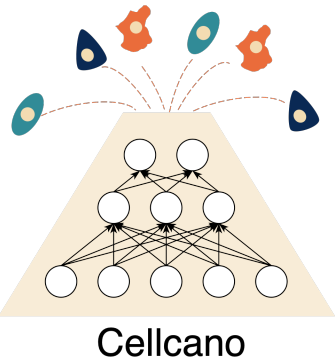
Cellcano: supervised cell type identification for single cell ATAC-seq data

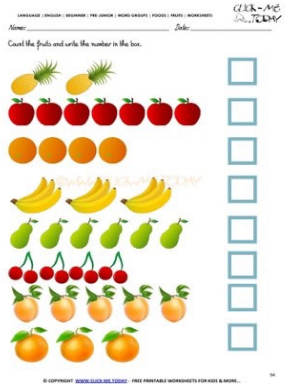
Hao Wu
Department of Biostatistics and Bioinformatics
Rollins School of Public Health
Emory University



Single cell omics data

Bulk seq

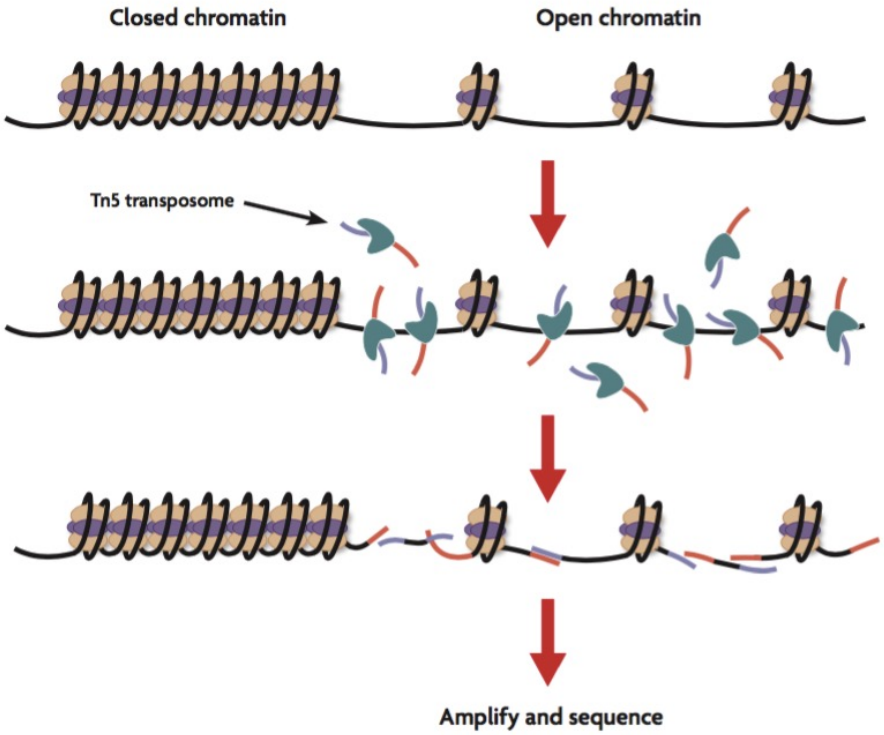
sc seq



ATAC-seq

Assay for **T**ransposase-**A**ccessible **C**hromatin using **seq**uencing

- Measures the chromatin accessibility.
- Learn regulatory mechanism
- Reveal cell type or cell state

