Mapping single-cell data to reference atlases by transfer learning

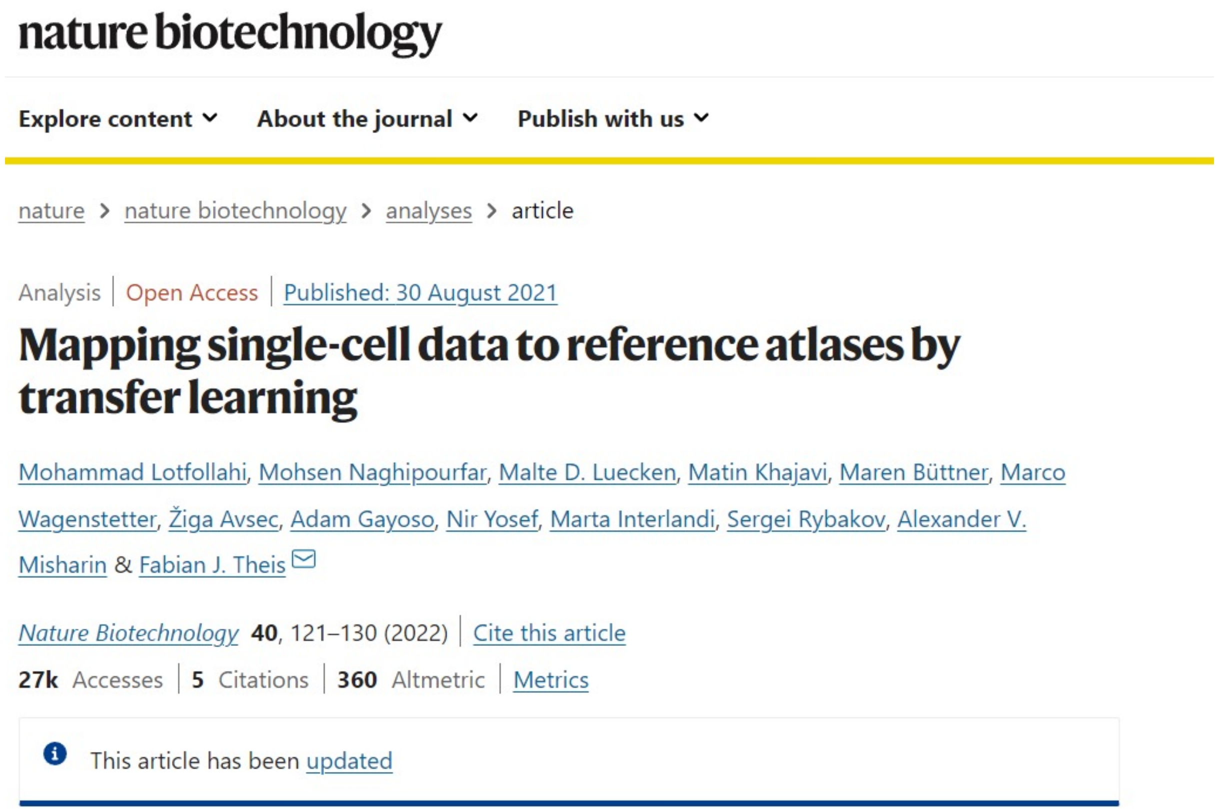
Shang Gao, Ph.D candidate

Advisor: Dr. Yang Dai, Dr. Jalees Rehman

CBQB journal club, Feb. 7th

Introduction

- The authors developed the deep learning strategy **scArches** for mapping query datasets onto reference datasets for the purpose of data integration and the annotation or identification of cell types.



nature biotechnology

nature › nature biotechnology › analyses › article

Analysis | Open Access | Published: 30 August 2021

Mapping single-cell data to reference atlases by transfer learning

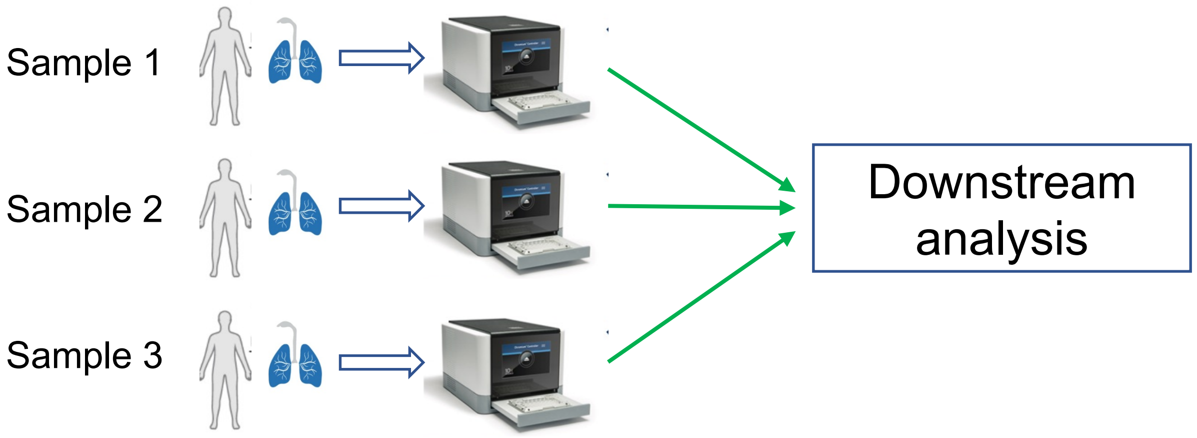
Mohammad Lotfollahi, Mohsen Naghipourfar, Malte D. Luecken, Matin Khajavi, Maren Büttner, Marco Wagenstetter, Žiga Avsec, Adam Gayoso, Nir Yosef, Marta Interlandi, Sergei Rybakov, Alexander V. Misharin & Fabian J. Theis ✉

