hat{y}) * z(1\text{-}z) * x$$

Deep learning is efficient because

- Monotonic activation function ensure that the global minimum of the cost function equals to its local minimum.
- Partial derivatives of each weight can be readily obtained and can be computed in parallel.
- Parameters (weights) optimized quickly along the tangent.





https://mc.ai/how-to-build-your-own-neural-network-from-scratch-in-python/

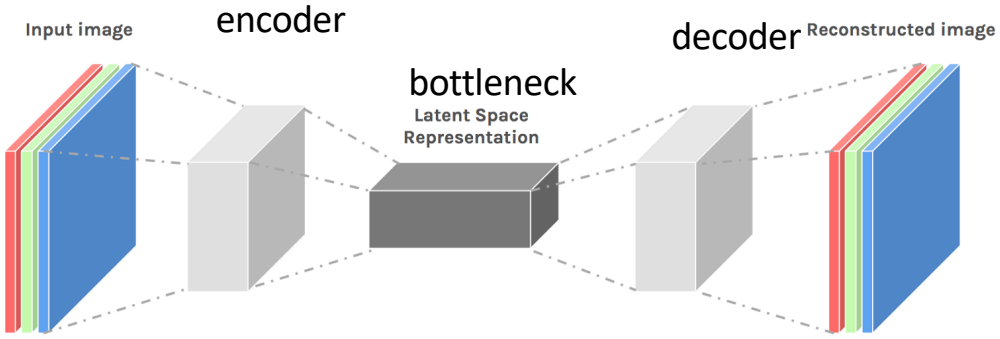
National Institute of Allergy and Infectious Diseases

How can we apply ANN to RNA-seq data?

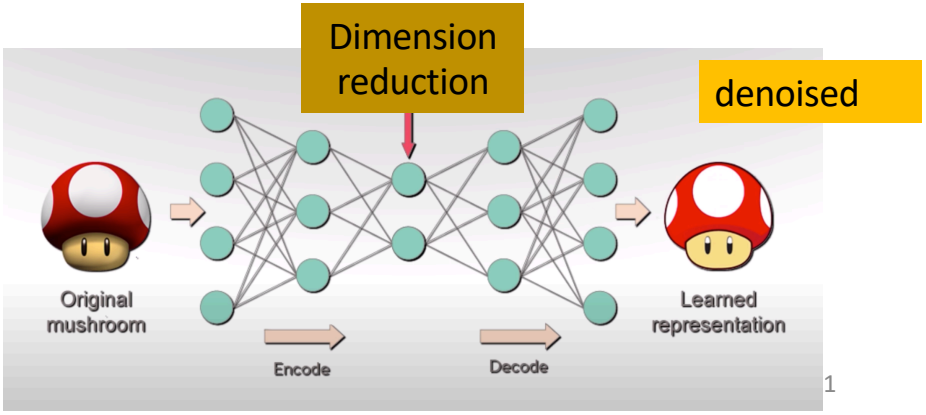
- NN do handle Large data, but mostly in image processing
- Most well known for supervised learning
  - Pattern recognition
  - Image generation
  - Disease classification/prediction using genome mutations
- The results in hidden layers are not easily interpretable
- How to do unsupervised learning?
  - Single cell study often uses Unsupervised learning, self organization
    - tSNE, UMAP, unsupervised clustering etc.
  - Auto-encoder

Autoencoder – one type of deep learning that use the input as the target to reproduce…

- Output is set to reconstruct the input
- Achieve few things
  - A latent space representing compressed information
  - Reconstructed output is cleaned from noise
  - Intermediate stages reflect some abstract features of the data
  - Through including penalties, hidden layers can be further engineered to desired features



Cost function = L_r ***(Input; Reconstructed)***



National Institute of Allergy and Infectious Diseases

The paper …

- Achieve many goals in one program
  - Scaling up
  - Dimension reduction
  - Imputation/denoising
- Implement penalties to achieve
  - Batch correction
  - Clustering

National Institute of Allergy and Infectious Diseases

nature methods

ARTICLES

https://doi.org/10.1038/s41592-019-0576-7

Exploring single-cell data with deep multitasking neural networks

Matthew Amodio[1,11], David van Dijk[1,2,11], Krishnan Srinivasan[1,11], William S. Chen[3], Hussein Mohsen[4], Kevin R. Moon[5], Allison Campbell[3], Yujiao Zhao[6], Xiaomei Wang[6], Manjunatha Venkataswamy[7], Anita Desai[7], V. Ravi[7], Priti Kumar[8], Ruth Montgomery[6], Guy Wolf[9,10,11] and Smita Krishnaswamy[1,2,11]*

It is currently challenging to analyze single-cell data consisting of many cells and samples, and to address variations arising from batch effects and different sample preparations. For this purpose, we present SAUCIE, a deep neural network that combines parallelization and scalability offered by neural networks, with the deep representation of data that can be learned by them to perform many single-cell data analysis tasks. Our regularizations (penalties) render features learned in hidden layers of the neural network interpretable. On large, multi-patient datasets, SAUCIE's various hidden layers contain denoised and batch-corrected data, a low-dimensional visualization and unsupervised clustering, as well as other information that can be used to explore the data. We analyze a 180-sample dataset consisting of 11 million T cells from dengue patients in India, measured with mass cytometry. SAUCIE can batch correct and identify cluster-based signatures of acute dengue infection and create a patient manifold, stratifying immune response to dengue.

Processing single-cell data of high dimensionality and scale is inherently difficult, especially considering the degree of noise, batch effects, artifacts, sparsity and heterogeneity in the data[1,2]. Furthermore, this effect becomes exacerbated as one tries to compare between samples, which themselves contain noisy heterogeneous compositions of cellular populations. Deep learning offers promise as a technique for handling the size and dimensionality of modern biological datasets. However, deep learning has been underused for unsupervised exploratory tasks.

In this paper, we use a regularized autoencoder, which is a neural network that learns to recreate its own input via a low-dimensional bottleneck layer that learns representations of the data and enables a denoised reconstruction of the input[3–7].

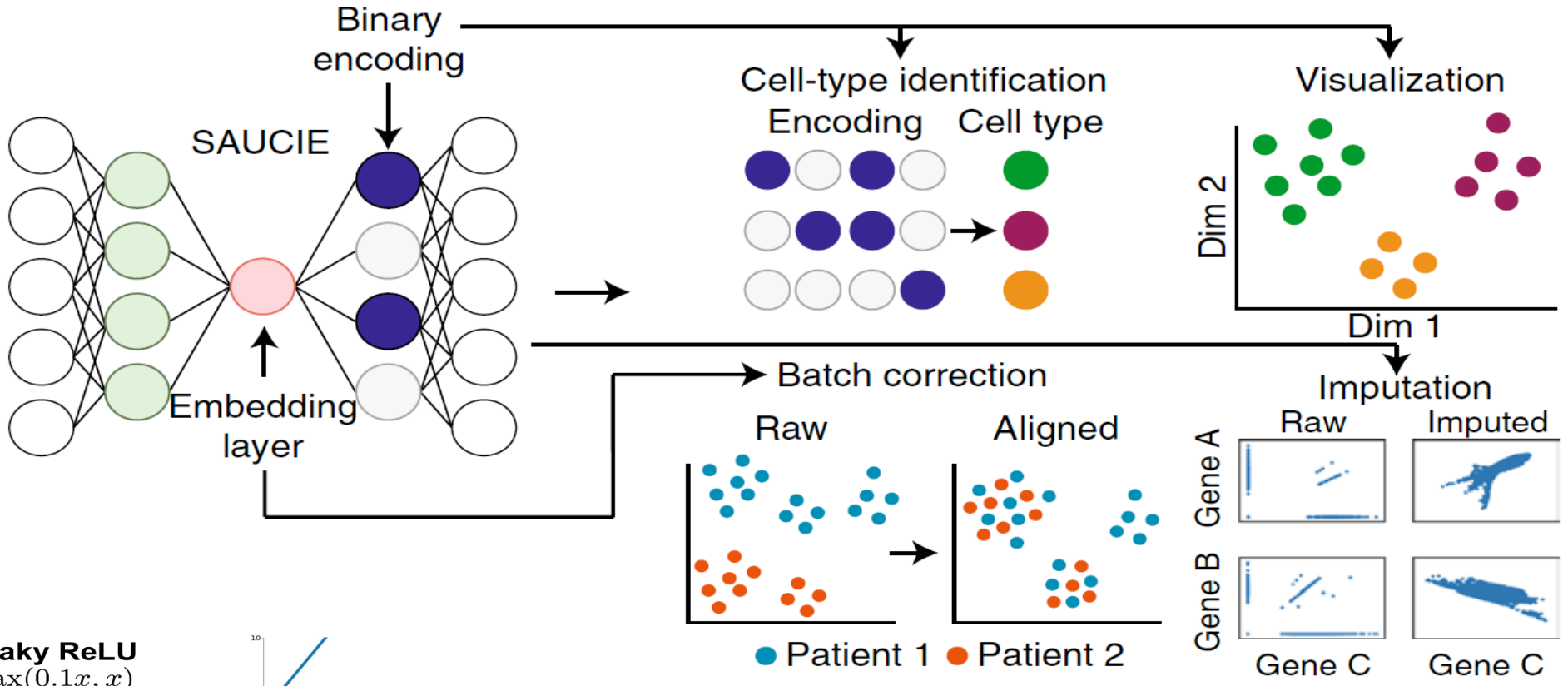
Since autoencoders learn their own features, they can reveal structure in the data without defining or explicitly learning a

We apply SAUCIE to the batch correcting, denoising and clustering of an 11-million cell mass cytometry dataset with 180 samples from 40 subjects in a study of the dengue flavivirus and see the proportions of subpopulations.