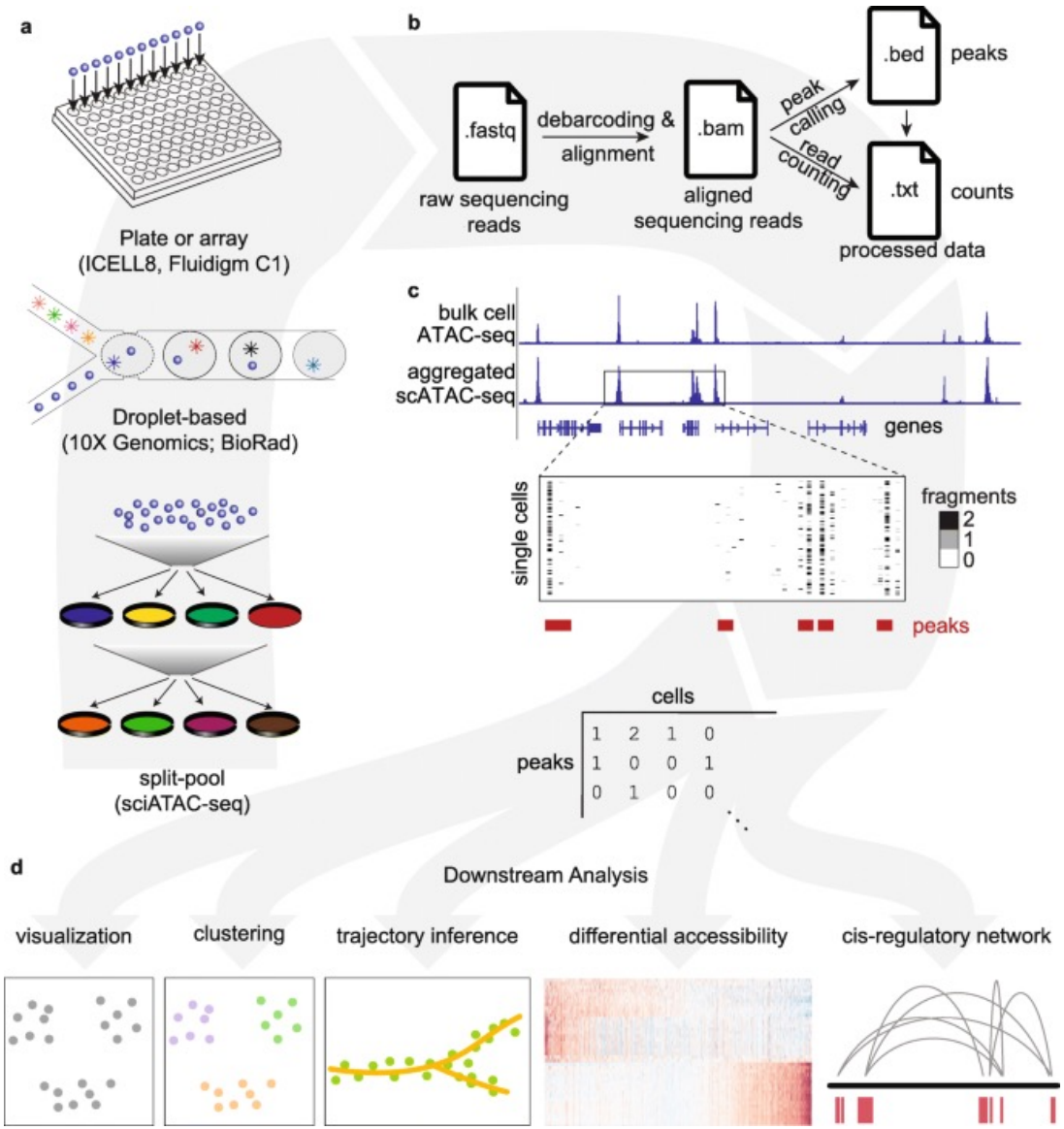


Source: https://www.activemotif.com/blog-atac-seq

scATAC-seq workflow

Single-cell ATAC-seq (scATAC-seq) measures chromatin accessibility in individual cells.

- Different from scRNA-seq: mRNA molecules vs DNA molecules
- Sparse and lower counts

