

# STAT 35510

## Lecture 9

Spring, 2024  
Jingshu Wang

# Outline

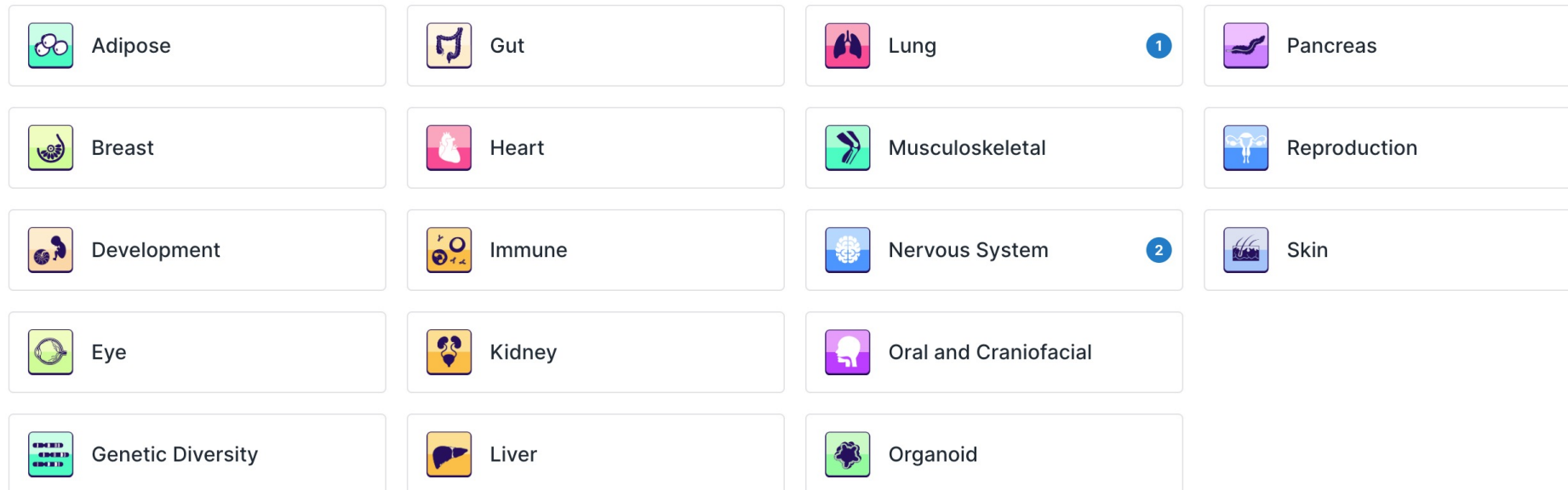
- Reference mapping and automatic cell type label transfer (annotation)
  - Collection of large-scale atlas data
  - Autoencoder-based methods
  - Cell-cell similarity based methods
  - More complicated models using transformer

# External data: Human Cell Atlas (HCA)

- Global collaboration to map all cells in a human body
- The HCA community collect multi-omics single-cell sequencing data
- Data publicly available for download