Results

The SAUCIE architecture and layer regularizations. To enable unsupervised learning in a scalable manner, we base our method on the autoencoder. A key challenge is to extract meaning from the model's internal representation of the data. Specifically, we seek representations in hidden layers that are useful for performing the various analysis tasks associated with single-cell data. Here, we introduce several design decisions and novel regularizations to our autoencoder architecture (Fig. 1) to constrain the learned representations for four key tasks: clustering, batch correction, denoising

Caution – these are not achieved in one step …



Leaky ReLU
$\max(0.1x, x)$

In fact, there are two steps to achieve different tasks…

- Using the first run with autoencoder to do batch correction and imputation
- Using the batch corrected/normalized expression matrix from the first step to train autoencoder again to do clustering and dimension reduction/2D visualization

Step 1, correction of batch effect (and imputation naturally)

L $_r$: Regular reconstruction loss (SSE)
L $_b$: regularization/paneity on the batch differences using MMD.

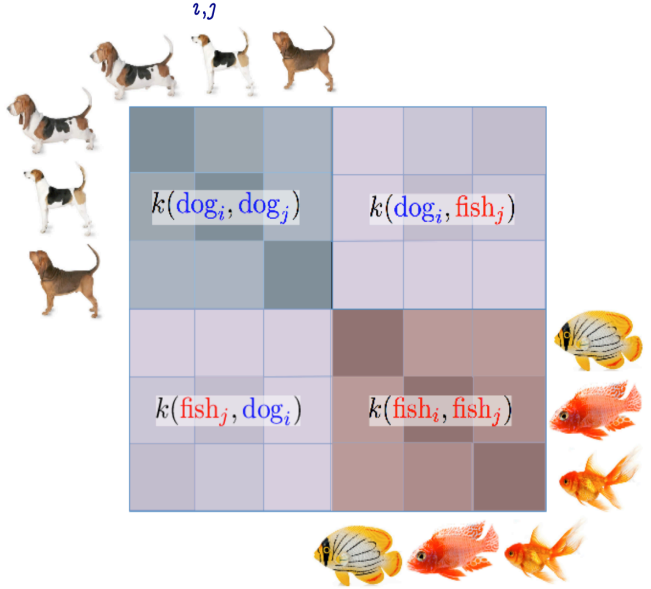
Cost function I: $L = L_r(X; \hat{X}) + \lambda_b \cdot L_b(V)$

$$L_r(X; \hat{X}) = \frac{1}{n} \sum_{i=1}^{n} ||\mathrm{xi} - \hat{\mathrm{x}}_\mathrm{i}||^2$$

Maxium Mean Difference:

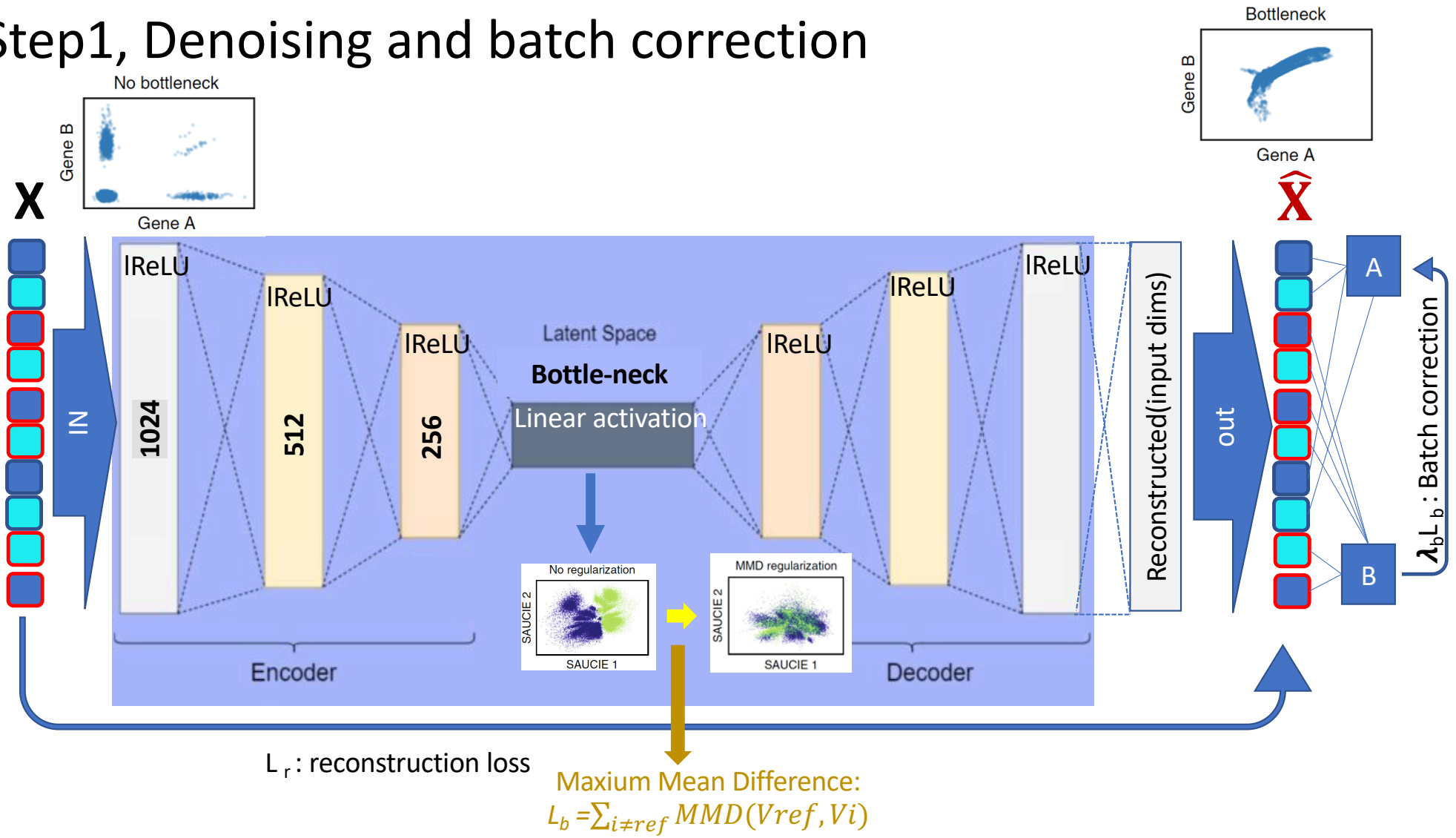
$$L_b = \sum_{i \neq ref} MMD(Vref, Vi)$$

$$\widehat{MMD}^2 = \frac{1}{n(n-1)} \sum_{i \neq j} k(\mathrm{dog}_i, \mathrm{dog}_j) + \frac{1}{n(n-1)} \sum_{i \neq j} k(\mathrm{fish}_i, \mathrm{fish}_j) - \frac{2}{n^2} \sum_{i,j} k(\mathrm{dog}_i, \mathrm{fish}_j)$$