Nature Biotechnology **40**, 121–130 (2022) | Cite this article

27k Accesses | **5** Citations | **360** Altmetric | Metrics

ℹ This article has been updated

Motivation



Sample 1

Sample 2

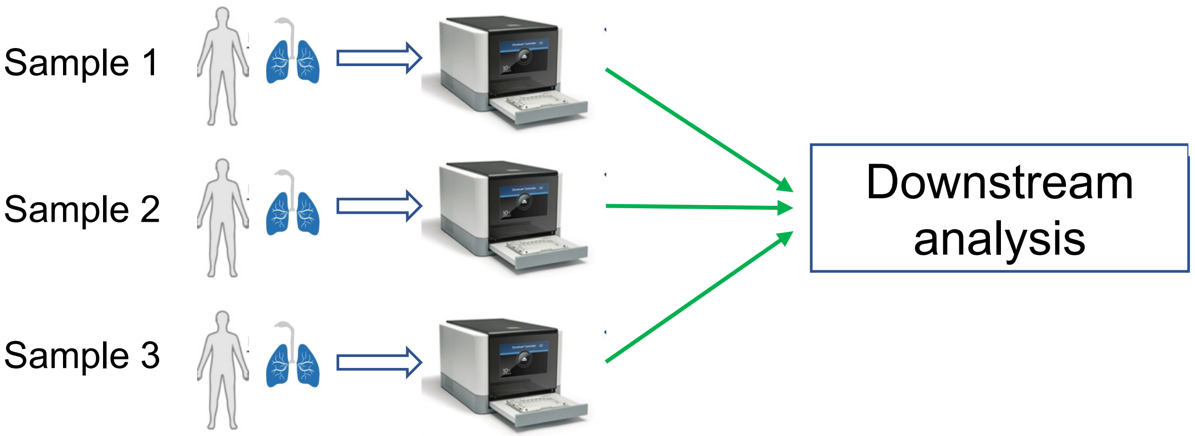
Sample 3

Downstream analysis

CHAN ZUCKERBERG BIOHUB Stanford

HUMAN LUNG CELL ATLAS

Motivation

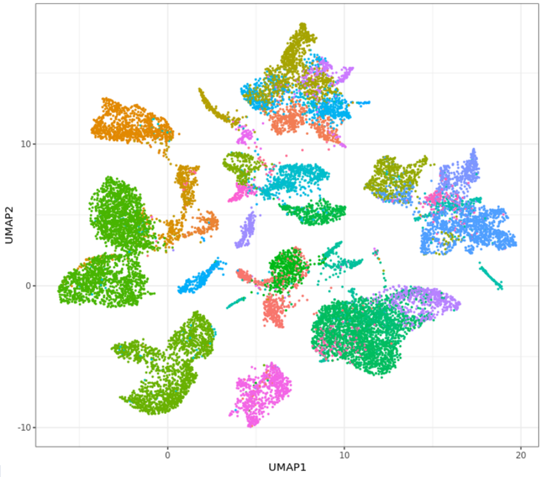

Sample 1

Sample 2

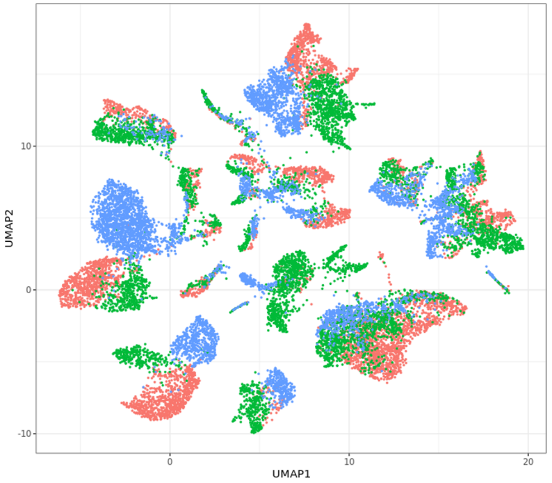
Sample 3

Downstream analysis

UMAP of 3 samples
Color: Cell types

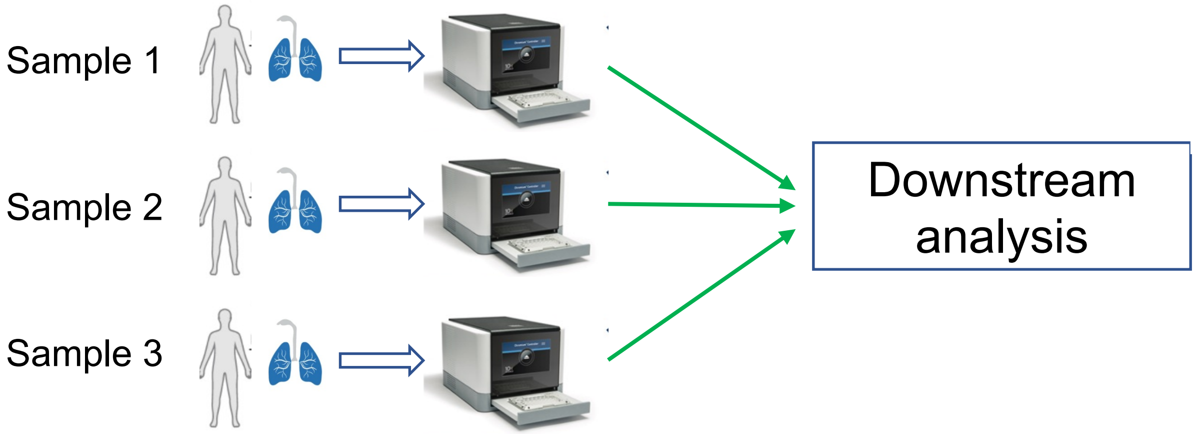


UMAP of 3 samples
Color: Batches

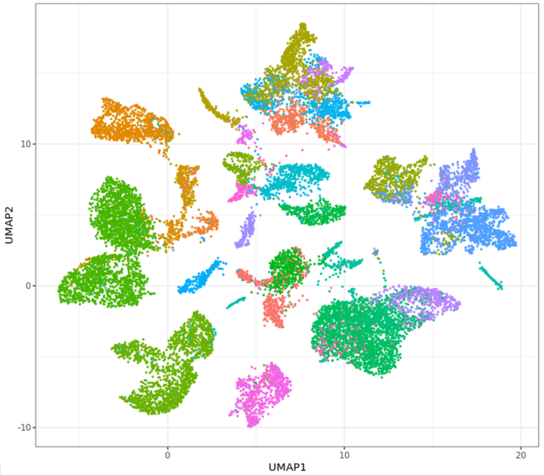


HUMAN LUNG CELL ATLAS

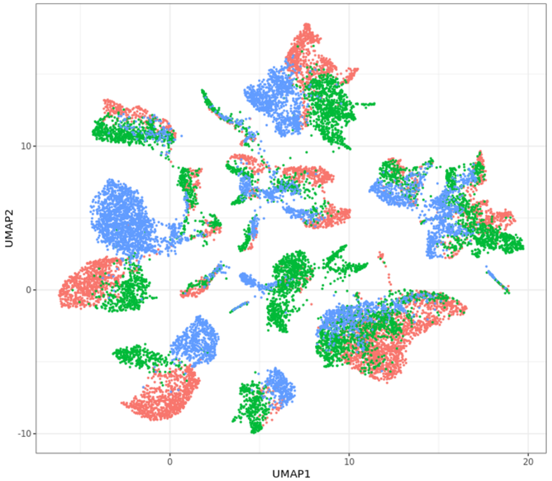
Motivation



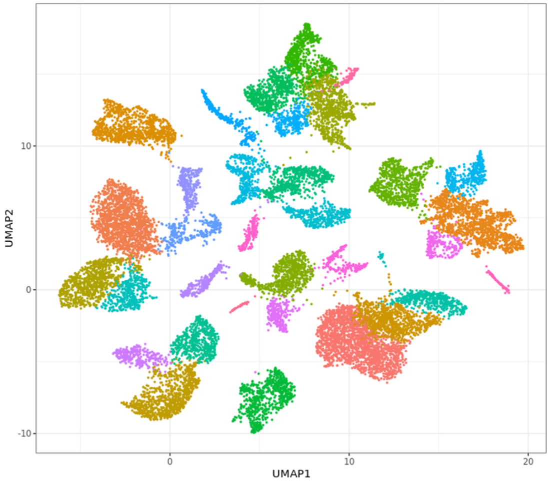
UMAP of 3 samples
Color: Cell types



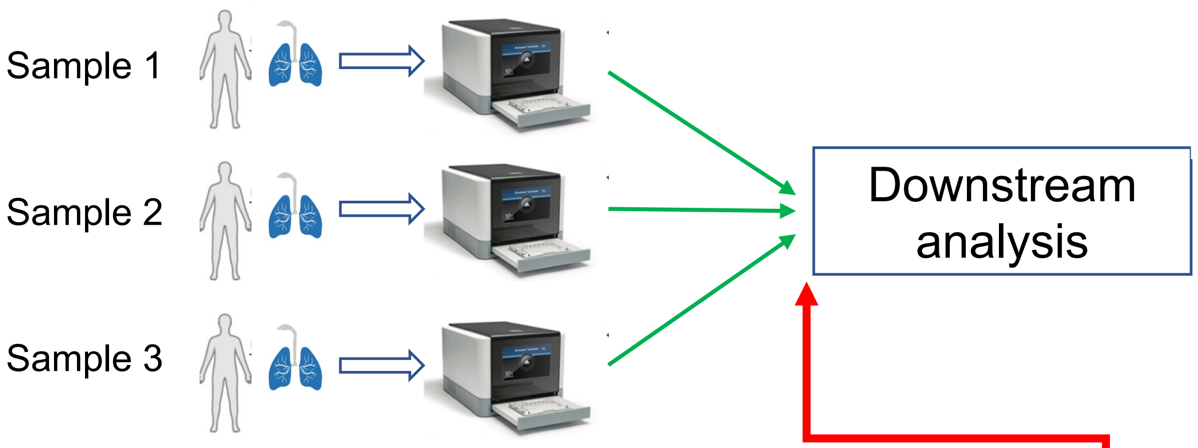
UMAP of 3 samples
Color: Batches



UMAP of 3 samples
Color: Seurat clusters



Motivation



Sample 1

Sample 2

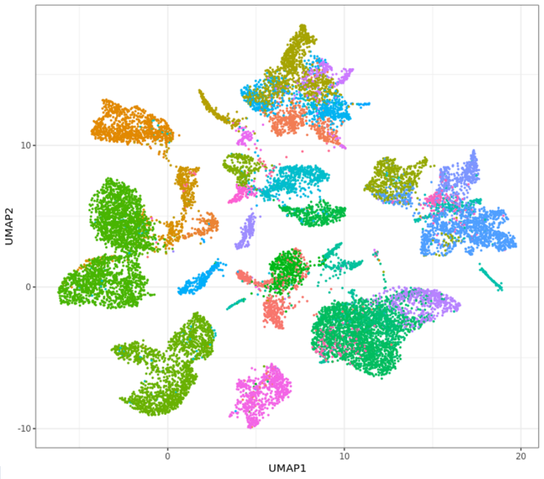
Sample 3

Downstream analysis

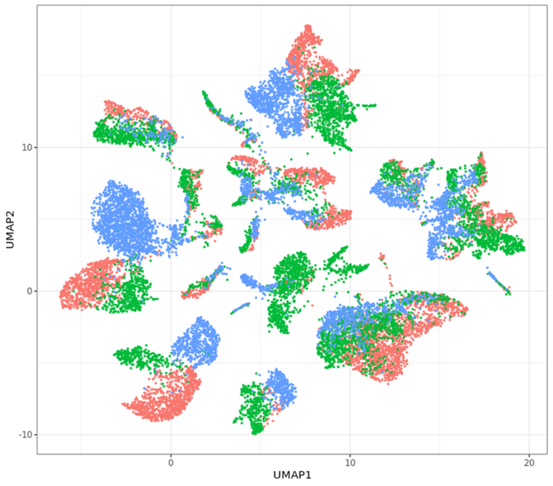
HUMAN LUNG CELL ATLAS

Integration tools should be applied to eliminate the batch effect before the downstream analysis.

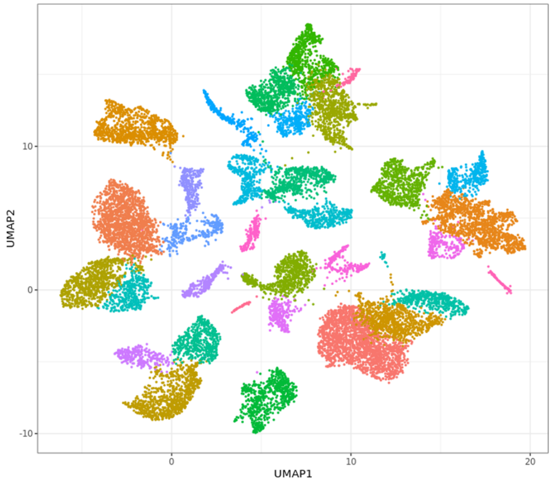
UMAP of 3 samples
Color: Cell types



UMAP of 3 samples
Color: Batches



UMAP of 3 samples
Color: Seurat clusters



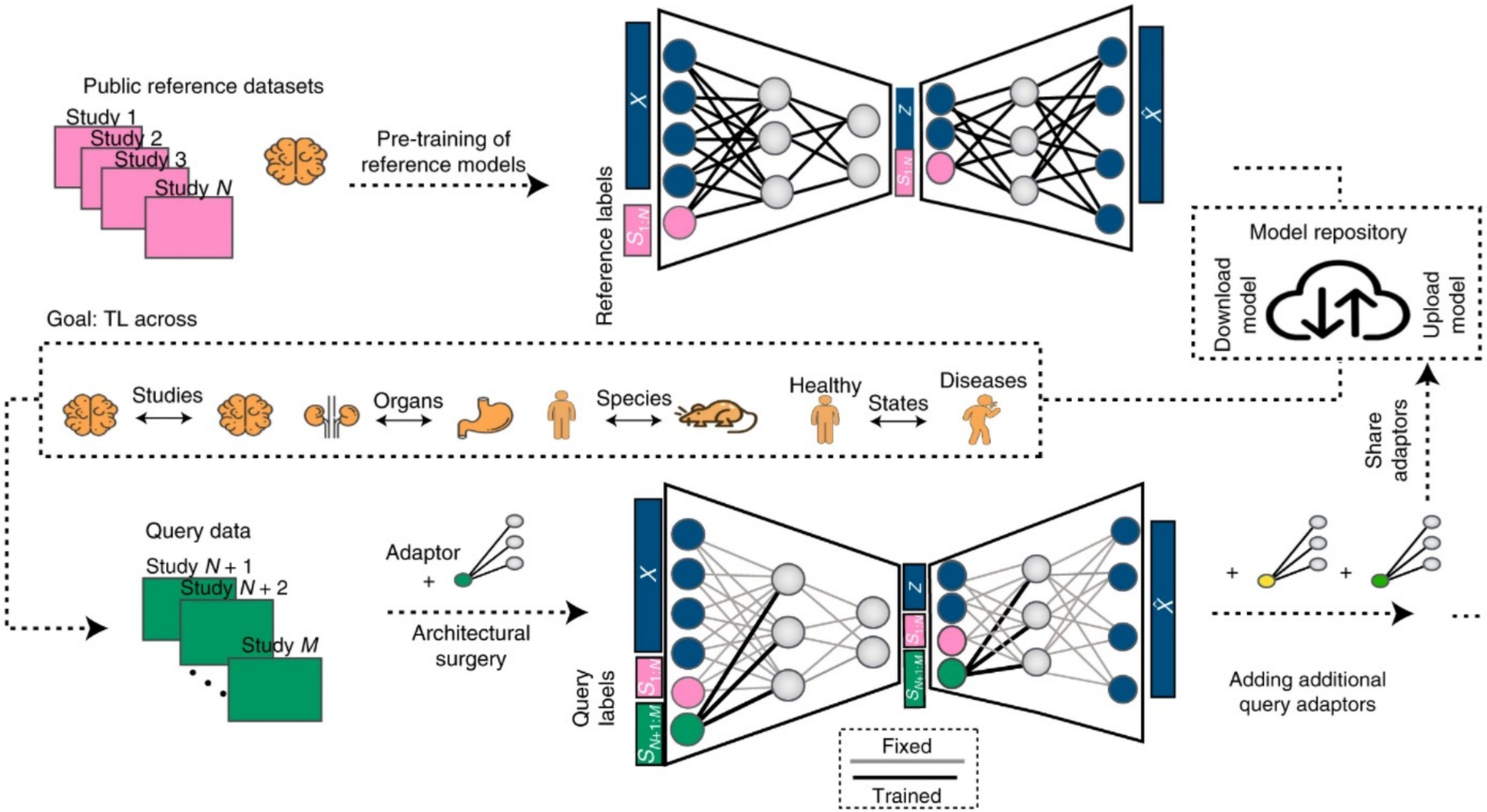
Motivation

- Existing approaches like the Seurat platform allow for integration of data but require that users run the complete pipeline on new datasets which requires excessive computational resources and time