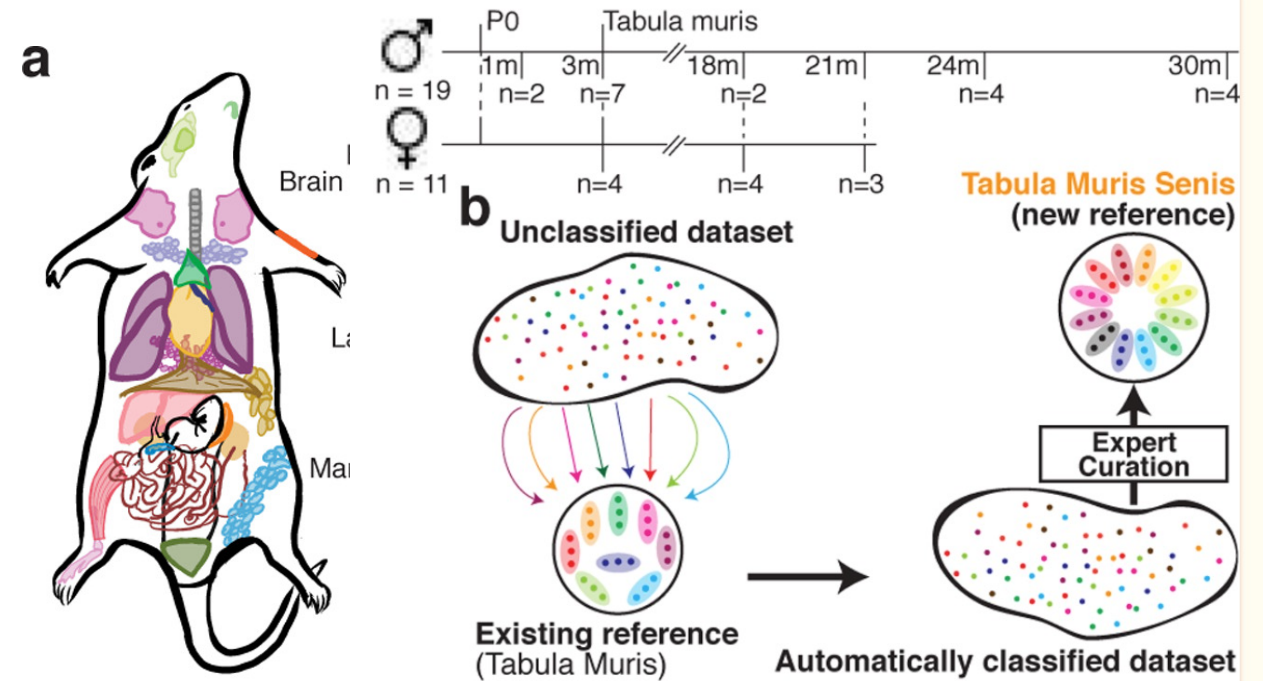


## HCA Biological Network Atlases



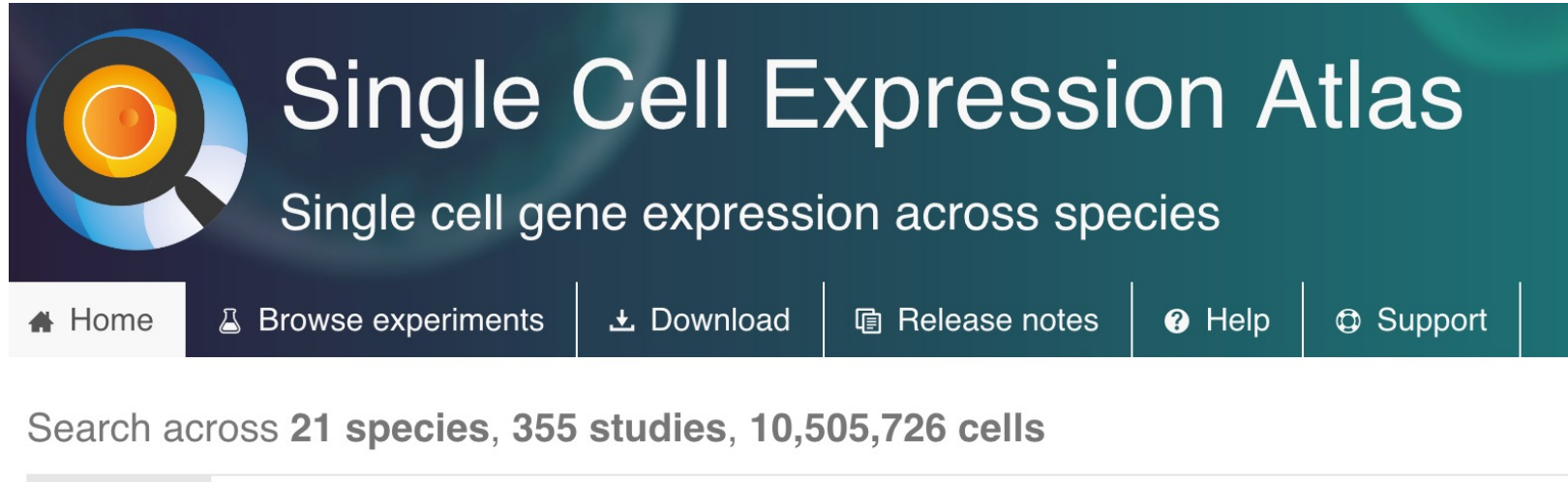
# External data for mouse

- Mouse Cell Atlas (Han et. al., Cell 2018):  
~ 500,000 cells, 40 tissues
- Data from Tabula Muris Consortium:  
multi-tissue atlas transcriptomics data  
along mouse lifespan to understand aging
  - (The Tabula Muris Consortium Nature 2018): 100K cells, 20 organs and tissues
  - (The Tabula Muris Consortium Nature 2020): 350K cells, 6 age groups (1 month – 30 months), 23 tissues and organs
- Various large-scale data for different mouse tissues (such as the brain)



# Many other atlas-scale data

- scRNA-seq atlas data across species including animals, plants and fungi



The image shows the header of the Single Cell Expression Atlas website. It features a dark teal background with a circular logo on the left containing a stylized orange and yellow cell. The main title "Single Cell Expression Atlas" is in large white font, with the subtitle "Single cell gene expression across species" below it. A navigation bar contains links for Home, Browse experiments, Download, Release notes, Help, and Support. Below the navigation bar, a search bar is visible with the text "Search across 21 species, 355 studies, 10,505,726 cells".

**Single Cell Expression Atlas**  
Single cell gene expression across species

Home | Browse experiments | Download | Release notes | Help | Support

Search across **21 species, 355 studies, 10,505,726 cells**

- Human protein atlas
  - Protein coding genes form 31 human tissues

# What can large-scale atlas data offer?

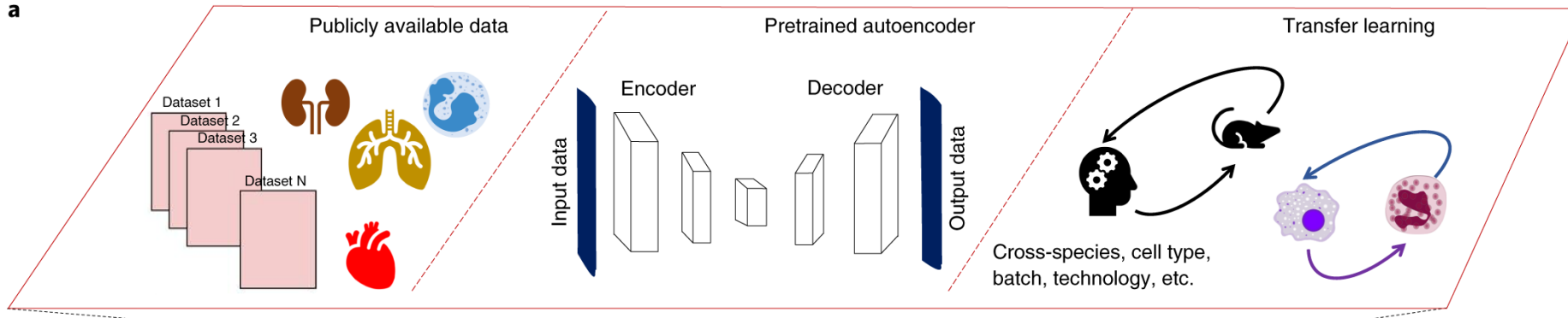
- Large number of cells characterizing the expression patterns of genes in various cell populations
- Expert curated annotations of the cells
  - Aiming to provide information on every cell type
- Understand gene expression and cell population variability across individuals / patients
- Data on mouse cells may provide a better understanding of human cells

## Goals:

- Create a reference atlas map that have corrected batch effects across individual datasets within the atlas data
- Reference mapping: transfer learning for analyzing new target data (small sample size, collected under a new condition)
  - Better visualization and clustering, especially for the rare cell types
  - Denoising of the target data
  - Automatic cell type annotation
- Comparison between the new target data and the reference
  - New cell type
  - Differentially expressed genes between target and reference within the same cell type

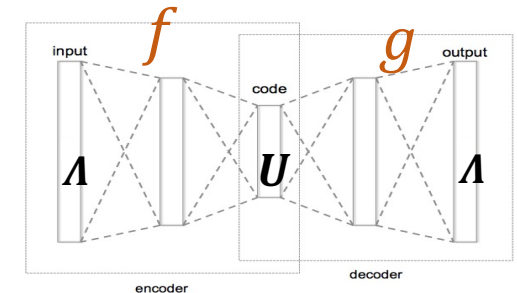
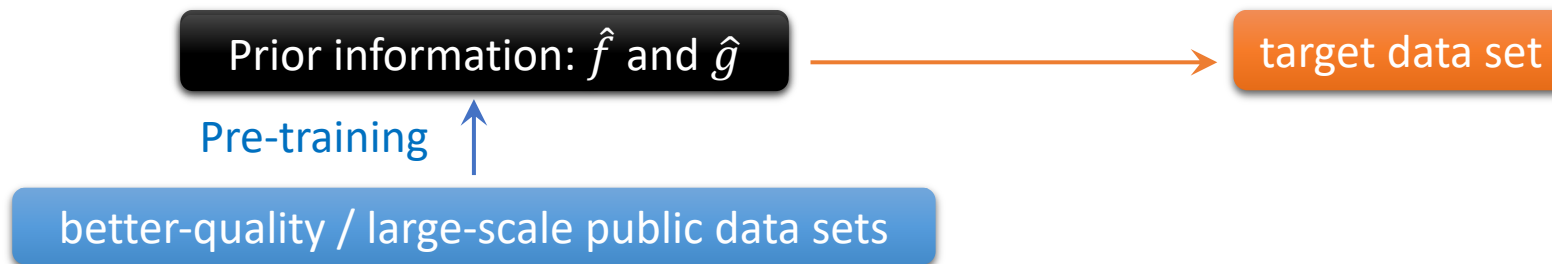
# SAVER-X (Wang et. al. Nature Methods 2019)