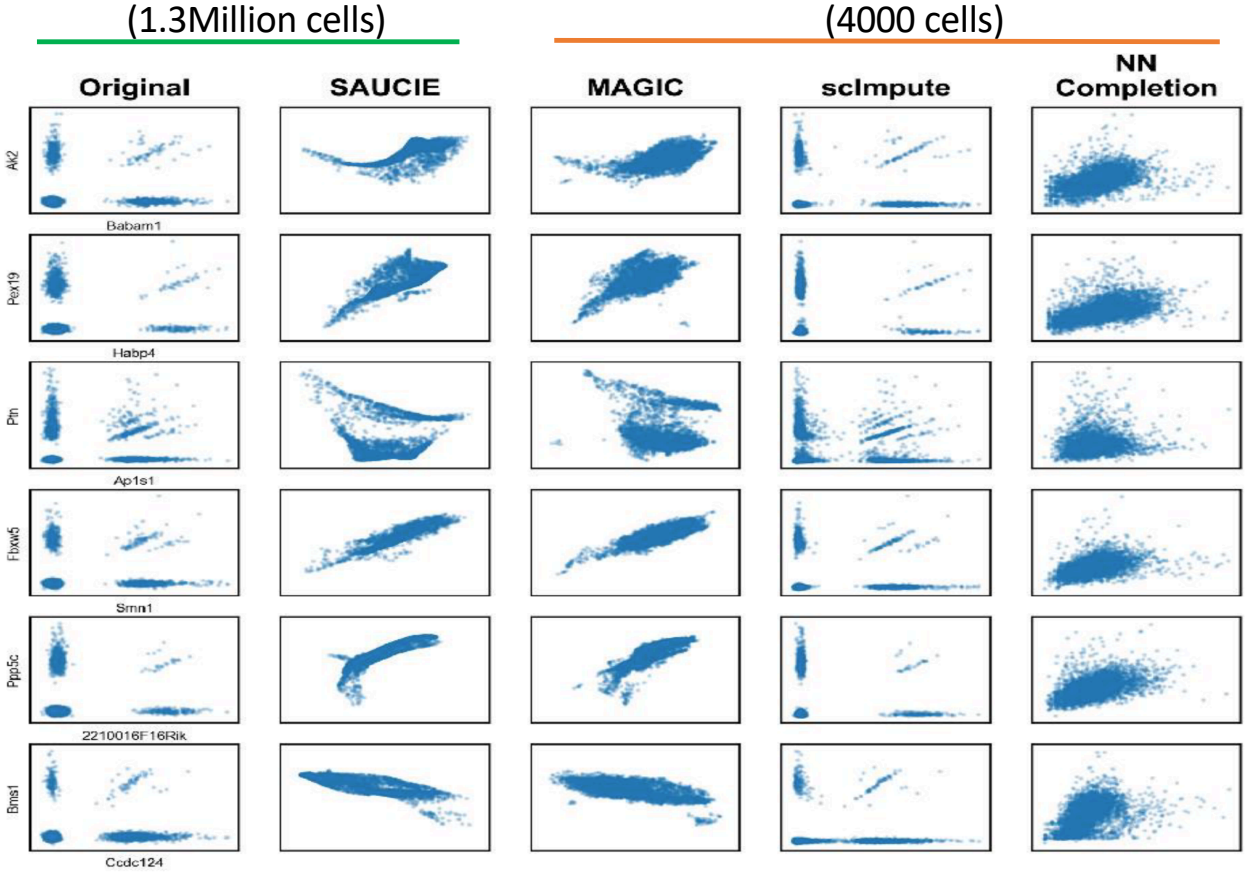
National Institute of Allergy and Infectious Diseases

Step1, Denoising and batch correction



Imputation results – no ground truth

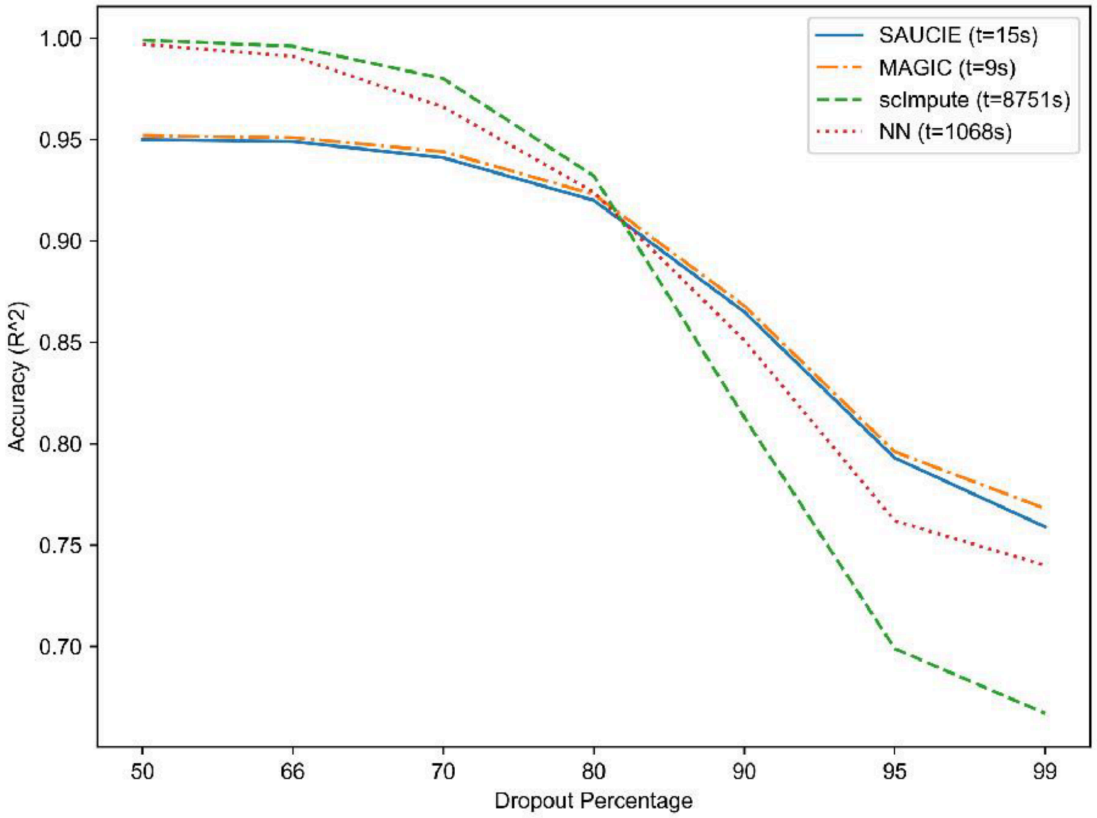
- Without ground truth
- SAUCIE Produced similar results as MAGIC, which is from a different algorithm
- SAUCIE Produced different relationships between genes, unlike NN completion



National Institute of Allergy and Infectious Diseases

Imputation with ground truth

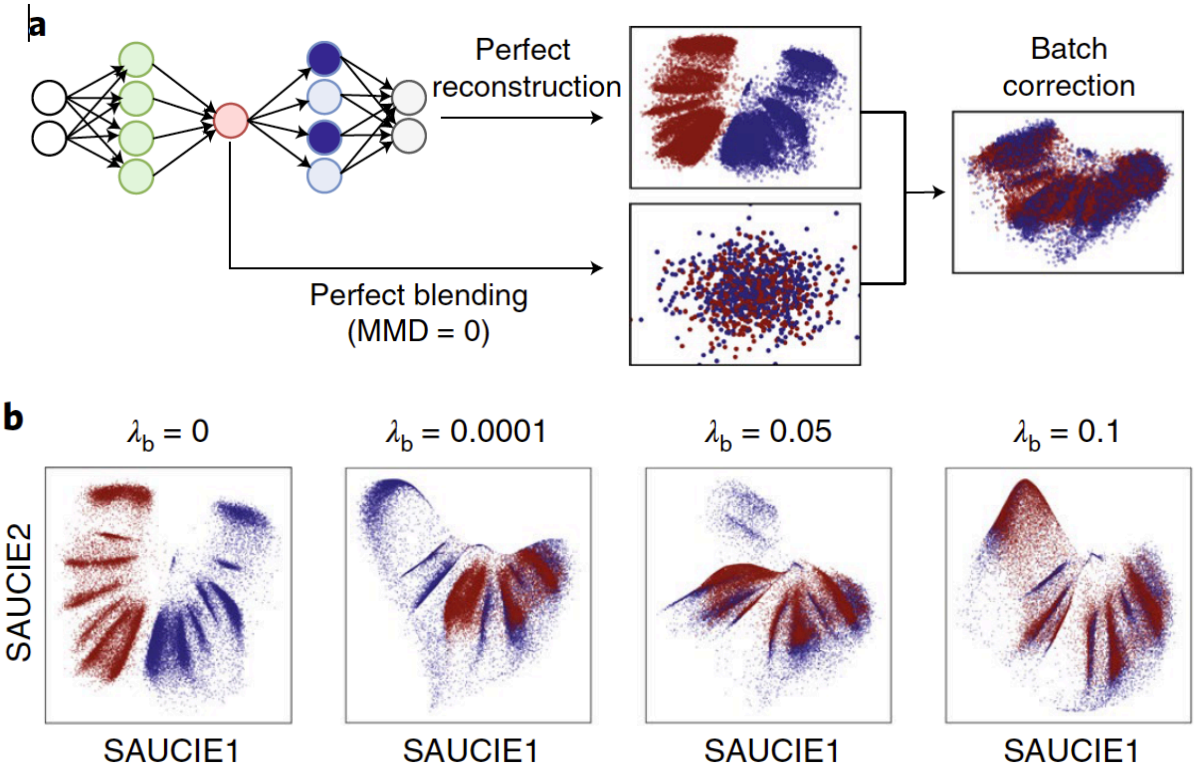
- With ground truth
  - Data from bulk RNA-seq data, where there is no/minimal dropout
  - Transcripts are throughout randomly creating artificial dropout
- SAUCIE is working fine with as much as 99% dropout rate



National Institute of Allergy and Infectious Diseases

For batch correction: one has to take good care about the hyper parameters

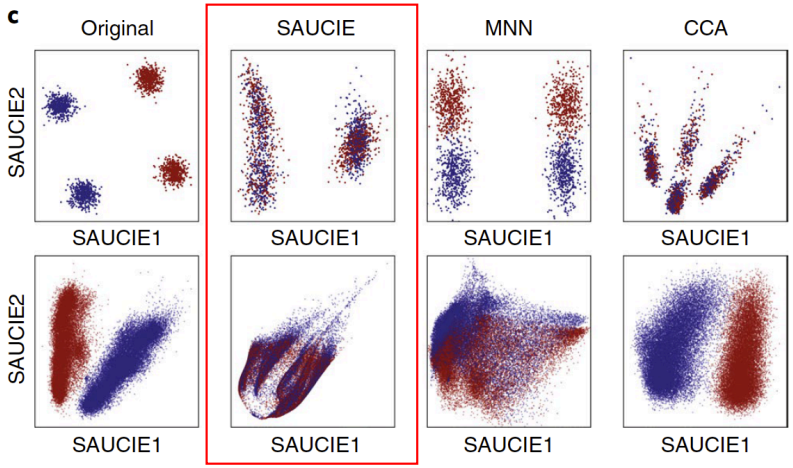
- This method make assumption that between batches, the cell type distribution should be similar. Which might not be the case in reality. Do based on MMD may lead to artificial merging.