- Unlike the existing tools, scArches uses a transfer learning approach to transfer the knowledge from pre-trained model to user specific data which not only reduces run time for integration but also leverages prior knowledge included in the preference dataset for the identification of cell types in the user-specific data.

scArches Overview

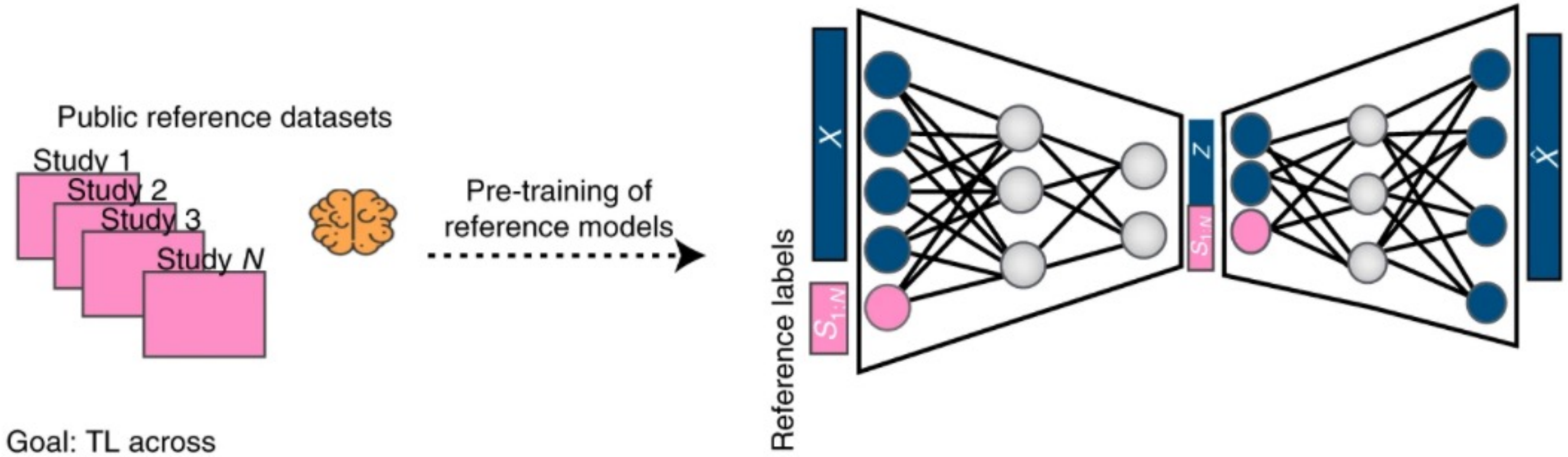
Deep generative model



Transfer learning, Architecture surgery

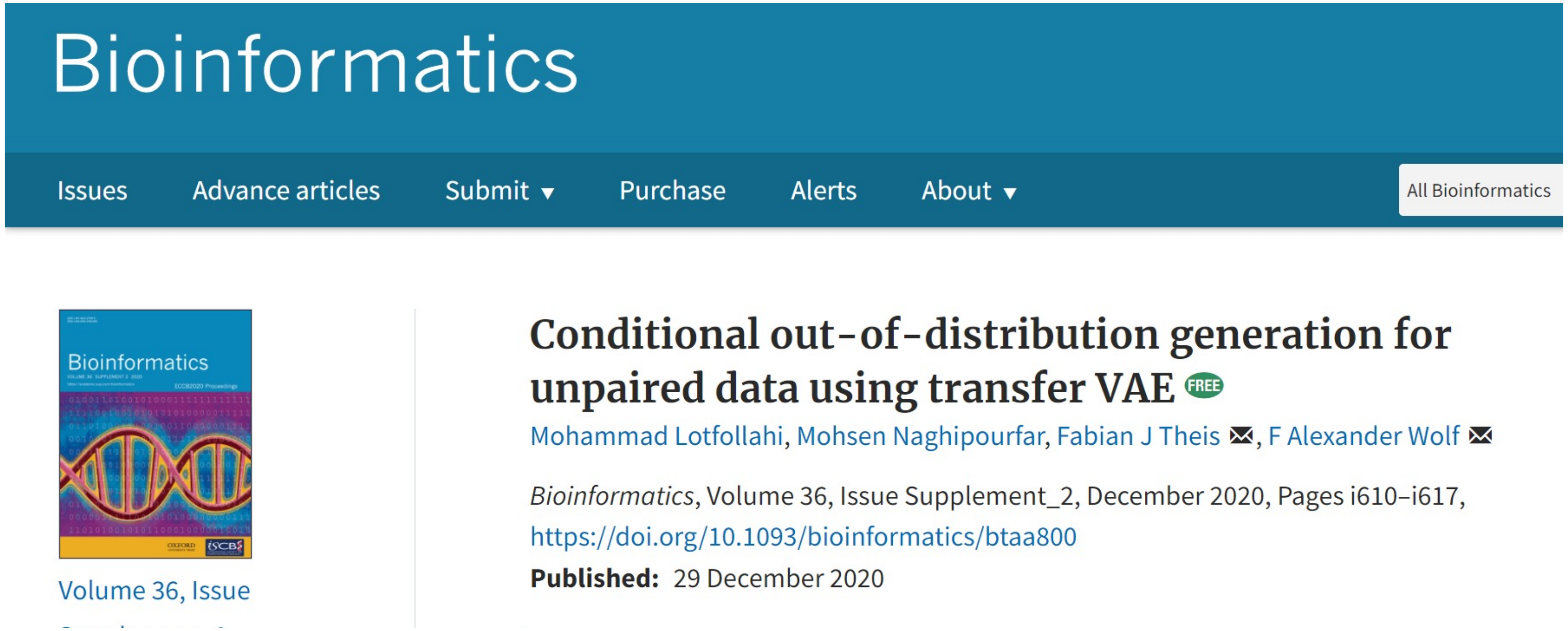
Deep generative model

- scArches integrates the reference data by deep generative model (DGM). It provided multiple variants of DGM including trVAE, scVI, scANVI, totalVI.



scArches trVAE

- trVAE builds upon conditional variational autoencoder (CVAE).



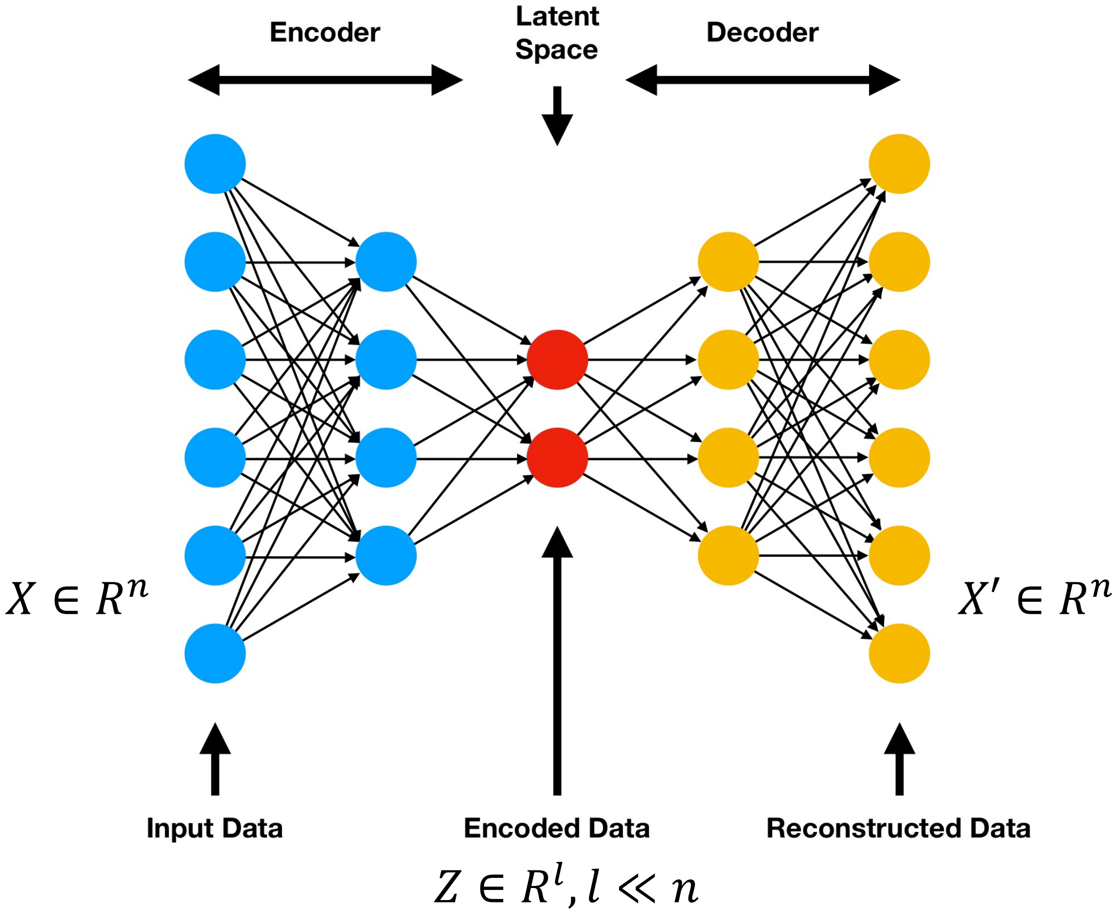
Conditional out-of-distribution generation for unpaired data using transfer VAE FREE

Mohammad Lotfollahi, Mohsen Naghipourfar, Fabian J Theis ✉, F Alexander Wolf ✉

Bioinformatics, Volume 36, Issue Supplement_2, December 2020, Pages i610–i617, https://doi.org/10.1093/bioinformatics/btaa800

Published: 29 December 2020

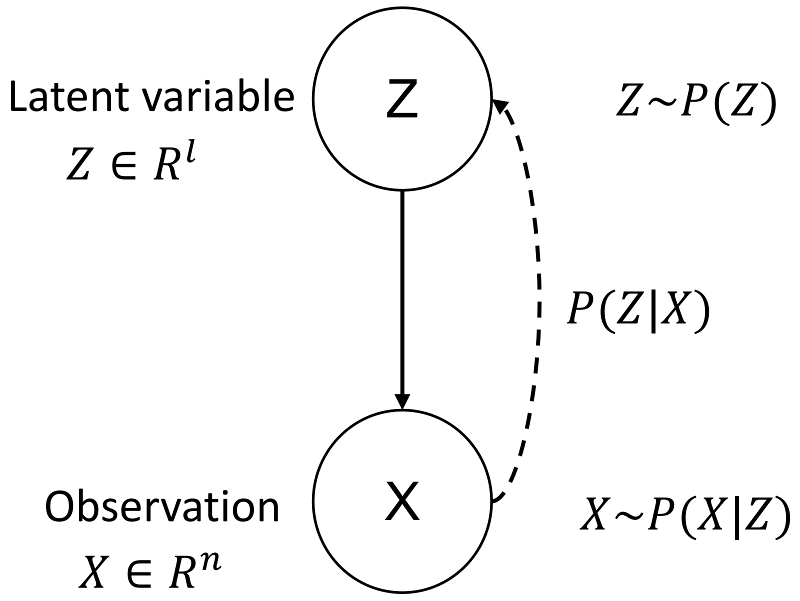
Autoencoder



Loss = $\frac{1}{M}\sum(X - X')^2$

- Dimension reduction methods.
- If the autoencoder has no hidden layer in Encoder and Decoder and no non-linearity activation on neurons, autoencoder is identity to PCA.
- Adding multiple hidden layers in Encoder and Decoder and the non-linearity activation on neuron will build a deep autoencoder.