- SAVER-X: transfer learning from reference data to help denoising

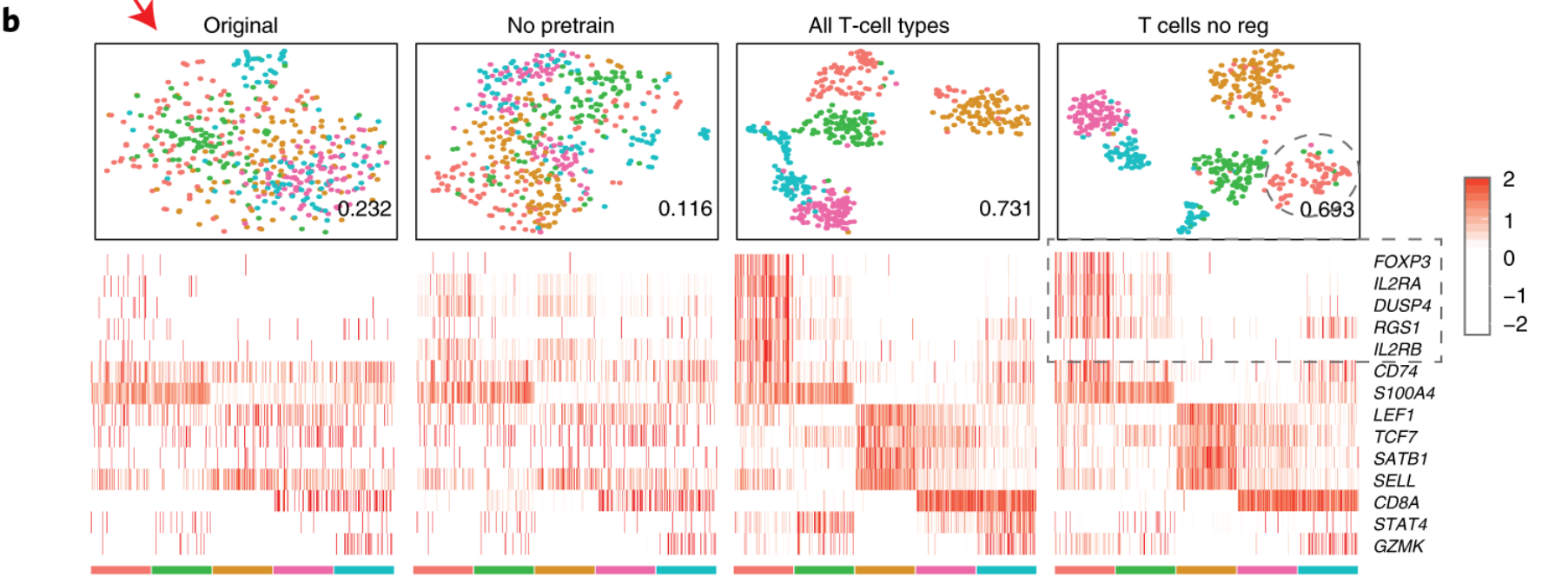
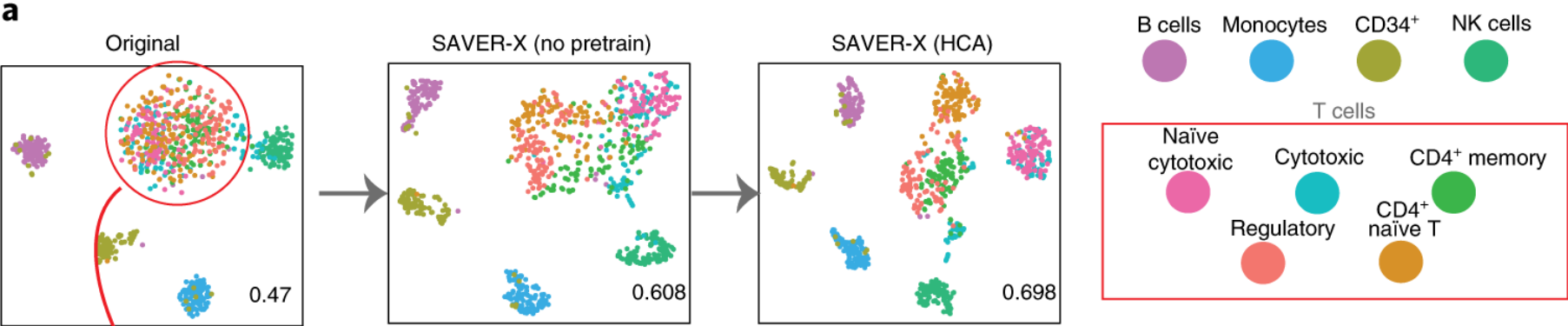


- Main idea: use reference data as better initialization autoencoder**
  - No adjustment of batch effect
    - Reference data should have similar tissue / cell types
    - Only focus on the target data (no comparison between reference and target)