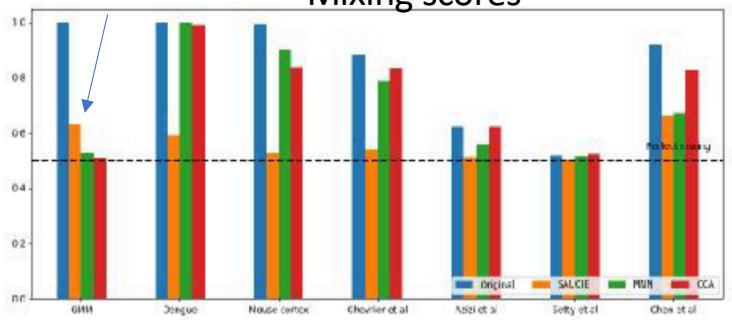
Cost function I: $L = L_r(X; \hat{X}) + \lambda_b \cdot L_b(V)$

Batch correction: Where is the ground truth?

- SAUCIE is most powerful in bring different batches together.
- Used quantitative measures
  - mixing score (matching batches)
  - shape preservation of the clusters

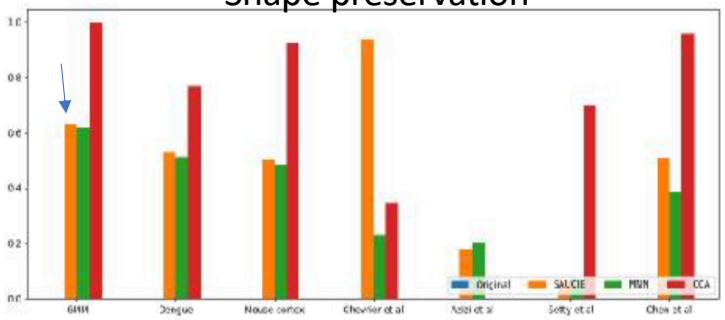


Mixing scores



Datasets

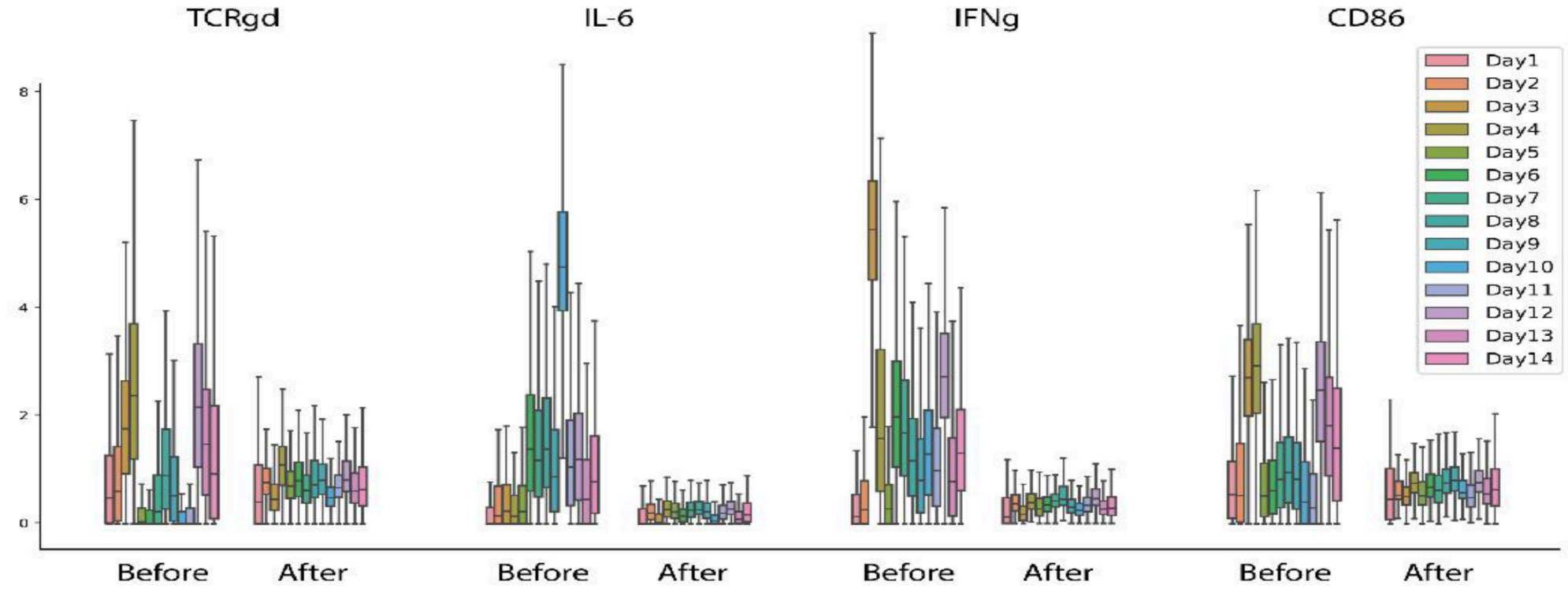
Shape preservation



Datasets

National Institute of Allergy and Infectious Diseases

Artificial effects corrected: Bias from processing dates



National Institute of Allergy and Infectious Diseases

Step 2, cluster identification and dimension reduction

Cost function II: $L = L_r(\hat{X}; \tilde{X}) + \lambda_c \cdot L_c(B) + \lambda_d \cdot L_d(B; \hat{X})$

Information dimension regularization:

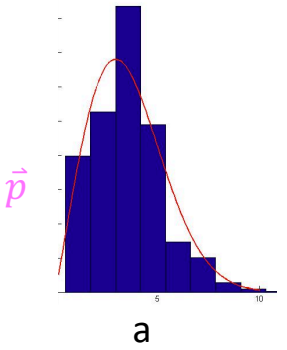
(λ_c control granularity of clusters: $L_c(B)$ encourage sparse and binary activation of neurons in the B layer)

$$L_c(B) = -\sum_{j=1}^{k} p_j \log p_j$$

Where p is the normalized distribution of neural activation in the B layer.

A normalized activation distribution

$\vec{p} = \vec{a} / ||\vec{a}||_1$, where $||\vec{a}||_1 := \sum_{i=1}^{n} |a_i|$



a

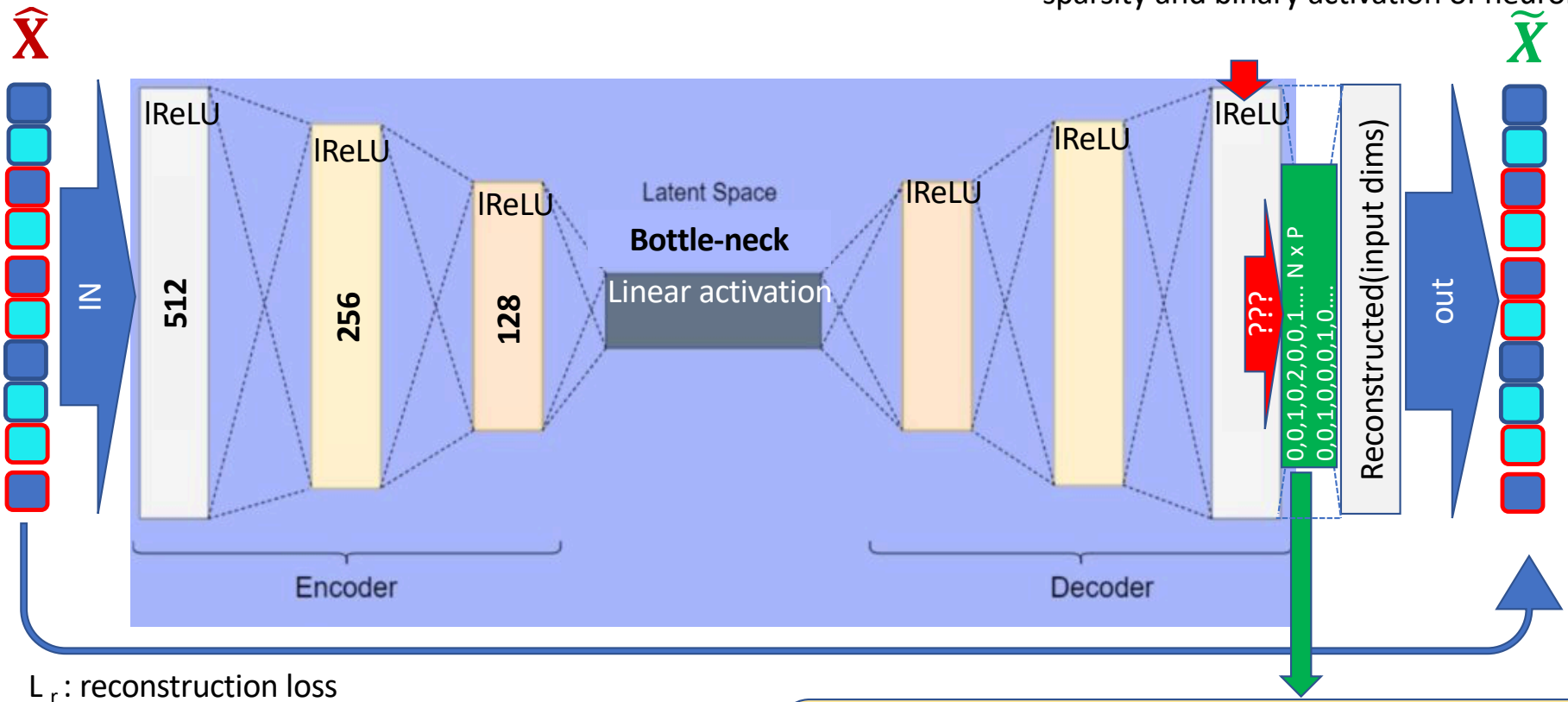
Minimize Intra cluster difference $L_d(B; \hat{X}) = \sum_{i,j:bi=bj} ||\hat{x}_i - \hat{x}_j||^2$