Variational autoencoder



Latent variable

$Z \in R^l$

Z

$Z \sim P(Z)$

$P(Z|X)$

Observation

$X \in R^n$

X

$X \sim P(X|Z)$

- Variational inference

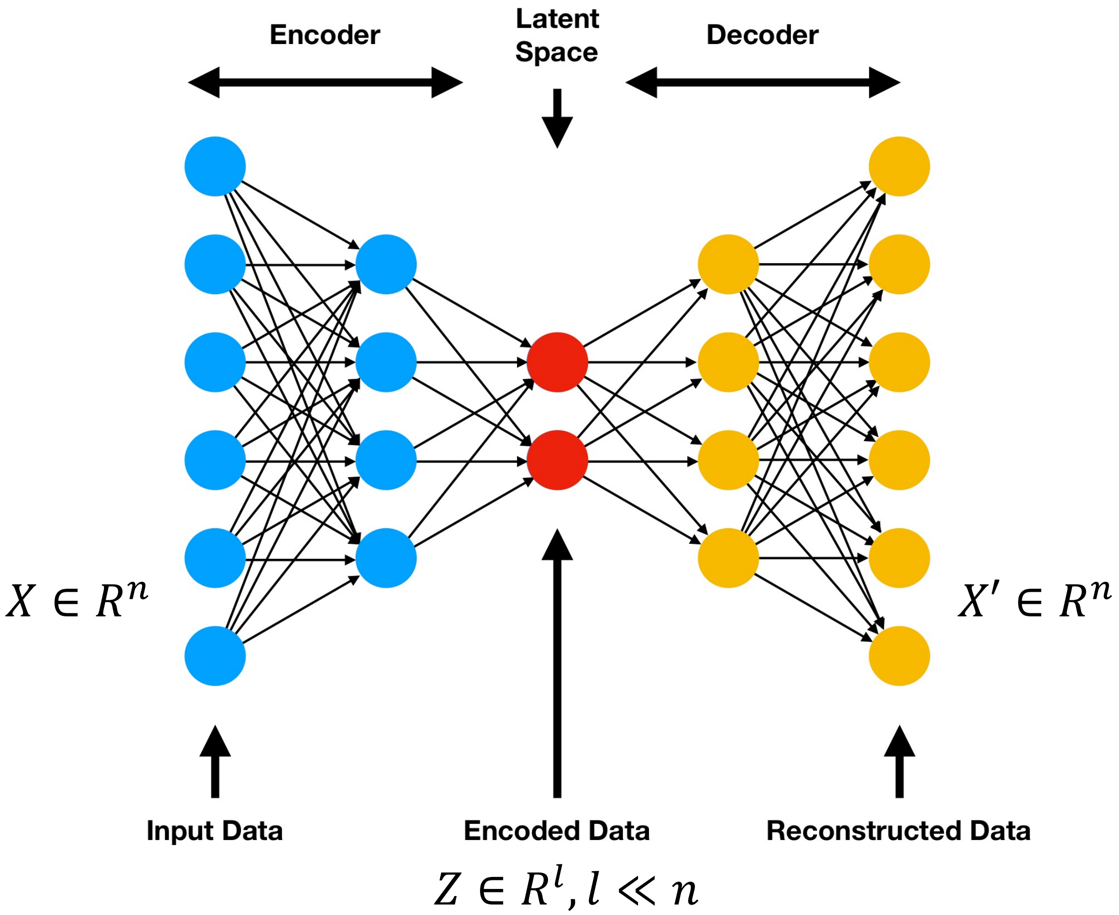
Find a variational distribution $Q(Z|X)$ which minimize,

$$D = KL(\ Q(Z|X)||P(Z|X)\ )$$

$$D = -E_{z\sim Q}[\log P(X|Z)] + KL(Q(Z|X)||P(Z)) + \log P(X)$$

Minimize $-E_{z\sim Q}[\log P(X|Z)] + KL(Q(Z|X)||P(Z))$

Variational autoencoder



Minimize $-E_{z \sim Q}[\log P\,(X|Z)] + KL(Q(Z|X)||P(Z))$

$Q(Z|X)$ is the Encoder; $P(X|Z)$ is the Decoder.

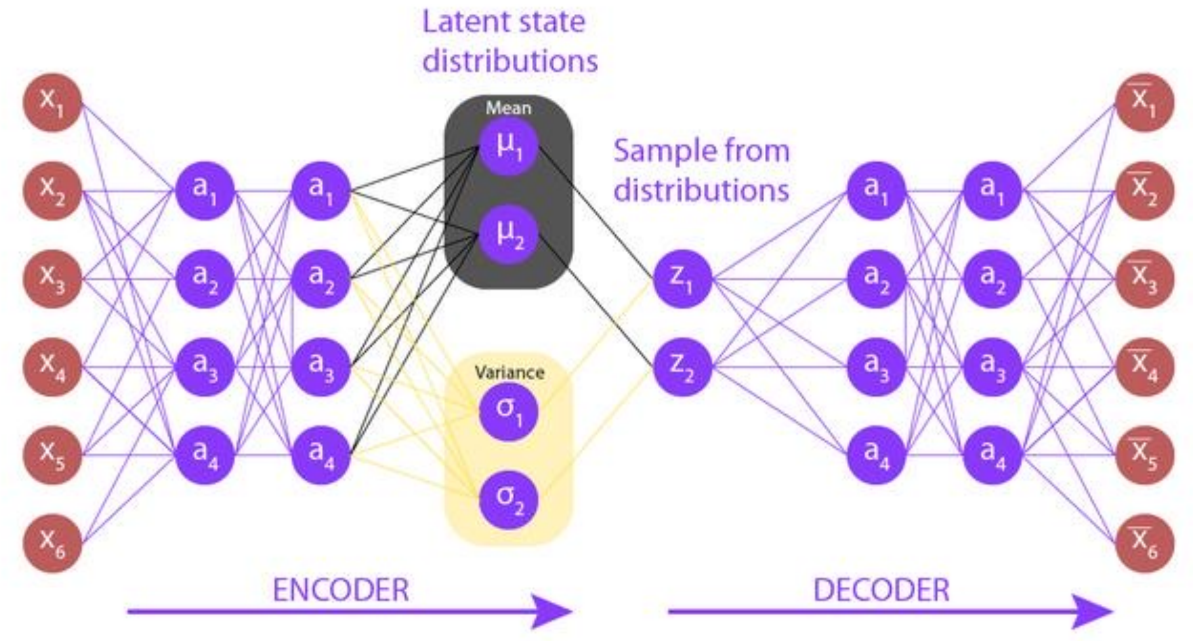
$KL(Q(Z|X)||P(Z))$ is the difference between latent space distribution $Q(Z|X)$ and prior distribution of Z $P(Z)$.

$P(X|Z)$ is equivalent to $P(X|X')$,

- If the X~ Normal, then $P(X|X')$ can be written as the form like $\exp(-|X - X'|^2)$. $\text{MSE} = \frac{1}{M}\sum(X - X')^2$.
- If the X~ Bernoulli, $P(X|X')$ is equivalent to cross entropy of X and X'.
- If we assume X follow other distributions like negative binomial distribution, the output layer will be the parameter of the NB distribution instead of reconstructed data. $-E_{z \sim Q}[\log P\,(X|Z)]$ is the negative loglikelihood of NB.

Variational autoencoder

Reparameterization


Latent state distributions
Mean
μ_1
μ_2
Variance
σ_1
σ_2
Sample from distributions
x_1 x_2 x_3 x_4 x_5 x_6
a_1 a_2 a_3 a_4
a_1 a_2 a_3 a_4
z_1 z_2
a_1 a_2 a_3 a_4
a_1 a_2 a_3 a_4
$\overline{x}_1$ $\overline{x}_2$ $\overline{x}_3$ $\overline{x}_4$ $\overline{x}_5$ $\overline{x}_6$
ENCODER
DECODER

Minimize $-E_{Z\sim Q}[\log P\,(X|Z)] + KL(Q(Z|X)||P(Z))$

$Q(Z|X)$ is the Encoder; $P(X|Z)$ is the Decoder.

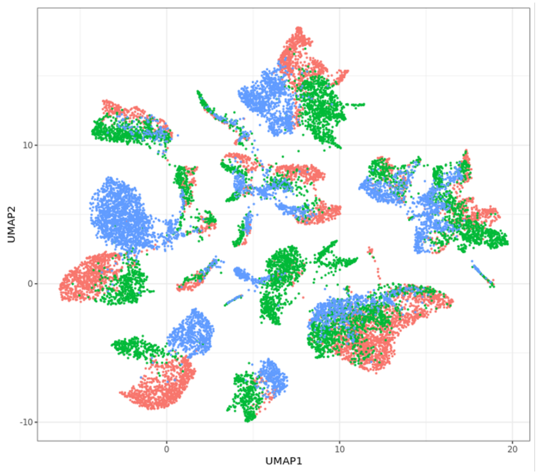
$KL(Q(Z|X)||P(Z))$ is the difference between latent space distribution $Q(Z|X)$ and prior distribution of Z $P(Z)$.

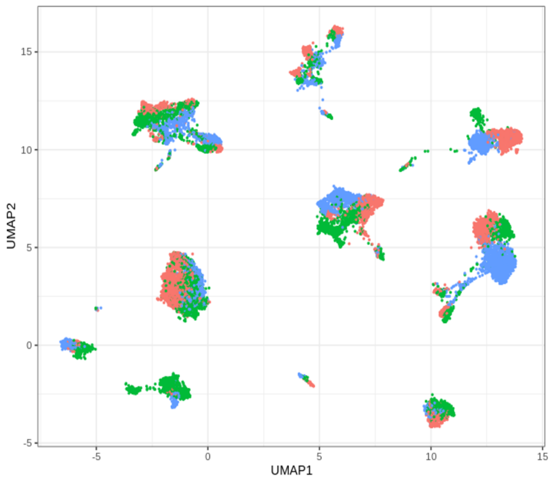
$P(X|Z)$ is equivalent to $P(X|X')$,

- If the X~ Normal, then $P(X|X')$ can be written as the form like $\exp(-|X-X'|^2)$. $\text{MSE} = \frac{1}{M}\sum(X-X')^2$.
- If the X~ Bernoulli, $P(X|X')$ is equivalent to cross entropy of X and X'.
- If we assume X follow other distributions like negative binomial distribution, the output layer will be the parameter of the NB distribution instead of reconstructed data. $-E_{Z\sim Q}[\log P\,(X|Z)]$ is the negative loglikelihood of NB.