Weight Initialization using  $\hat{f}$  and  $\hat{g}$



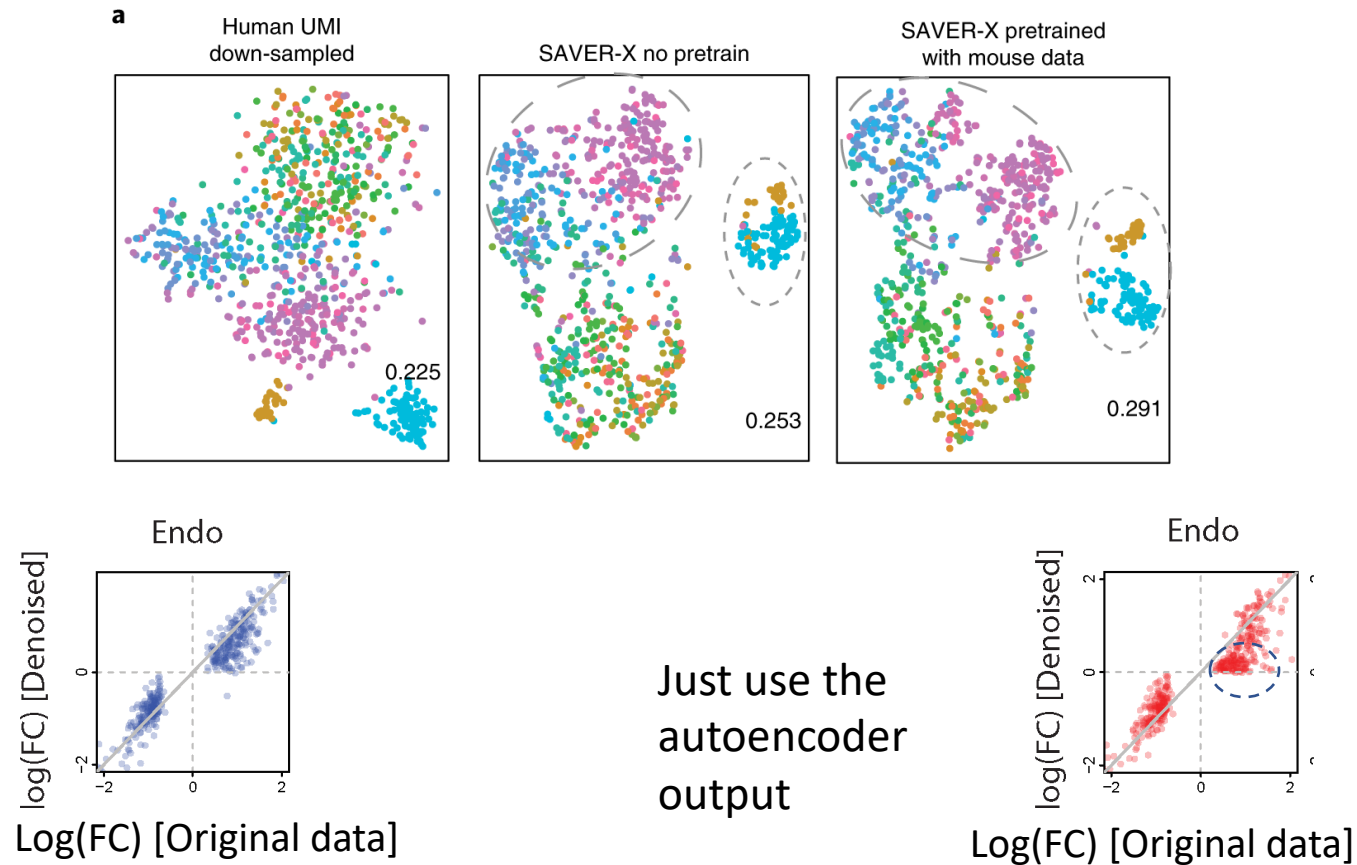
# SAVER-X (Wang et. al. Nature Methods 2019)





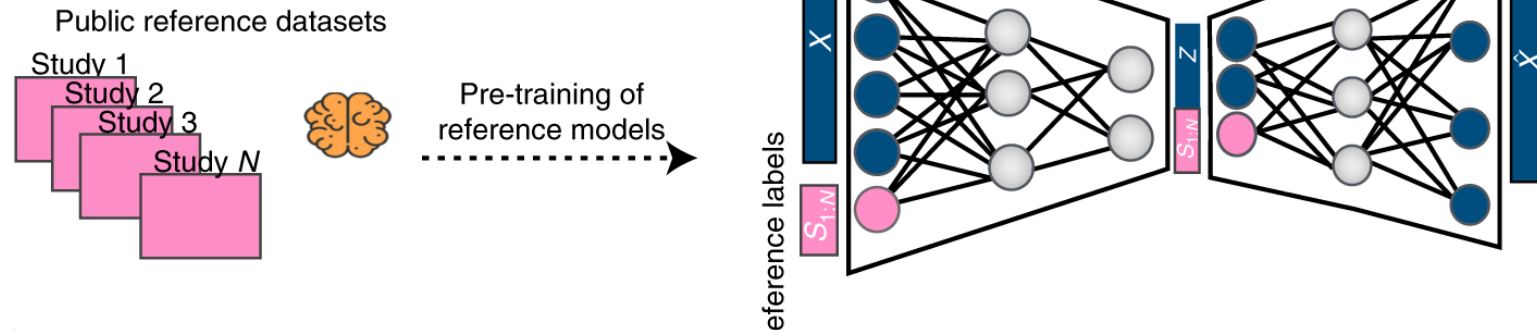
# SAVER-X (Wang et. al. Nature Methods 2019)

- Bayesian model makes final denoised value a weighted average between autoencoder output and observed data
  - Help removing biased from reference data
- Example: mouse to human transfer



# scArches (Lotfollahi et. al., Nature Biotech 2022)

- Uses a similar VAE framework but adjust for batch effects
- Focus on low-dimensional representation of the cells
  - Can also obtain “reference-corrected” gene expression matrix
- Main idea
  - Pretrain reference data using a similar framework as scVI
    - Add reference labels (such as batches, datasets, conditions, tissues, species ...) both in input layer and bottleneck layer
    - Can pre-train the reference model with other deep learning framework like scANVI
    - Can also add an extra MMD penalty in the loss function to further encourage that data from different batches are mixed in  $Z$  [reduce correlation between  $Z$  and batches]