(sum of Euclidean distance)

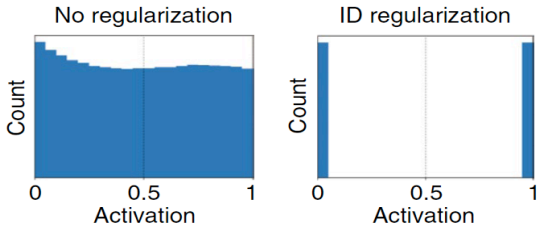
Norms: https://en.wikipedia.org/wiki/Norm_(mathematics)

Step2, 2D-visualization and Clustering

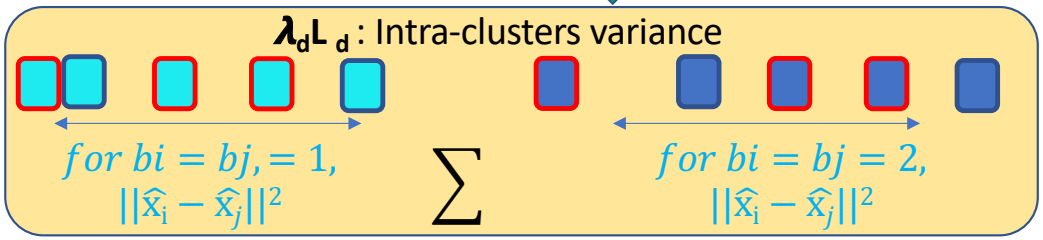
$\lambda_c L_c$: ID regularization, Encourage sparsity and binary activation of neurons



L_r : reconstruction loss



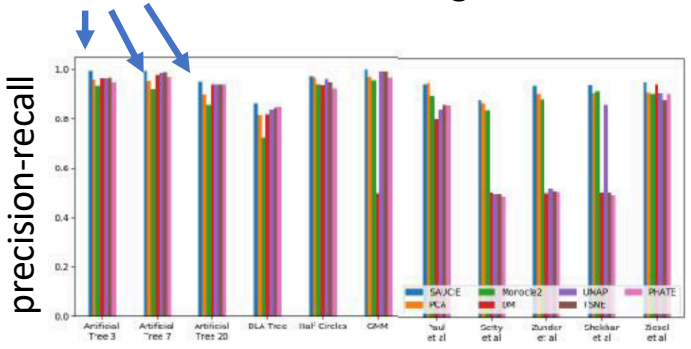
$\lambda_d L_d$: Intra-clusters variance



Dimension reduction for 2D visualization

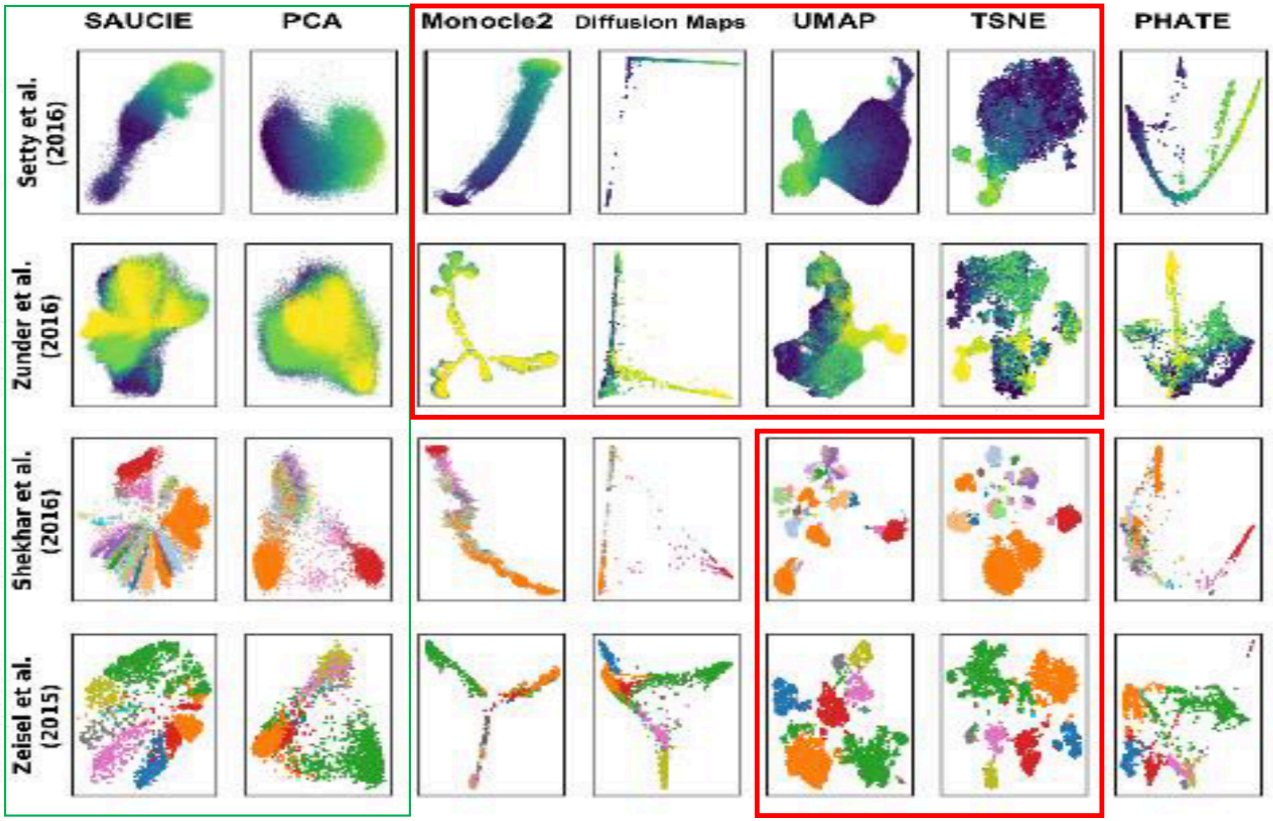
- Use precision-recall
  - Consistency of nearest neighbors in original space vs in 2d