VAE on healthy lung data

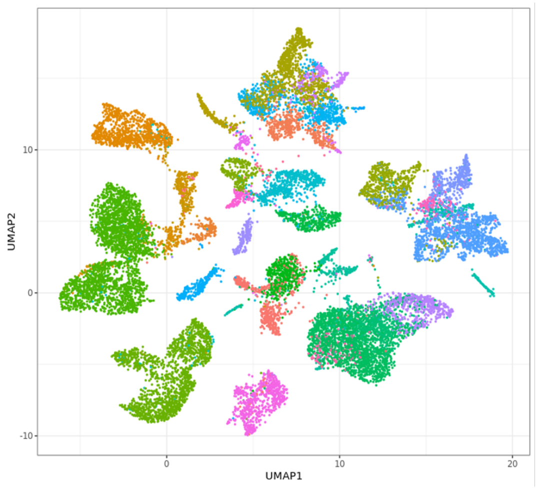
UMAP on PCA
Color: Batches

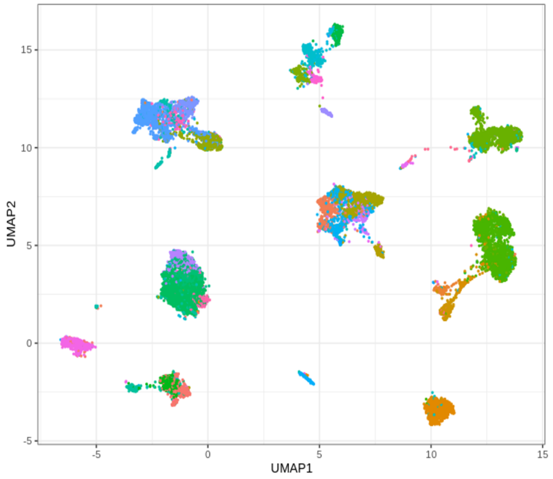




UMAP on VAE
Color: Batches

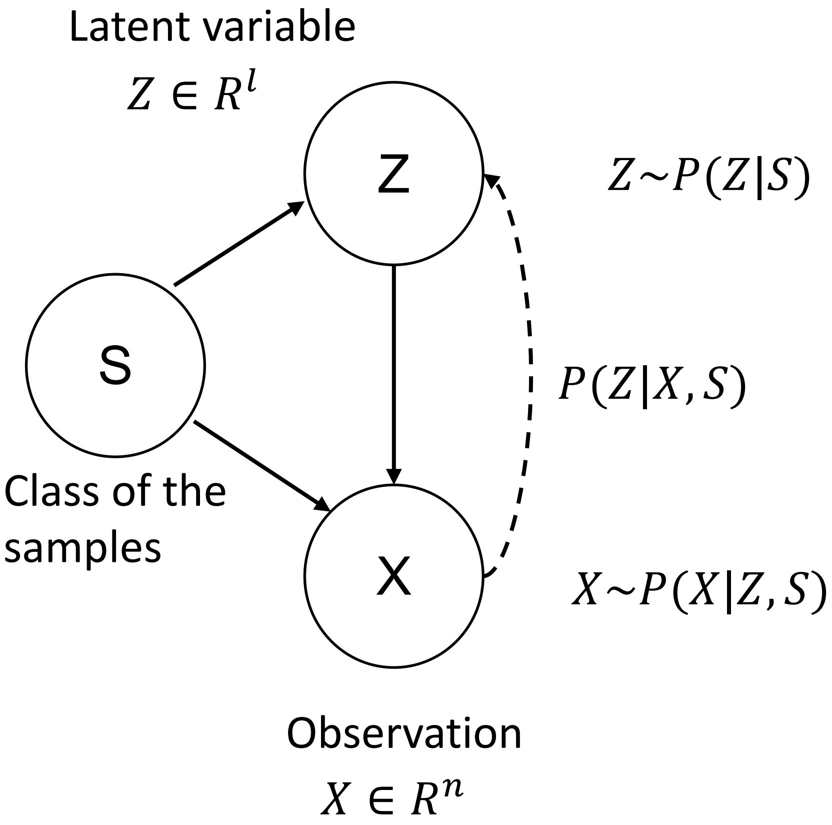
UMAP on PCA
Color: Cell types





UMAP on VAE
Color: Cell types

Conditional variational autoencoder(CVAE)



Latent variable

$Z \in R^l$

$Z \sim P(Z|S)$

$P(Z|X,S)$

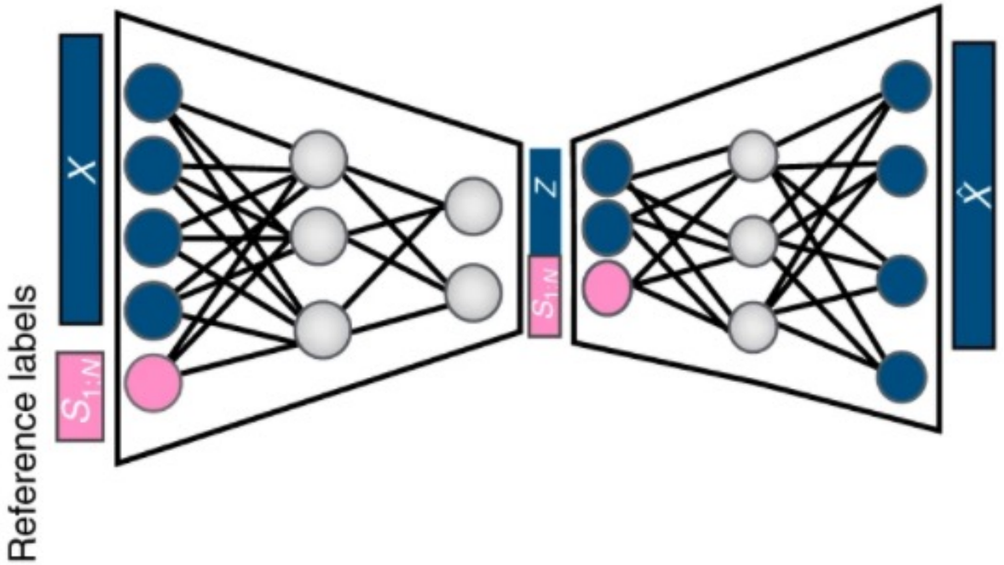
$X \sim P(X|Z,S)$

Class of the samples

Observation

$X \in R^n$

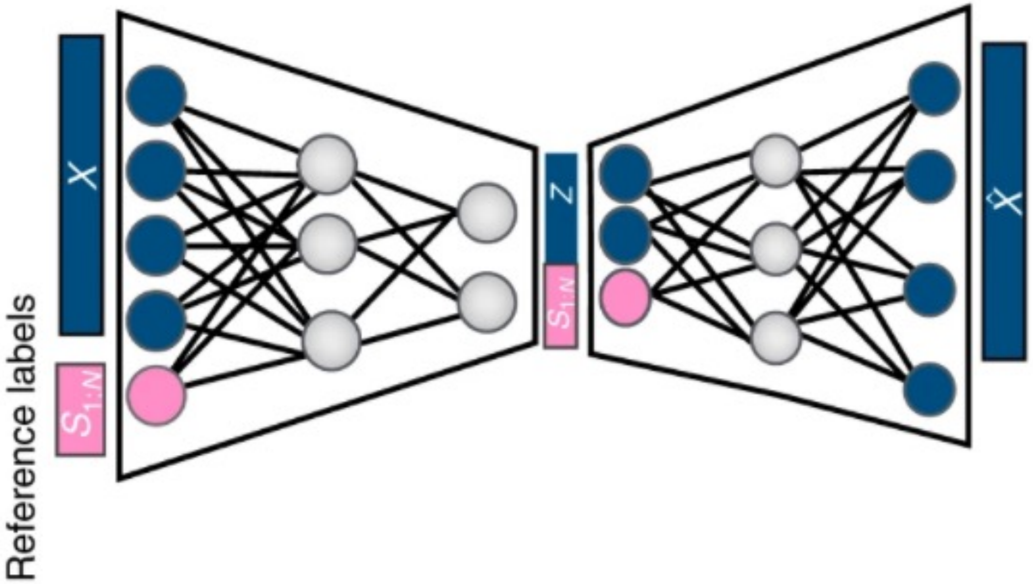
Minimize $-E_{z\sim Q}[\log P\,(X|Z,S)] + KL(Q(Z|X,S)||P(Z|S))$



trVAE

Loss of conditional variational autoencoder

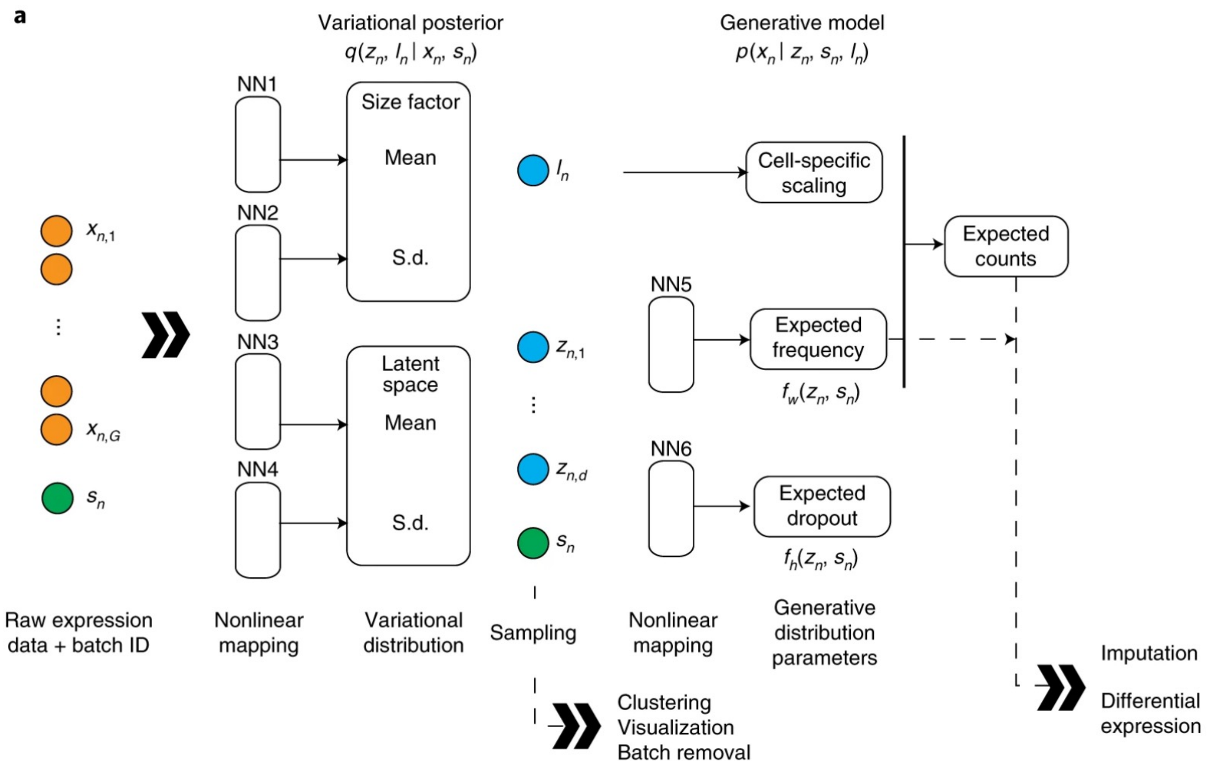
Minimize $-E_{z\sim Q}[\log P\,(X|Z,S)] + KL(Q(Z|X,S)||P(Z|S))$ $-MMD(y_i, y_j)$