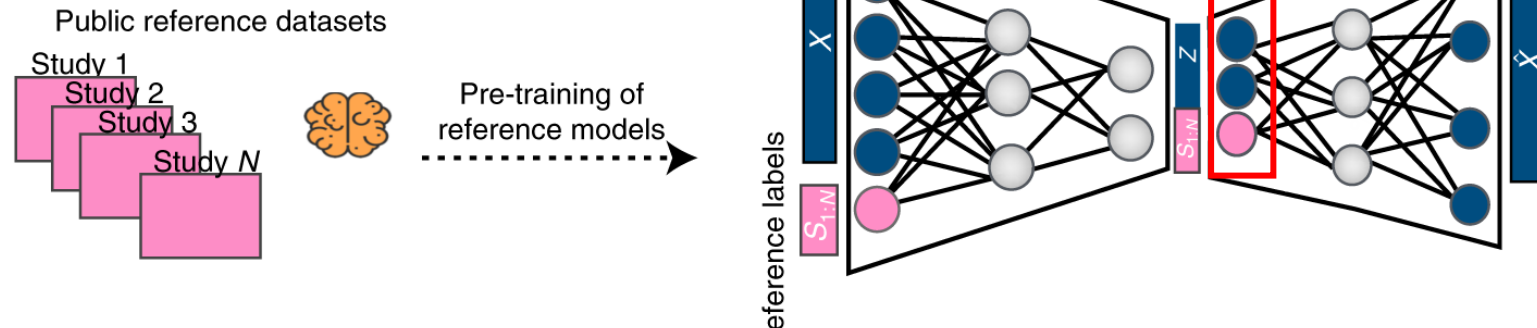


# scArches (Lotfollahi et. al., Nature Biotech 2022)

- MMD penalty between two datasets  $X$  and  $X'$

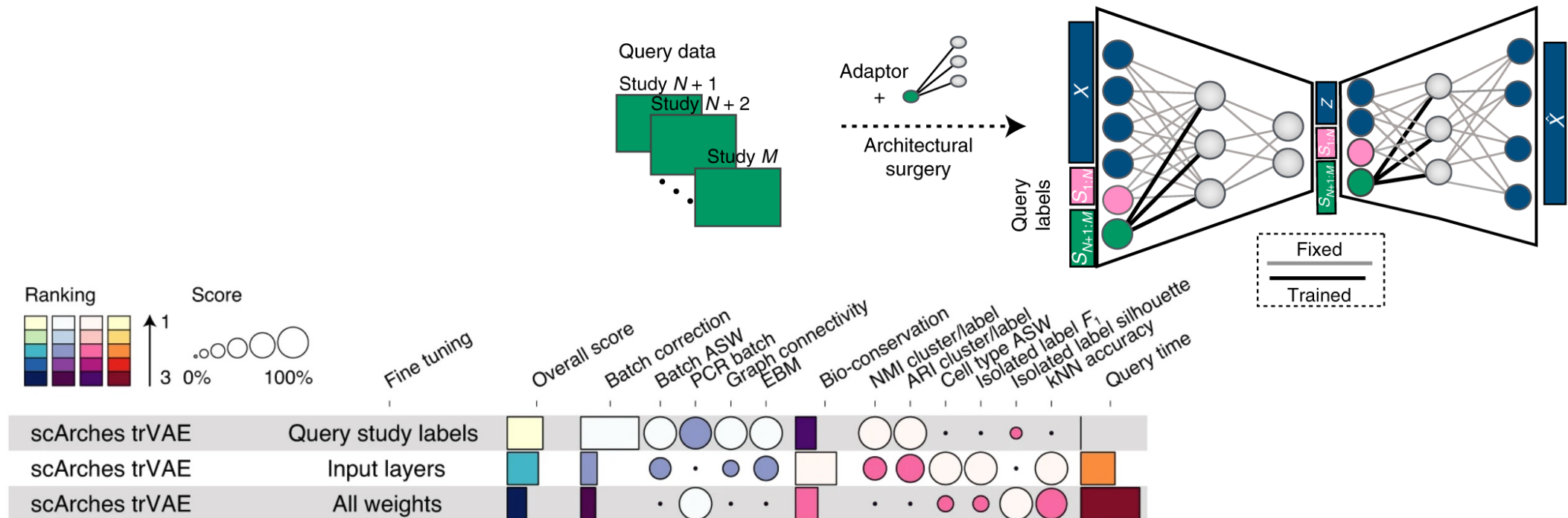
$$l_{\text{MMD}}(X, X') = \frac{1}{N_0^2} \sum_{n=1}^{N_0} \sum_{m=1}^{N_0} k(x_n, x_m) + \frac{1}{N_1^2} \sum_{n=1}^{N_1} \sum_{m=1}^{N_1} k(x'_n, x'_m) - \frac{2}{N_0 N_1} \sum_{n=1}^{N_0} \sum_{m=1}^{N_1} k(x_n, x'_m).$$

- $k(x, y)$ : Gaussian kernel similarity between two points
- Larger MMD  $\rightarrow$  more separation between the two datasets
- MMD loss can lead to over-correction if different datasets are biologically very different
- The authors suggest putting the MMD penalty on the first decoder layer instead of the bottleneck to further reduce correlation between  $Z$  and  $S$



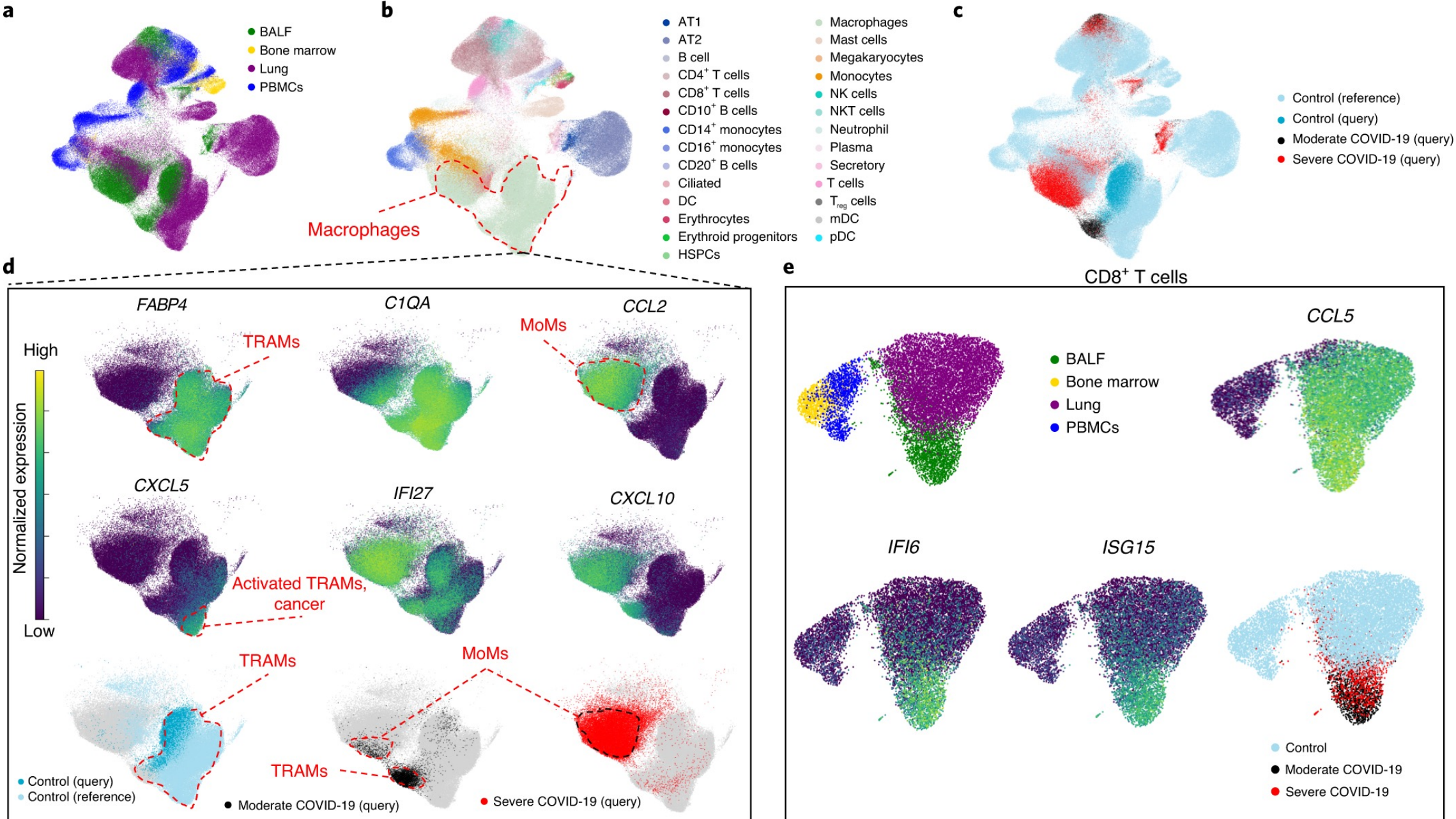
# scArches (Lotfollahi et. al., Nature Biotech 2022)