Different methods, first being SAUCIE

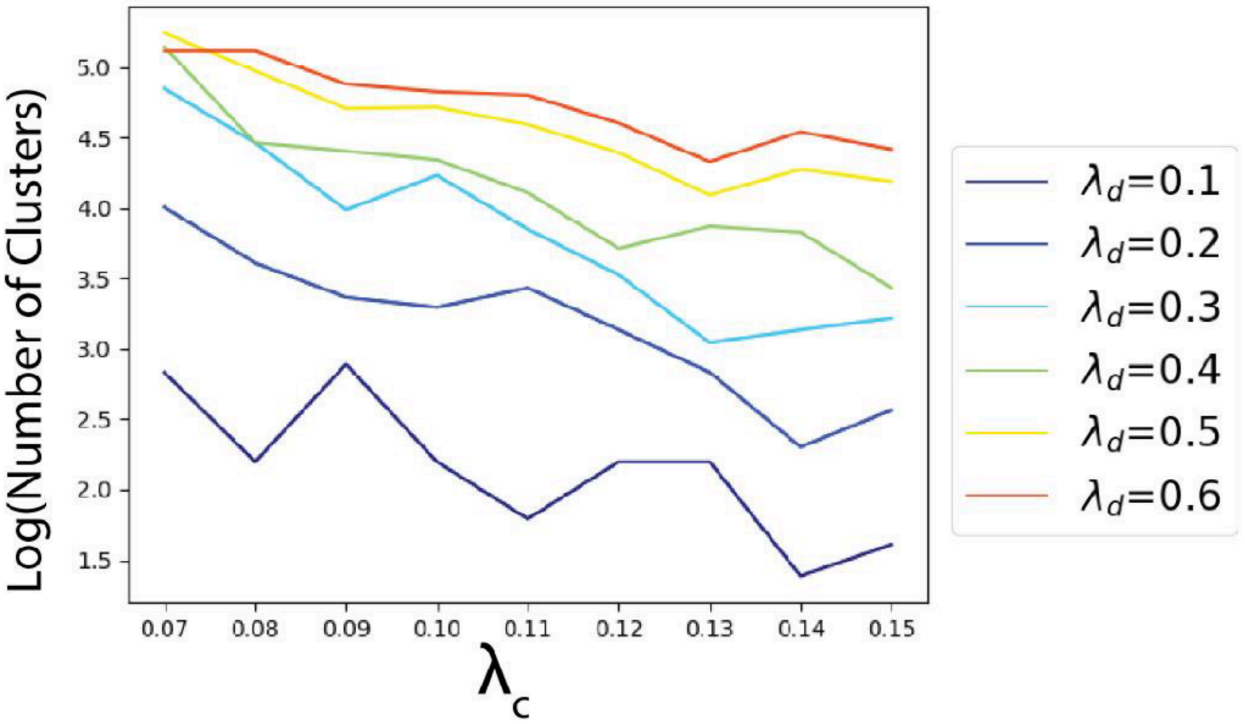


Datasets

National Institute of Allergy and Infectious Diseases



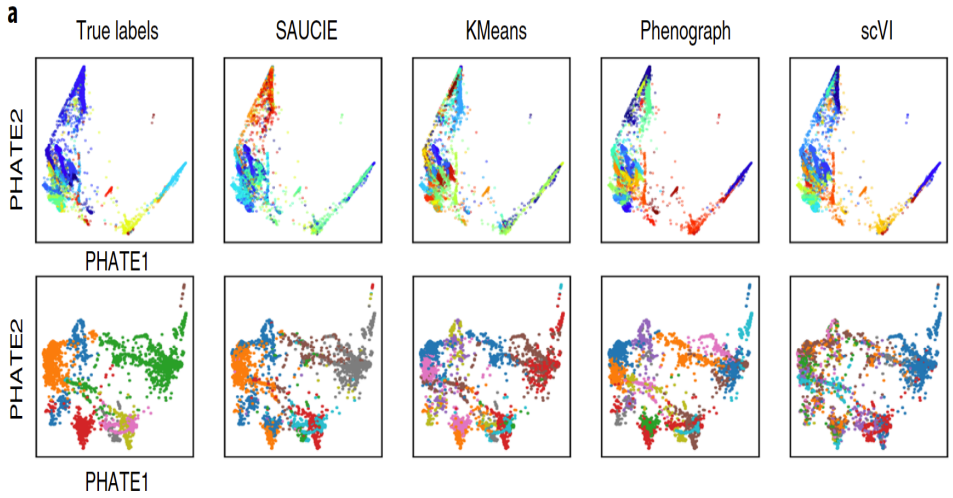
How parameters affects the number of clusters

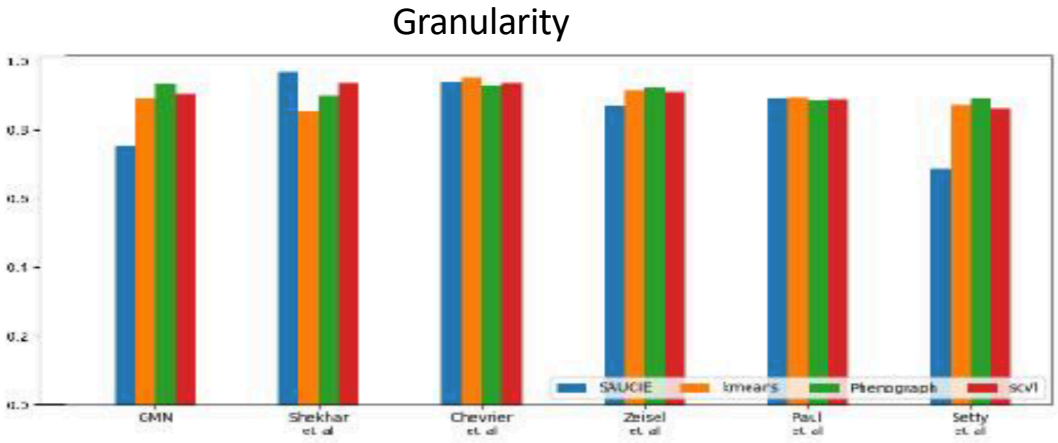


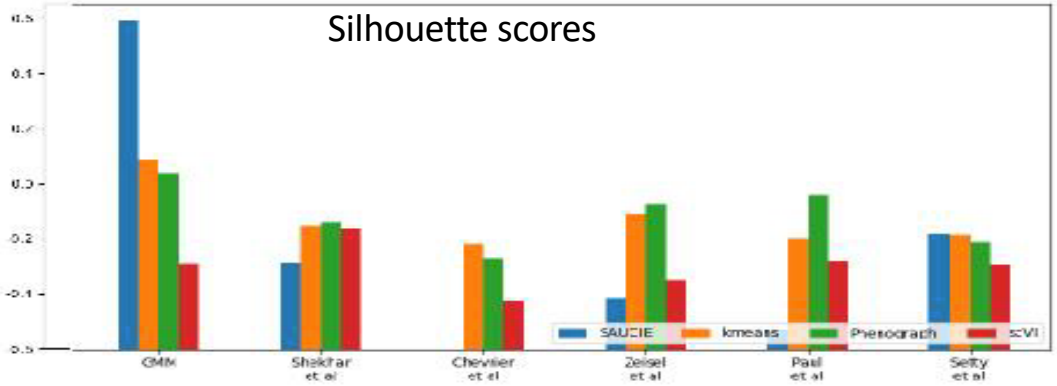
National Institute of Allergy and Infectious Diseases

Clustering results

- Visualization is not very good
- Quantitative measures
  - Granularity
  - Silhouette scores





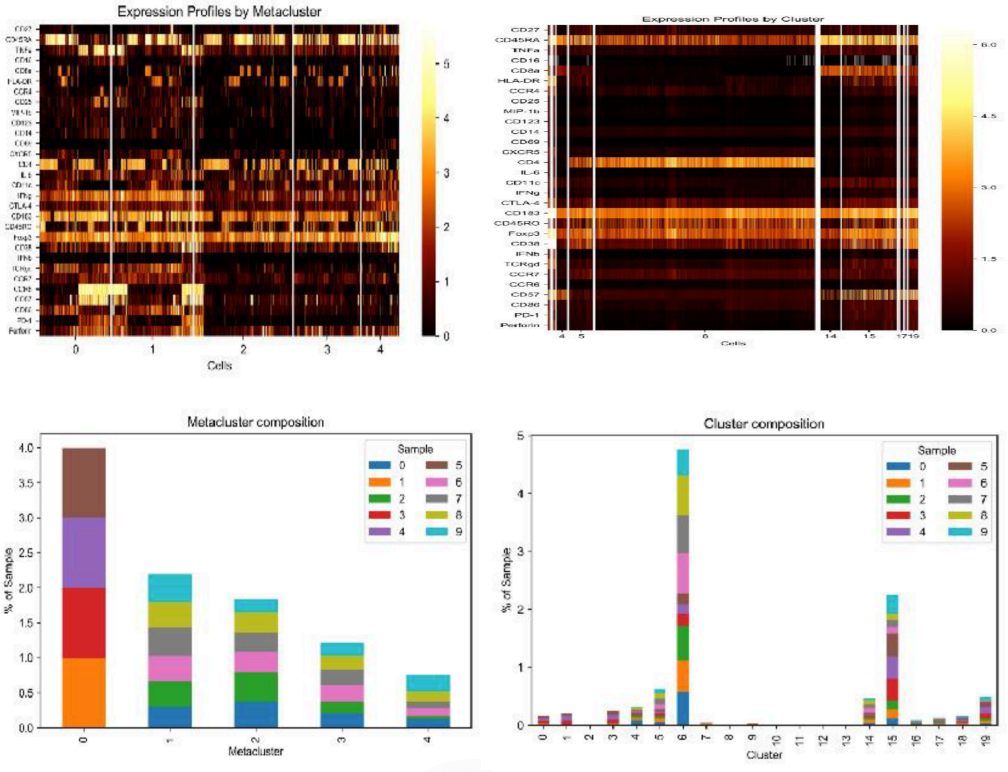


Clustering: Able to bring different patients together

- Phenograph's clusters are strongly biased by between-patient variations
- SAUCIE cluster get more even distribution of patients to each cluster
- The SAUCIE heatmap is also smoother