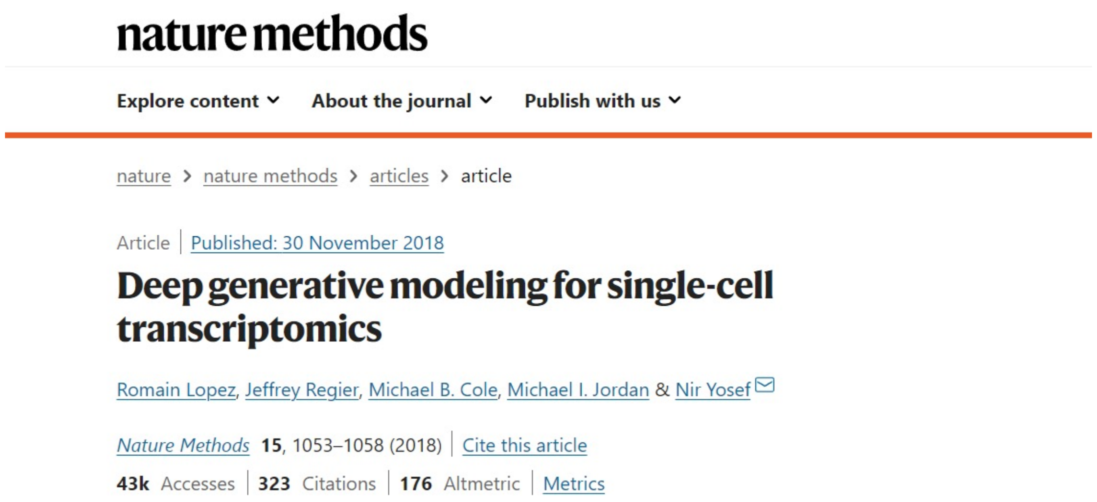


MMD is the Maximum Mean Discrepancy. It is the similarity of two distributions.

y is the values of the first hidden layer of Decoder. i and j are the labels of samples.

scVI





nature methods

Explore content ⌄ About the journal ⌄ Publish with us ⌄

nature › nature methods › articles › article

Article | Published: 30 November 2018

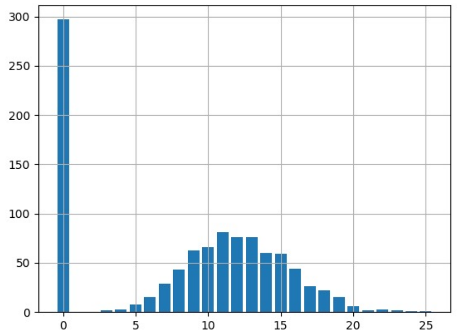
Deep generative modeling for single-cell transcriptomics

Romain Lopez, Jeffrey Regier, Michael B. Cole, Michael I. Jordan & Nir Yosef

Nature Methods **15**, 1053–1058 (2018) | Cite this article

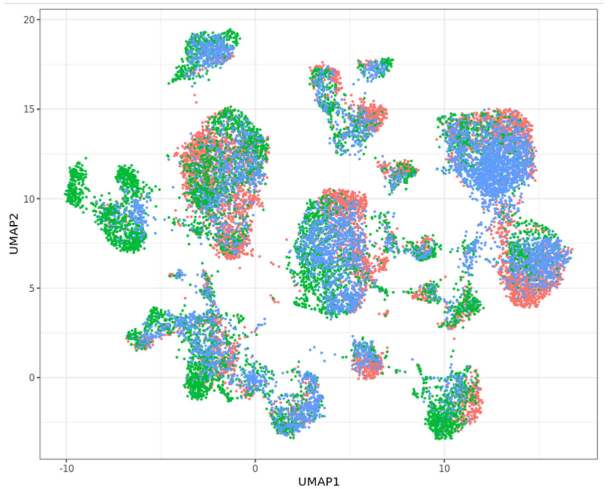
43k Accesses | **323** Citations | **176** Altmetric | Metrics

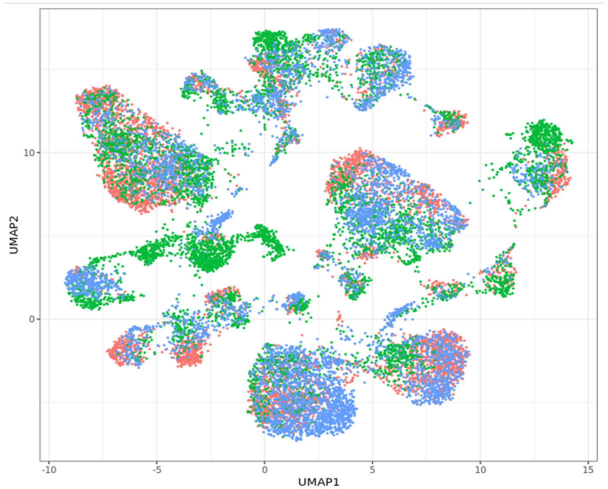
Zero inflated negative binomial distribution