- Main idea
  - Pretrain reference data using a similar framework as scVI
  - Map target data onto reference data by minimal fine-tuning the pre-trained model
    - Add extra nodes in input and bottleneck layer to indicate new dataset (and also new batches)
    - Only train weights from the new nodes
    - Their empirical experiments suggest that keeping all weights related to reference data frozen performs the best in mixing reference data with query (target) data





# scArches (Lotfollahi et. al., Nature Biotech 2022)



# Reference mapping in Seurat V4 (Hao et. al. Cell, 2021)

- Can integrate multi-modal data (we only describe the version for scRNA-seq data here)
- Low-dimensional projection using sPCA
  - Project the reference data by  $Z = U^T X$ , and then project the query data using the same  $U$ 
    - Can not use CCA any more
  - How to find  $U$ ?
    - Construct a cell-cell similarity matrix  $L$  (for example from KNN)
    - Find  $U$  that maximized the Hilbert-Schmidt Independence Criterion (HSIC):

$$HSIC \left( (U^T X)^T U^T X, L \right)$$
$$= \frac{1}{(n-1)^2} tr \left( X^T U U^T X H L H \right)$$

where  $H$  is the centering matrix  $H_{ij} = I - n^{-1} e e^T$ .

- This is equivalent to

$$\operatorname{argmax}_U tr(U^T X H L H X^T U)$$

$$\text{subject to } U^T U = I$$

- Solution:  $U$  is the eigenvector of matrix  $X H L H X^T$  (PCA: eigenvector of  $X H H X^T = X H X^T$ )
- In Seurat V5 they will use Laplacian eigen decomposition (will discuss in later lectures)

# Reference mapping in Seurat V4 (Hao et. al. Cell, 2021)

- Can integrate multi-modal data (we only describe the version for scRNA-seq data here)
- Low-dimensional projection using sPCA
- Problem with CCA: can not keep the reference embeddings fixed
- Find anchor cell pairs between the reference data and the query data
- Project the query data onto the reference using the kernel weighting of anchor differences vectors as in Seurat CCA V2 (Seurat V3)
  - Define the weight matrix between all query cells and anchor cells as matrix  $W$
- Cell type label transfer:
  - Assign the same cell type label to anchor cells in the query data by the cell type labels of their pairs in the reference dataset
    - Prediction score of the transferred labels:

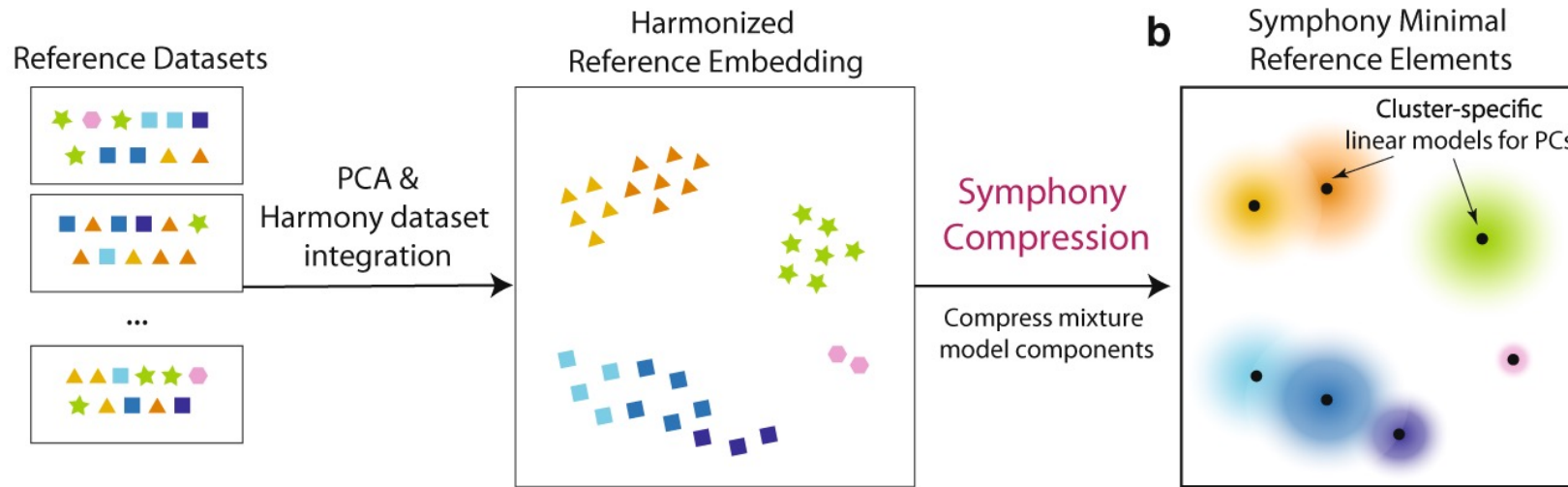
$$P_l = LW^T$$

$L$  are the labels of reference anchors

- Should be easy to assign an anchor similarity score to each cell to identify cells that can not be assigned well (unknown new cell types) [Similar idea implemented in scArches]

# Symphony (Kang et. al., Nature Communications 2021)

- Cell-cell similarity based reference mapping for joint visualization and label transfer
- Main Steps
  - Integrate reference data from different batches using Harmony



- Project the query data on the PC space of reference data by linear rotation

$$\mathbf{Z}_q = \mathbf{U}^T \mathbf{G}_{qs}$$

- Soft assign cells to reference clusters
- Move query cells within each cluster by subtracting the batch and cluster specific mean effect