This effect may due to batch correction

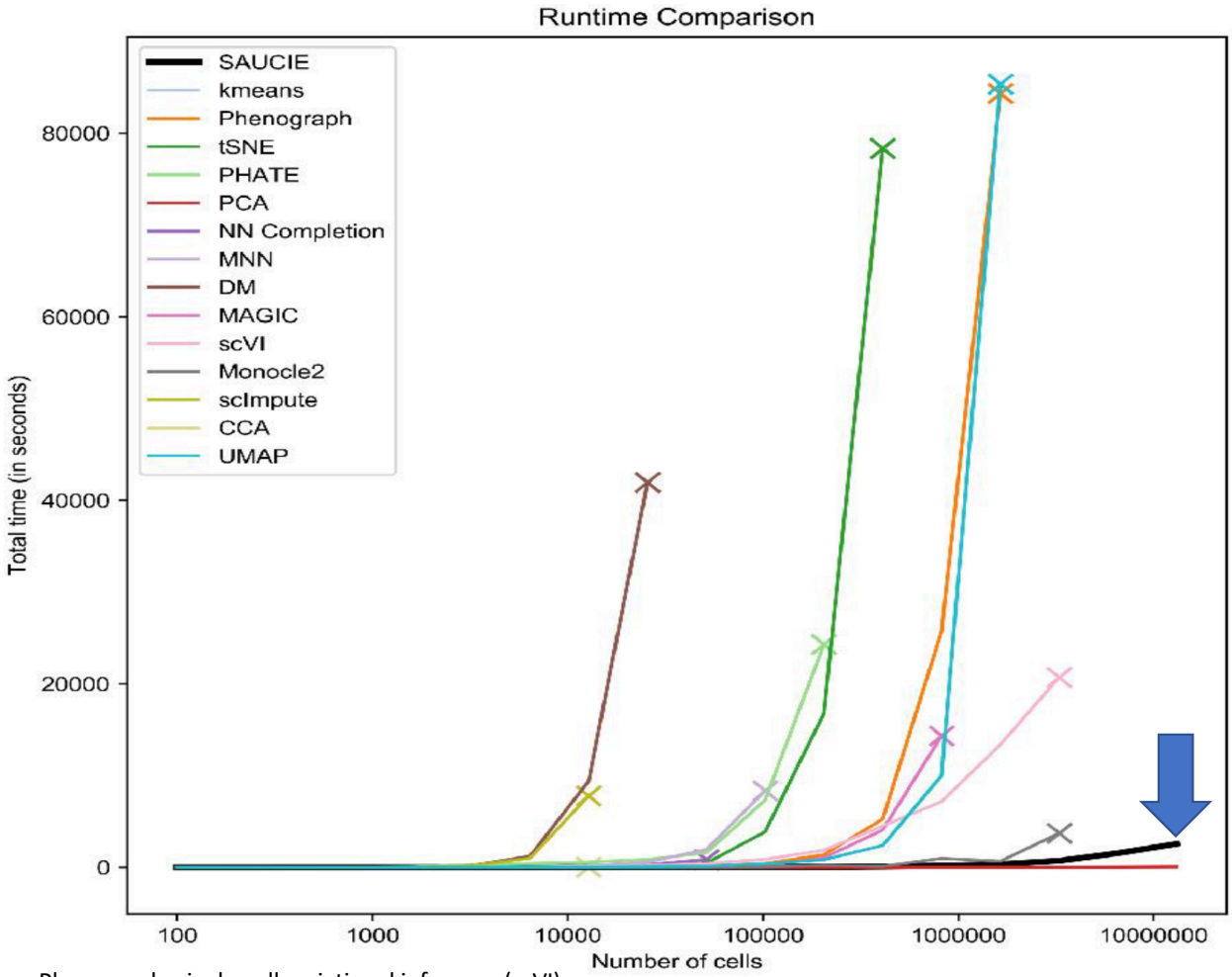
Not from clustering



National Institute of Allergy and Infectious Diseases

Key Advantages

- Advantages
  - Performed comparatively and better
  - Perform all functions in one method

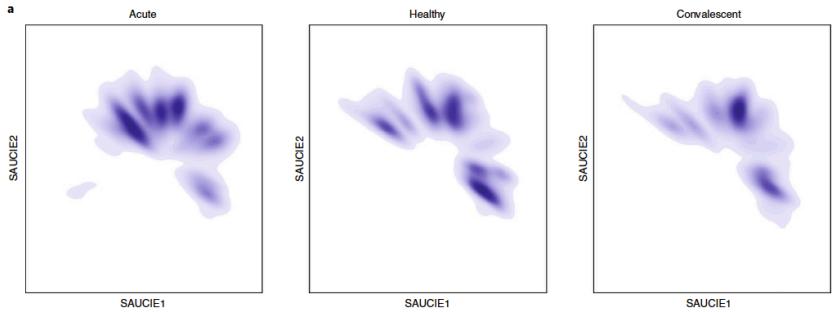


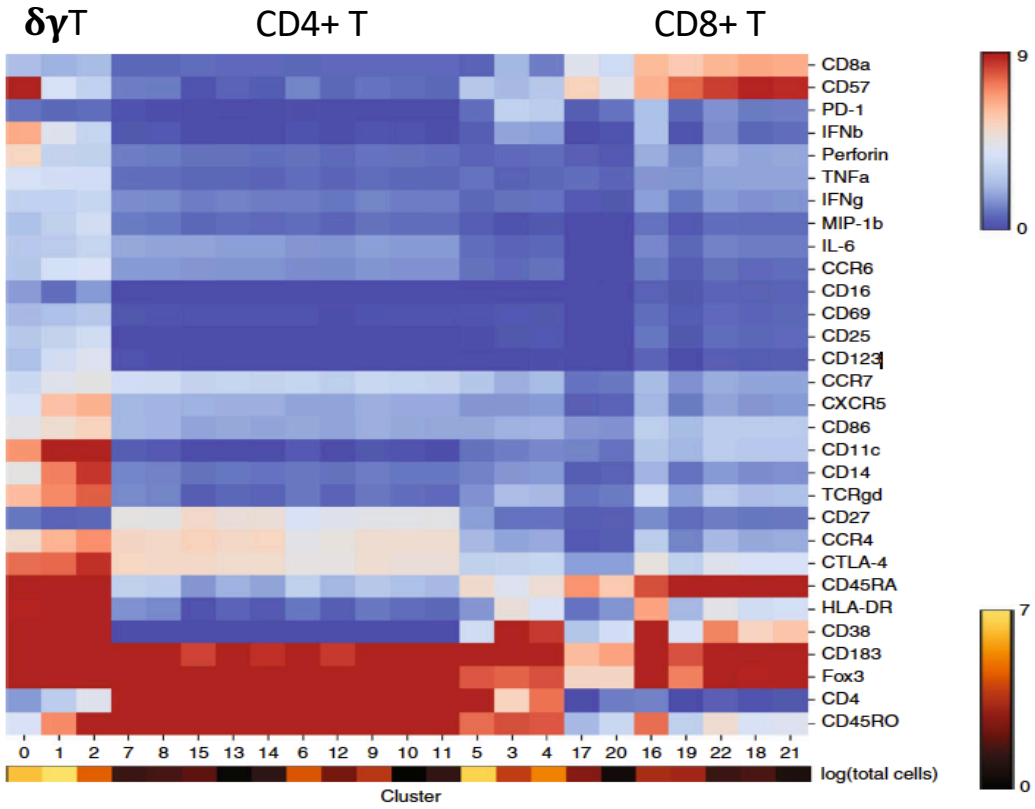
National Institute of Allergy and Infectious Diseases

Clustering: Phenograph, single-cell variational inference (scVI)
Batch correction: mutual nearest neighbors (MNN), canonical correlation analysis (CCA)
Visualization: PCA, Monocle2, diffusion maps, UMAP, tSNE and PHATE
Imputation: MAGIC, scImpute and nearest neighbors completion (NN completion)

CyTOF data on dengue patients

- An manifold at cell level showed differences across different conditions.
- Clusters identified including subpopulations of Cd4, Cd8 and rare regulatory **δγ**T cells
- The biggest difference being different batch of sequencing