Results

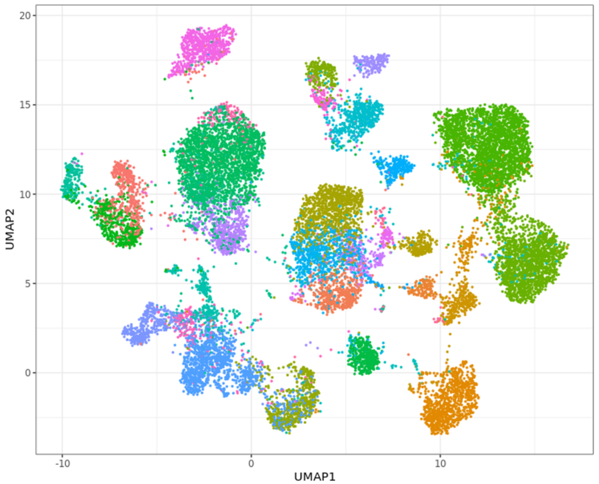
UMAP on trVAE
Color: Batches

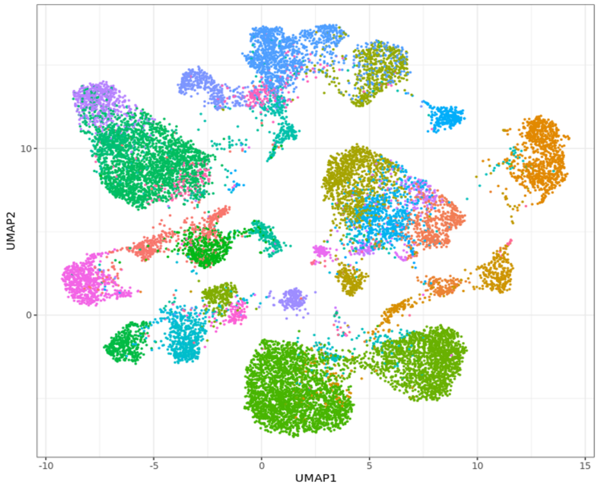




UMAP on scVI
Color: Batches

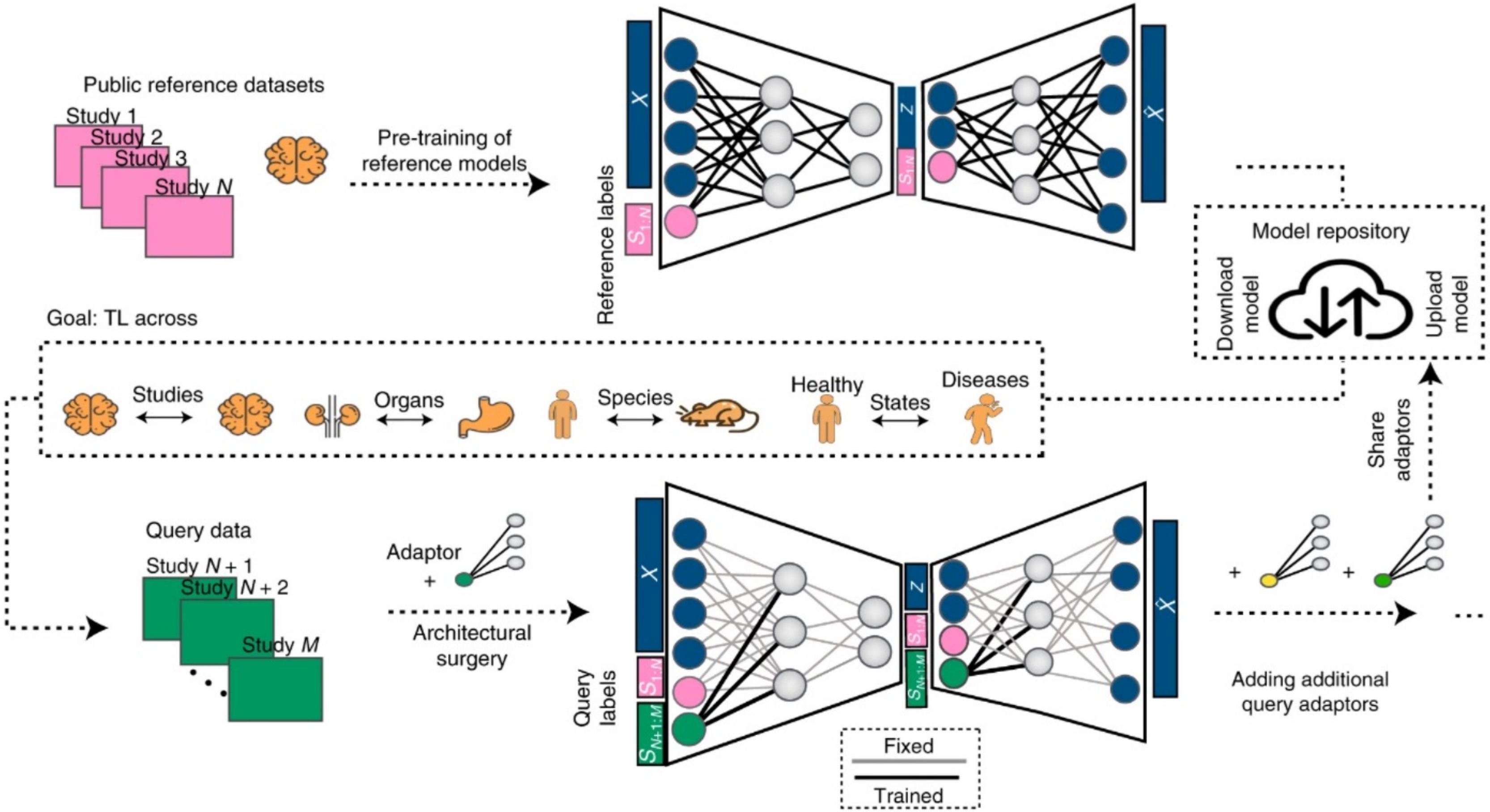
UMAP on trVAE
Color: Cell types





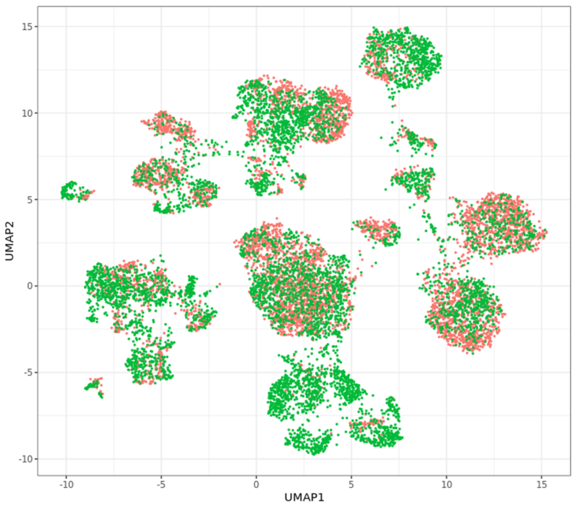
UMAP on scVI
Color: Cell types

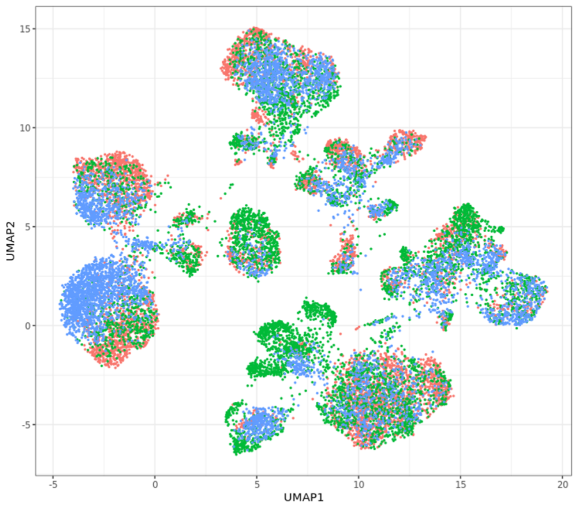
Transfer learning



Results: transfer learning of trVAE

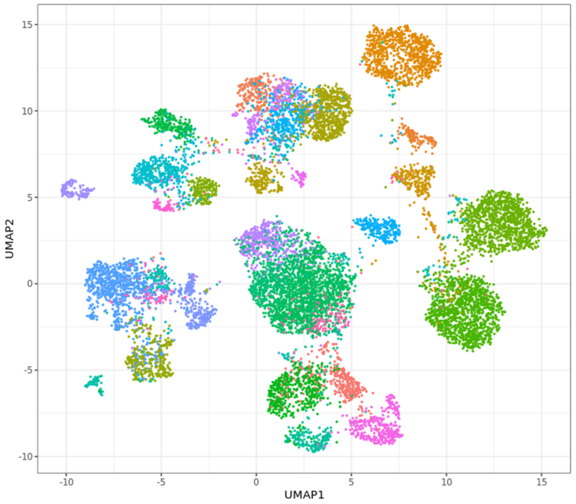
trVAE on ref data
Color: Batches

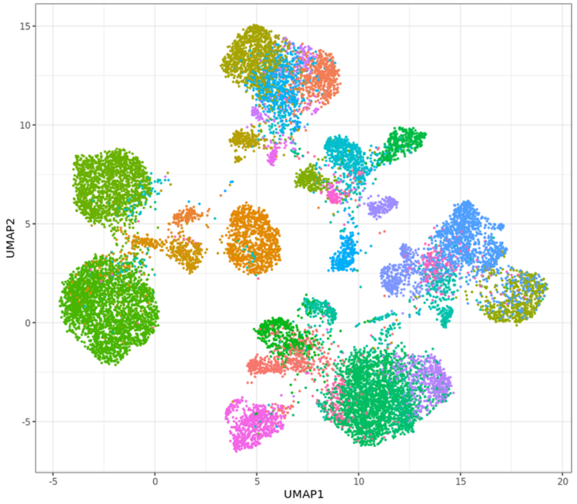




Mapping of query data
Color: Batches

trVAE on ref data
Color: Cell types

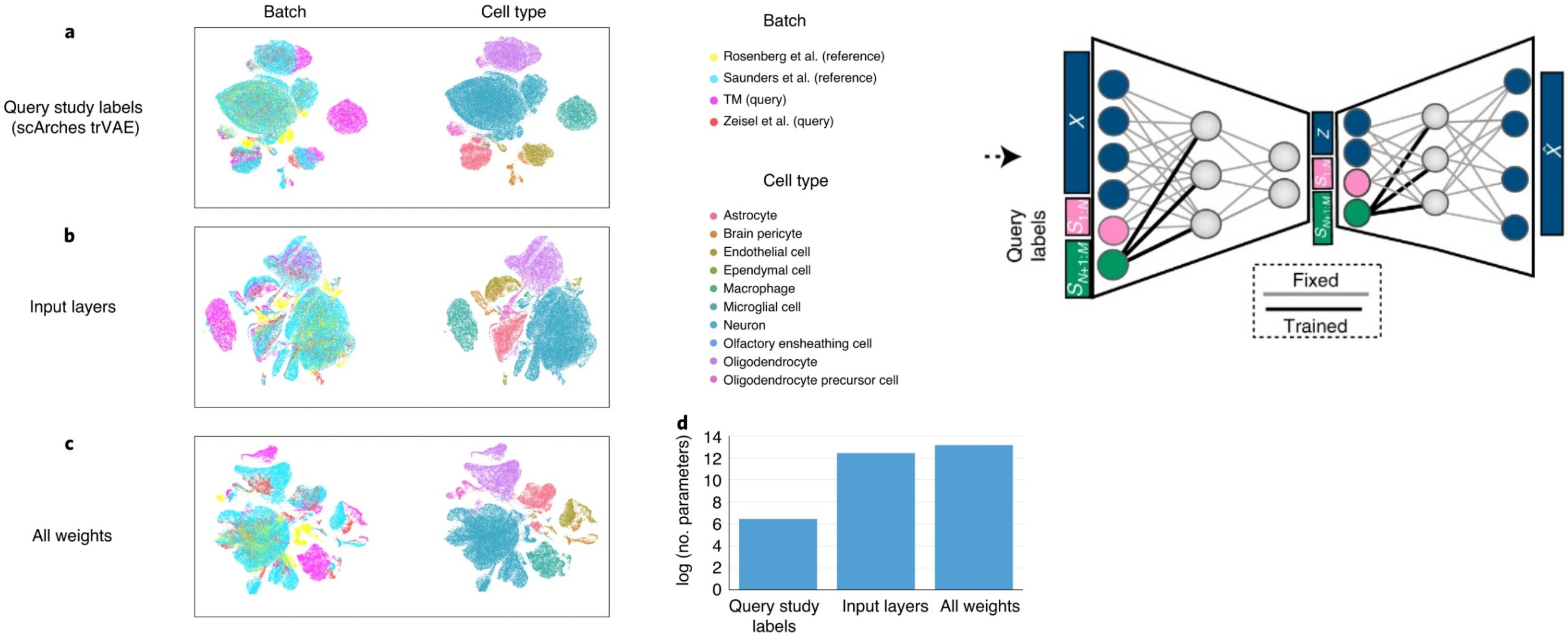




Mapping of query data
Color: Cell types

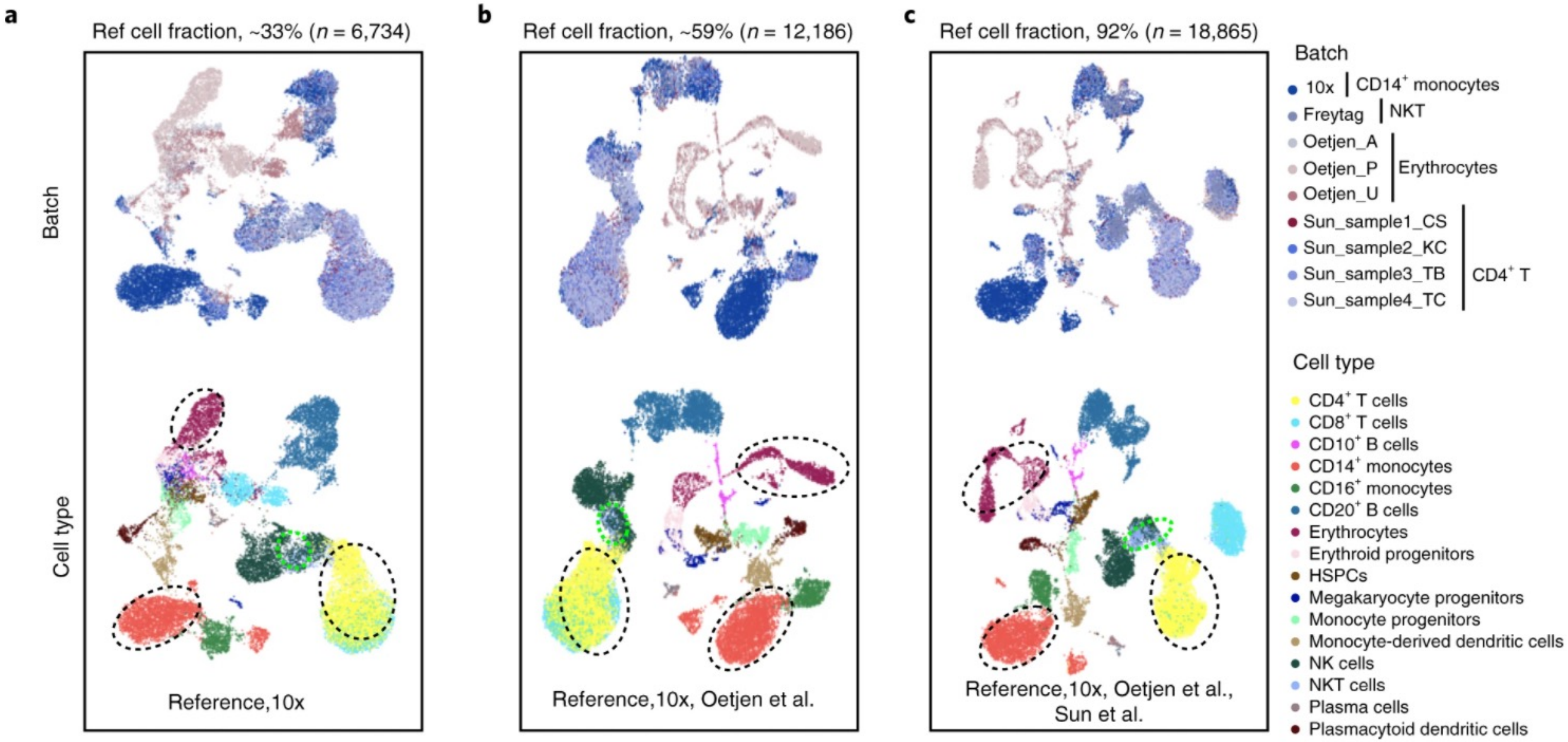
Results

TL and architecture surgery allow fast and accurate reference mapping.