# Symphony (Kang et. al., Nature Communications 2021)

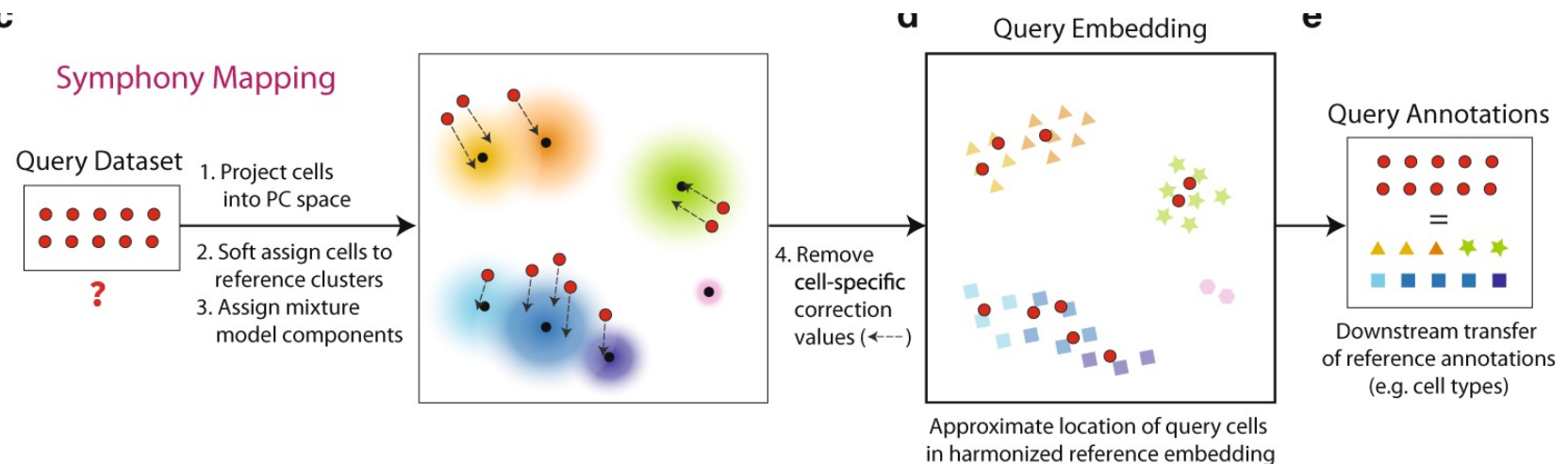
- Cell-cell similarity based reference mapping for joint visualization and label transfer

- Main Steps

- Integrate reference data from different batches using Harmony
- Project the query data on the PC space of reference data by linear rotation

$$\mathbf{Z}_q = \mathbf{U}^T \mathbf{G}_{qs}$$

- Soft assign cells to reference clusters
  - Assumes that there is no new unknown cell type
- Move query cells within each cluster by subtracting the batch and cluster specific mean effect

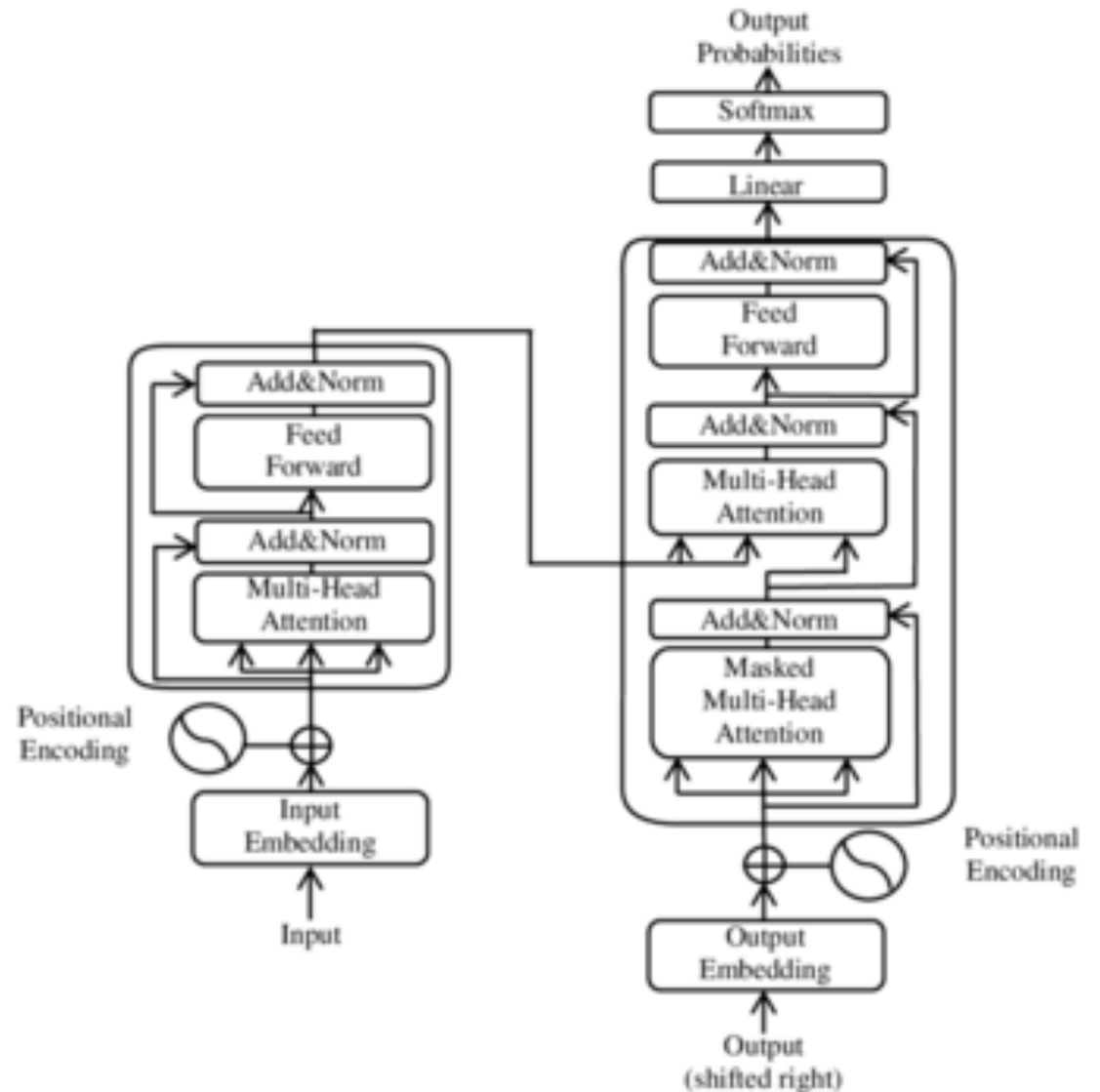


# New deep learning-based methods using transformer

- Instead of using autoencoder, researcher have also tried using more complicated deep learning models like transformer
- Youtube video from StatQuest for an relatively easy introduction of transformer:

<https://www.youtube.com/watch?v=zxQTK8quyY>

- Compared to autoencoder
  - Provides embedding of each gene
  - Explicitly make use of gene-gene similarity by self-attention
  - Multi-head attention sounds like bagging?