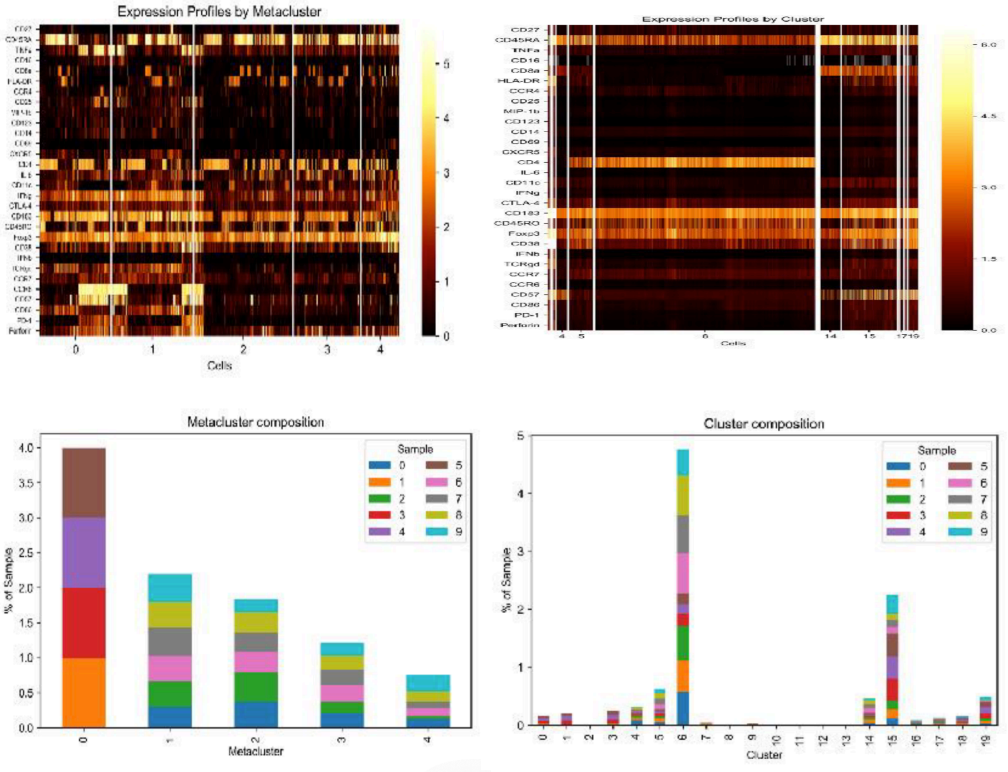
National Institute of Allergy and Infectious Diseases

Clustering: Able to bring different patients together

- Phenograph's clusters are strongly biased by between-patient variations
- SAUCIE cluster get more even distribution of patients to each cluster
- The SAUCIE heatmap is also smoother

This effect may due to batch correction

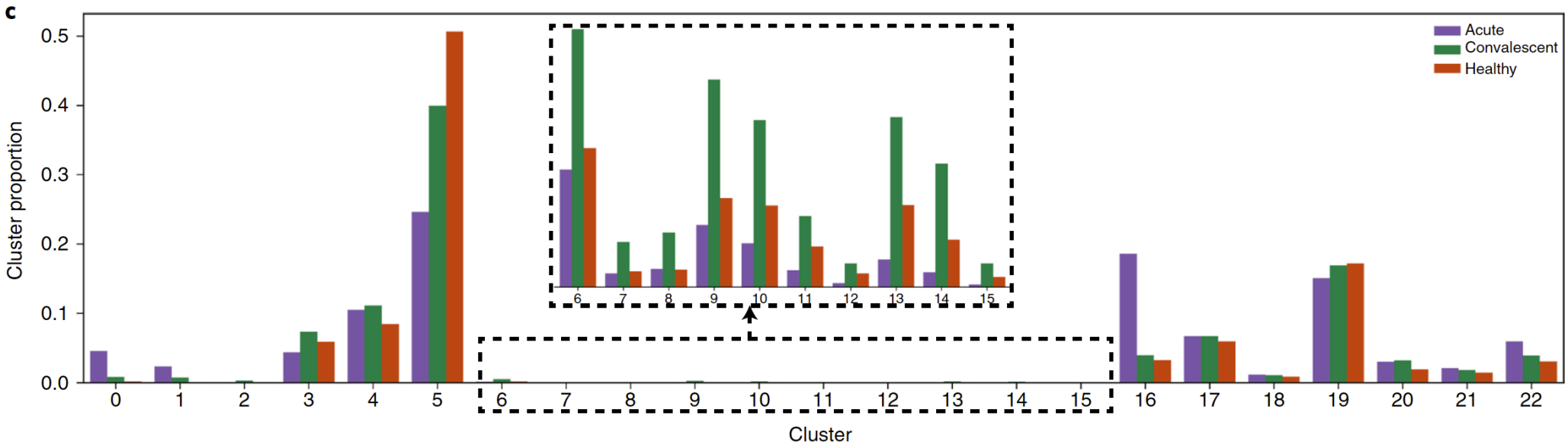
Not from clustering



National Institute of Allergy and Infectious Diseases

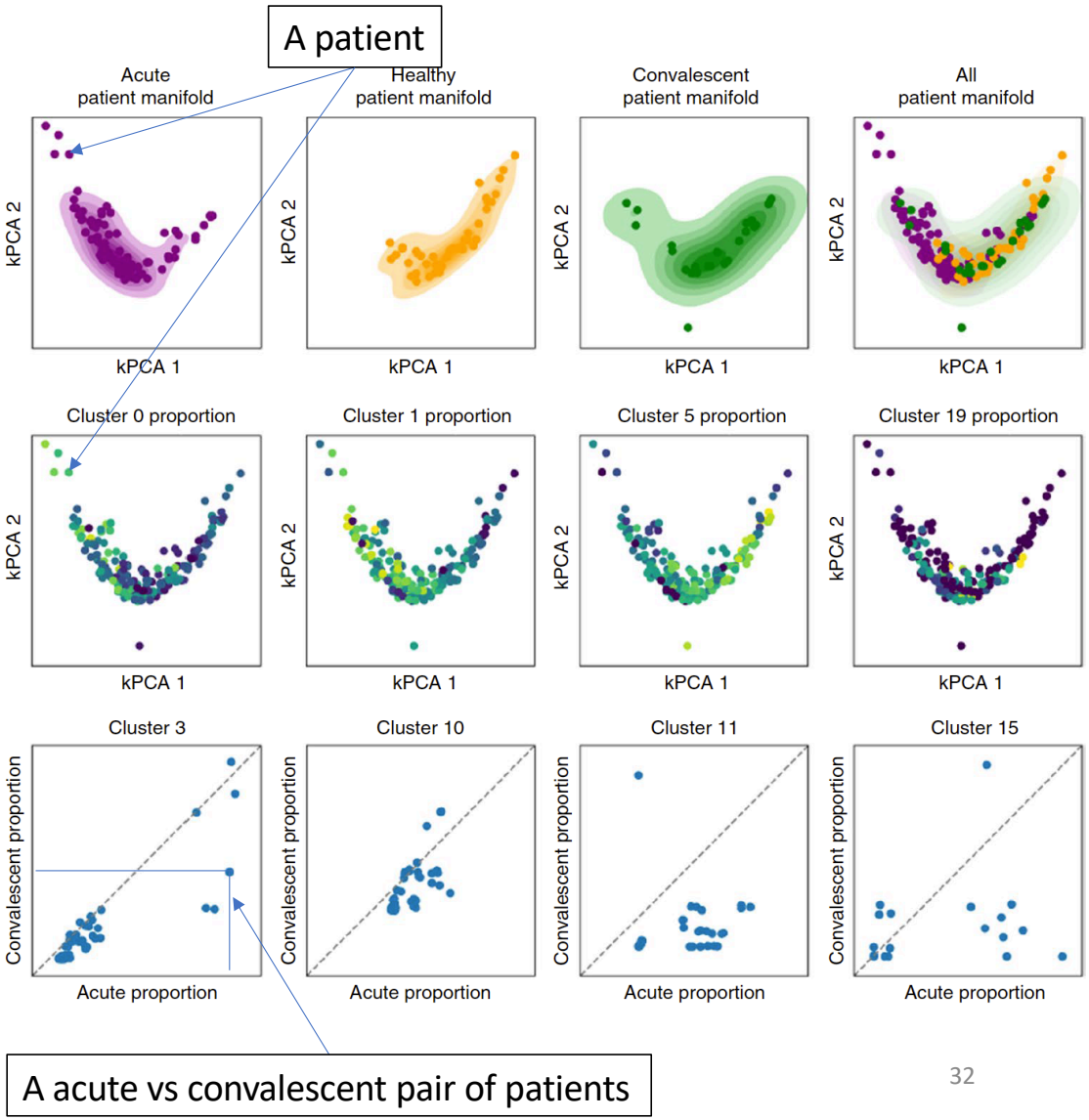
Presence of cell clusters in disease states

- Each patient has different proportion of cells in each cluster
- 180 patients x 23 proportion of cells in each cluster: forms a matrix
- A PCA of patient as a datapoint would show the manifold over patients.



Patient manifold

- Acute patients and healthy controls segregate differently.
- Cluster 0 and 1 contribute to acute patients
  - Delta-Gama T cells characterize them
- Cluster 5 contribute to healthy patients
  - CD4 T cells that has no activation/inflammation markers
- Examine patient pairs identify relative changes in proportion of cell types along disease progression



National Institute of Allergy and Infectious Diseases

Summary—an nice experiment but not perfect

- Autoencoder could extract useful information for single cell data.
- Autoencoder could support multi-tasking
- The results shows some utility of it on real data, including
  - Multiple batch correction, Imputation give similar but much faster results than MAGIC and many others, visulization, clustering is not so well
  - Caution: batch correction has to go together with a well balanced batches
- Scales well, can handle millions of single cell RNA-seq data and tens of millions of single cell CyTOF data
- A lot of parameters to twist.

Conclusion

- A nice idea with some good results.
- Caution with application, batch-correction should combine with good design.

Thank you!

- Ziv and Karthik helped me to understand U net and neural network
- Claire and Dan helped to determine the interest
- Dan and Poorani helped me to trouble-shoot the python codes to run SAUCIE