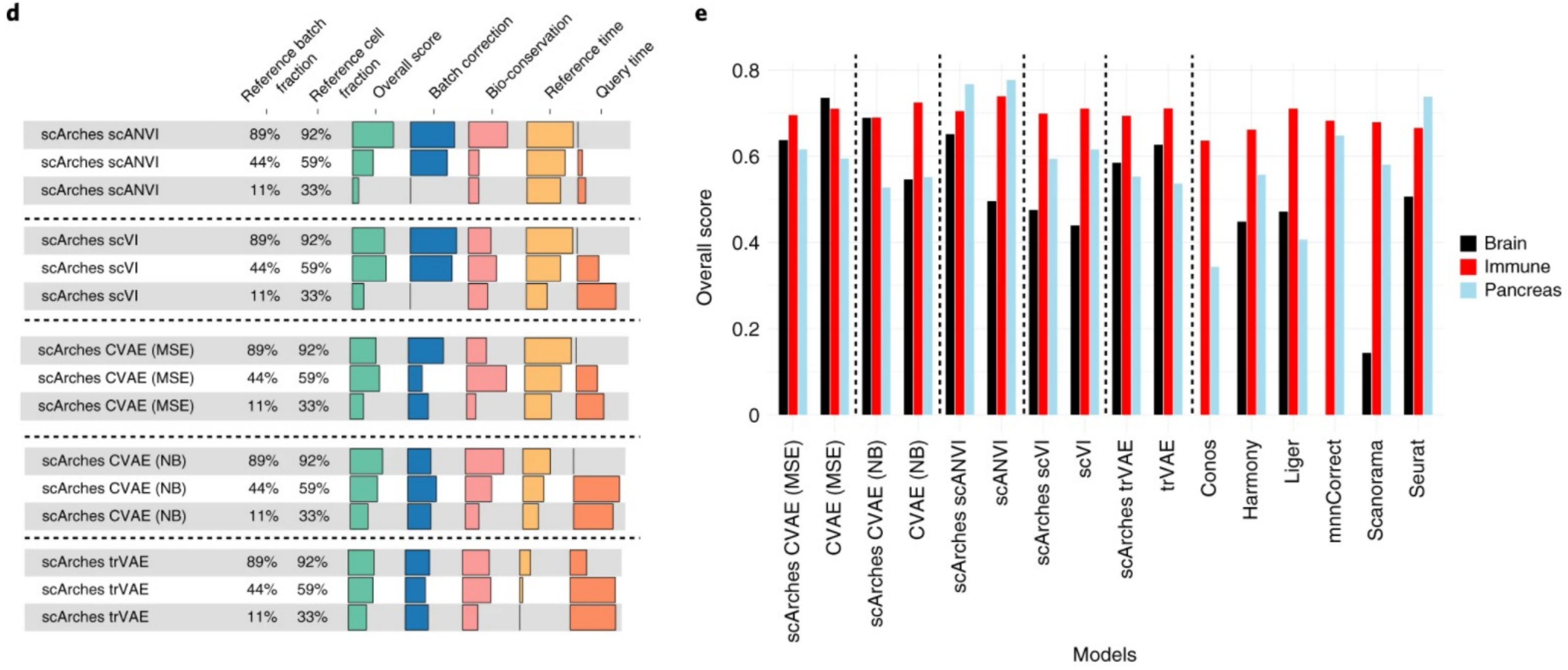
Results

scArches remains robust even when cell fractions in the reference data are varied



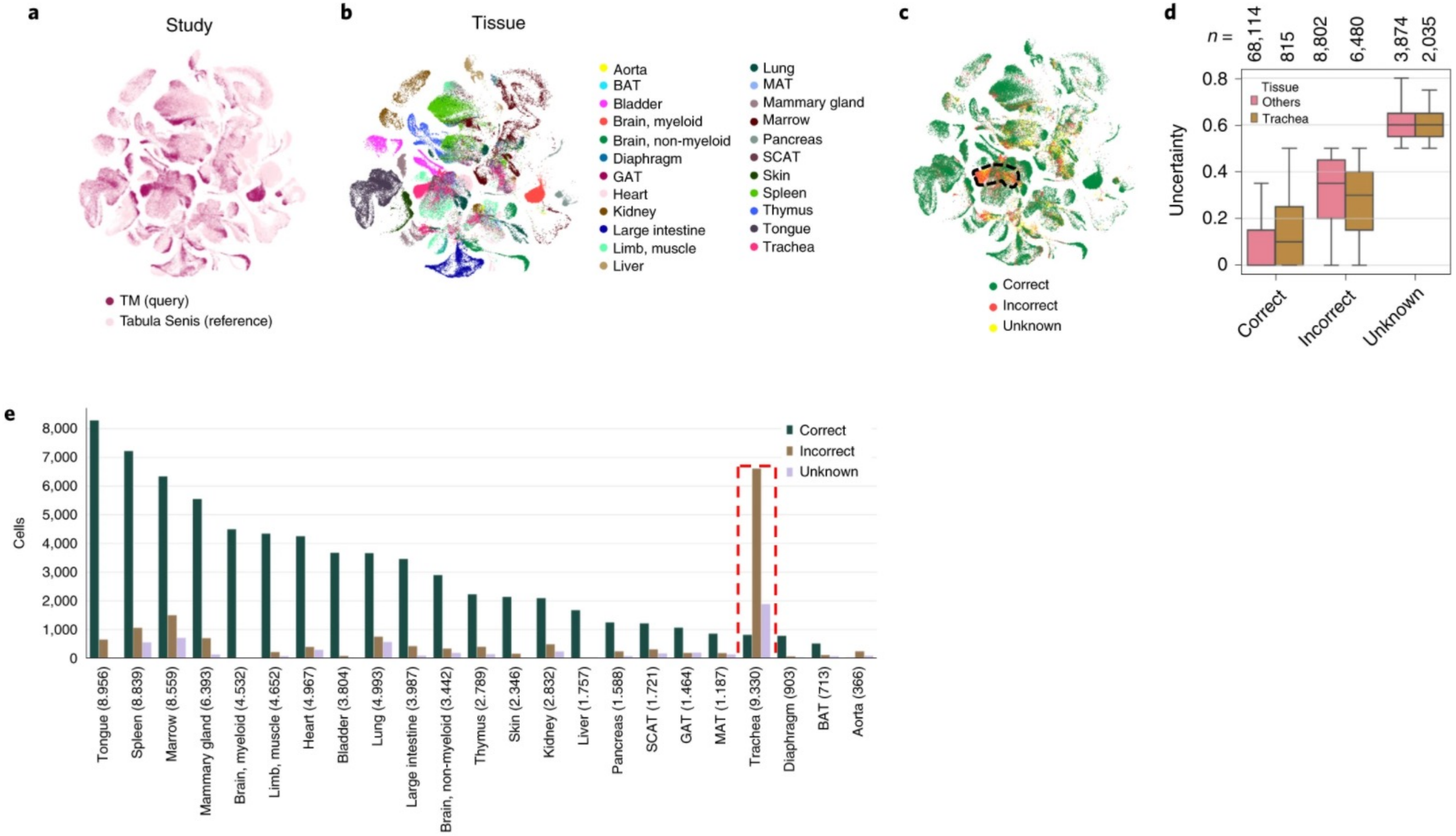
Results

scArches enables efficient reference mapping compared existing data-integration methods.



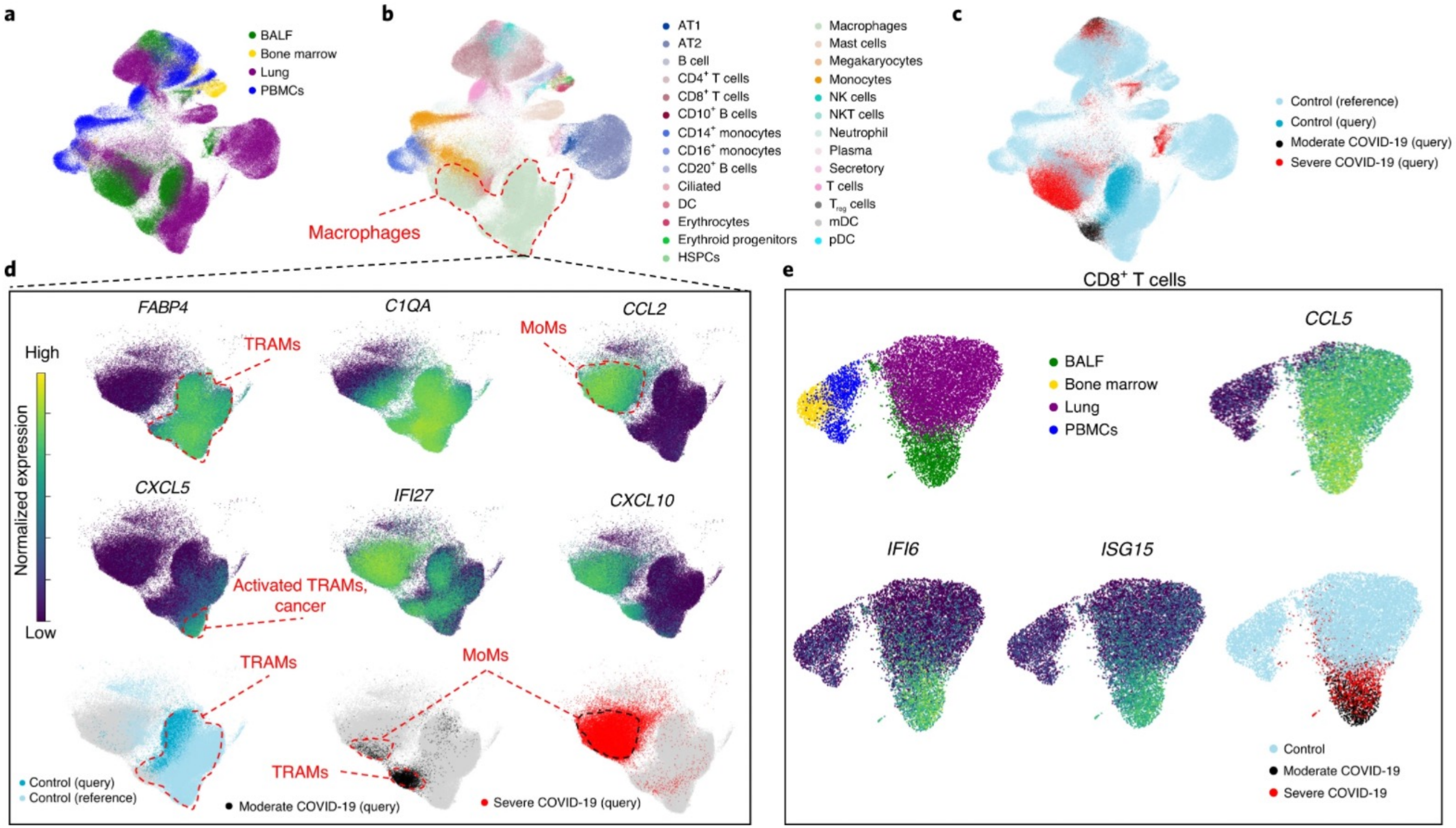
Results

scArches discovers cell types in query data, even when they such cell types are removed from reference data



Results

scArches resolves severity in COVID-19 query data mapped to a healthy reference and reveals emergent cell states.



Summary

- Deep generative model

“Things may shift their forms ten thousand times, but the principle remain unchanged.” *Xunzi, third century BC*

The goal of these different DGMs are trying to learn a low dimensional space of data by adding constraints on the model and modifying the learning target of the model.

Single cell studies are trying to understand the similarities between cells. By applying DGMs on single cell data, we can define the cell-cell distance on this new space. This will provide us information for clustering and trajectory building analysis.

- Transfer learning

Large number of single cell data are generated. But the analysis tools which making use of these big data is a few.

Unanswered Questions and new directions

- Learn biological meaningful latent variables
- distinguish the technical batch effect from biological effect
- Integrate more prior biological knowledge into the deep generative model

Acknowledgements

Rehman lab:

Xinge Wang
Mark Sanborn
Priyanka Gajwani
Arundhati Jana
Young-Mee Kim
Li Wang
Joseph Leasure
Shubhi Srivastava
Zijing Ye
Lissette Magana

Dai lab:

Yongchao Huang
Mehrdad Zandigohar

Mentors:

Yang Dai
Jalees Rehman