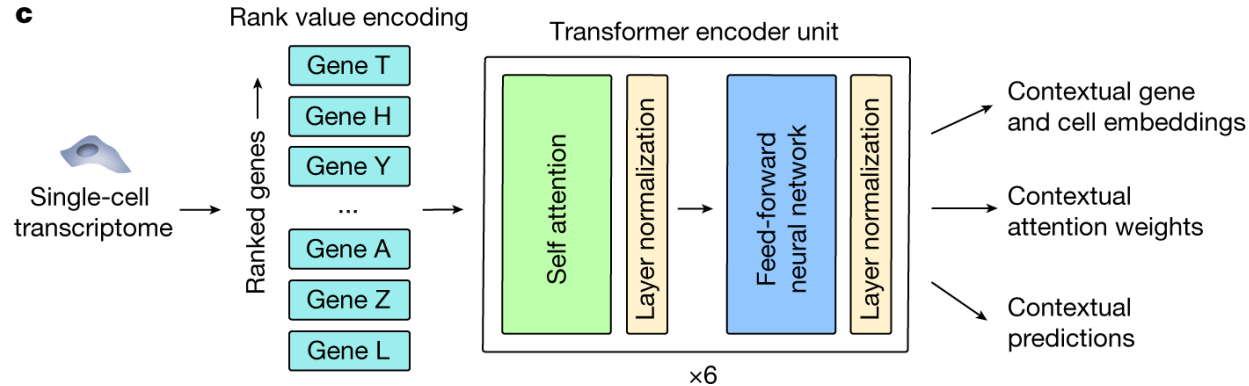


# Geneformer (Theodoris et. al., Nature 2023)

- Pre-trained model is based on 40M human cells from 561 datasets using droplet-based platforms
- Labels of a cell include: organ, platform, cell type (if provided by the original order)

## Pretraining

- Instead of using the original gene expression, use the ranking of genes (after scaling) within a cell type as the input (similar to quantile normalization)
  - That creates a position of a gene (word) within a cell (sentence)



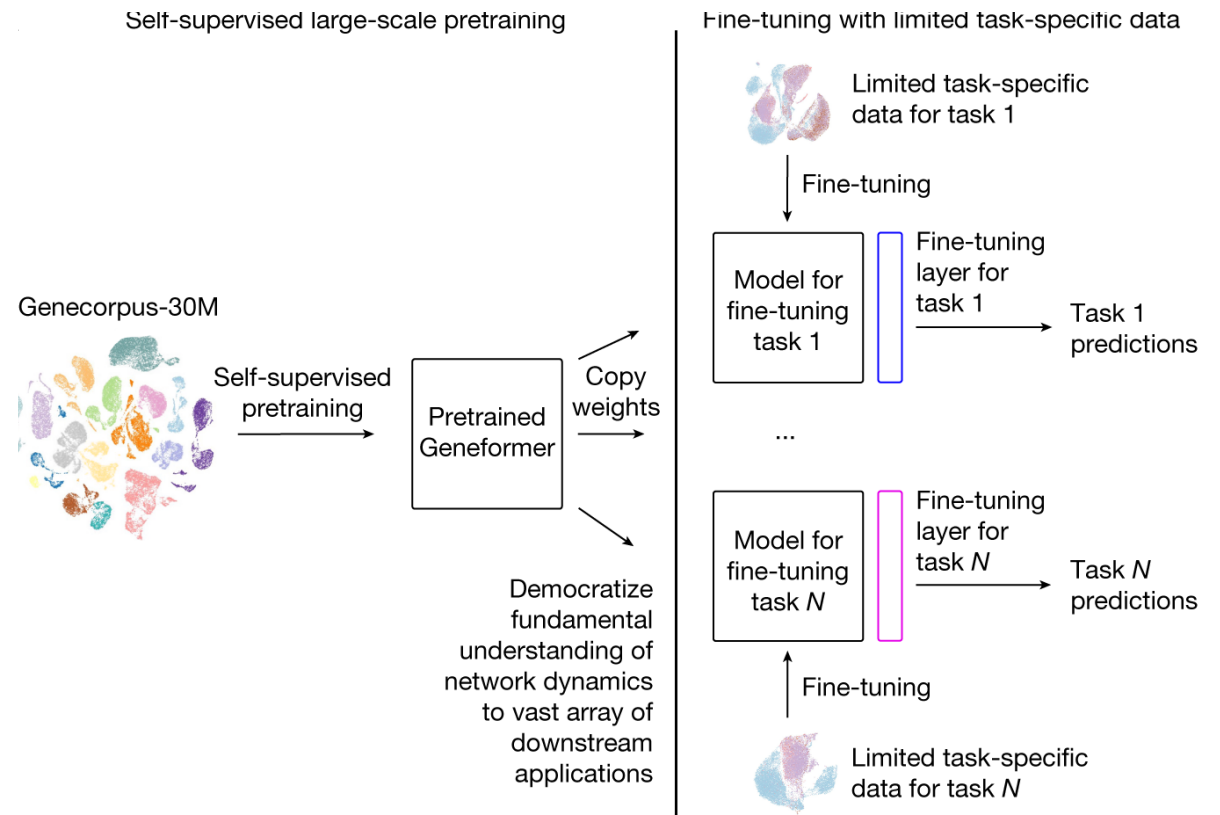
- The self-attention layers create embeddings of each gene
- Cell embedding can be obtained by weighted average of gene embeddings
- Unsupervised learning (no decoder units)
  - Objective function: prediction accuracy of randomly masked genes

# Geneformer (Theodoris et. al., Nature 2023)

- Pre-trained model is based on 40M human cells from 561 datasets using droplet-based platforms
- Labels of a cell include: organ, platform, cell type (if provided by the original order)

## Pretraining

- Fine-tuning
  - Specific tasks: gene classification, cell classification
  - Add a final task-specific transformer layer
  - Initialize the model with pretrained weights



# Related papers

- Han, X., Wang, R., Zhou, Y., Fei, L., Sun, H., Lai, S., ... & Guo, G. (2018). Mapping the mouse cell atlas by microwell-seq. *Cell*, 172(5), 1091-1107.
- Schaum, N., Karkanias, J., Neff, N. F., May, A. P., Quake, S. R., Wyss-Coray, T., ... & Weissman, I. L. (2018). Single-cell transcriptomics of 20 mouse organs creates a Tabula Muris: The Tabula Muris Consortium. *Nature*, 562(7727), 367.
- "A single-cell transcriptomic atlas characterizes ageing tissues in the mouse." *Nature* 583, no. 7817 (2020): 590-595.
- Wang, J., Agarwal, D., Huang, M., Hu, G., Zhou, Z., Ye, C., & Zhang, N. R. (2019). Data denoising with transfer learning in single-cell transcriptomics. *Nature methods*, 16(9), 875-878.
- Lotfollahi, M., Naghipourfar, M., Luecken, M. D., Khajavi, M., Büttner, M., Wagenstetter, M., ... & Theis, F. J. (2022). Mapping single-cell data to reference atlases by transfer learning. *Nature biotechnology*, 40(1), 121-130.
- Hao, Y., Hao, S., Andersen-Nissen, E., Mauck, W. M., Zheng, S., Butler, A., ... & Satija, R. (2021). Integrated analysis of multimodal single-cell data. *Cell*, 184(13), 3573-3587.
- Kang, J. B., Nathan, A., Weinand, K., Zhang, F., Millard, N., Rumker, L., ... & Raychaudhuri, S. (2021). Efficient and precise single-cell reference atlas mapping with Symphony. *Nature communications*, 12(1), 5890.
- Theodoris, C. V., Xiao, L., Chopra, A., Chaffin, M. D., Al Sayed, Z. R., Hill, M. C., ... & Ellinor, P. T. (2023). Transfer learning enables predictions in network biology. *Nature*, 618(7965), 616-624.