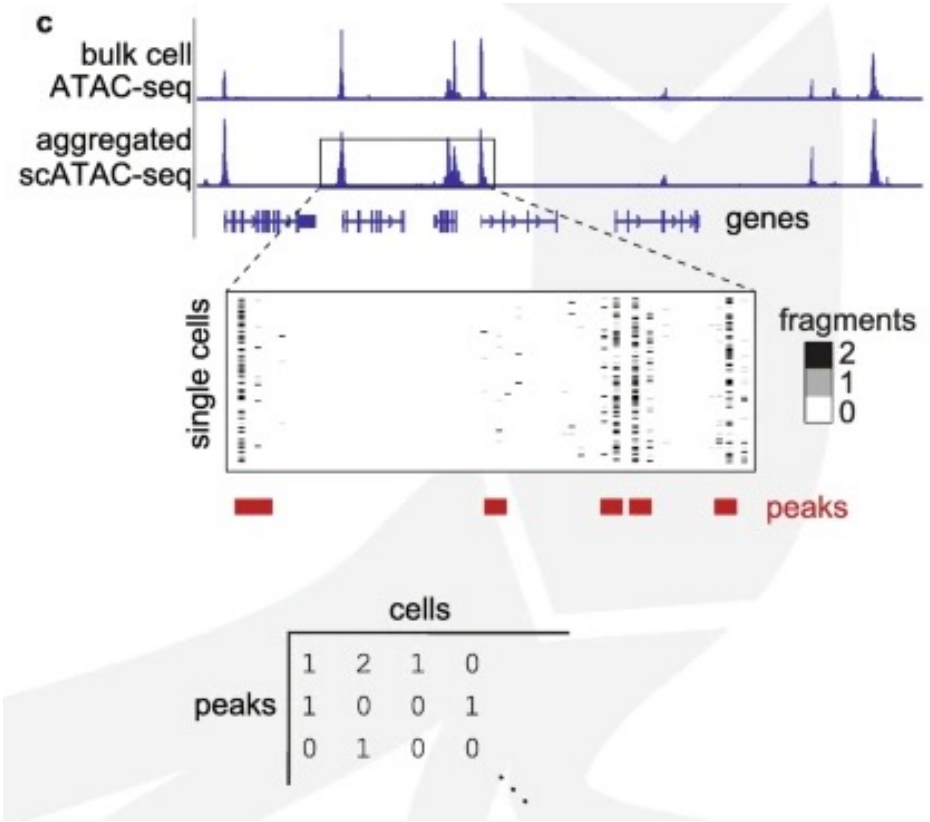


Chen et al. Genome Biology (2019).

scATAC-seq data

Raw data (sequence reads) can be summarized at different levels:

- Bin counts: read counts on genome-wide fixed-size bins.
- Peaks counts: call peaks first, then count number of reads in all peaks.
- Gene scores: summary the counts to gene levels.



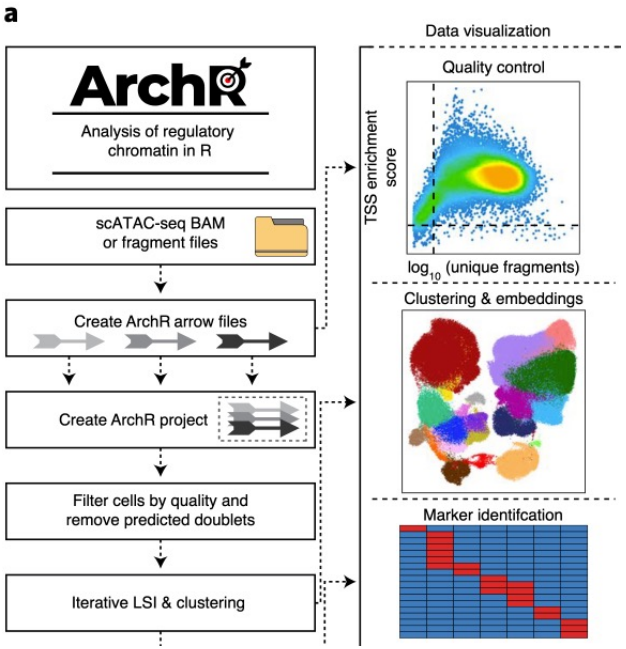
Chen et al. Genome Biology (2019). Granja et al. Nature Genetics (2021).

Computational cell type identification

- A fundamental task in single cell omics data.
- Commonly used methods in scRNA-seq:
  - Unsupervised methods: cell clustering + manual annotation.
  - Supervised methods: train a prediction model based on reference with known cell types.
  - Semi-supervised methods.
  - Supervised methods usually performs better.
- Method is limited in scATAC-seq.

Current practice in annotating cells in scATAC-seq

Unsupervised clustering + known peak markers / marker gene expression



Laborious and Slow!

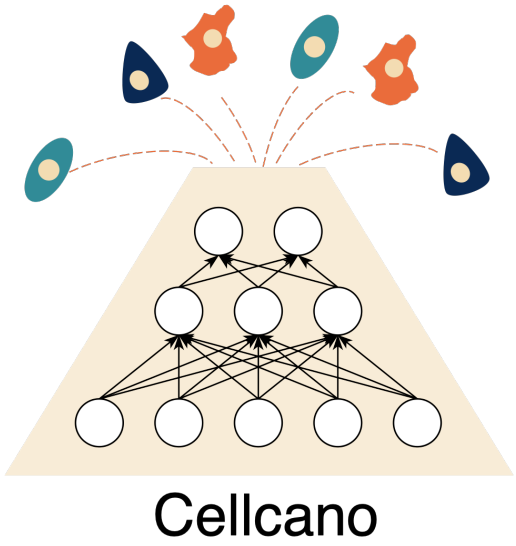
Granja et al. Nature Genetics (2021).

Supervised

- **Seurat**: borrow information from another modality (scRNA-seq).
- **SnapATAC**: Co-embed data into a low-dimension space and transfer cell labels with Seurat.
- **EpiAnno**: use peak-by-cell matrix and map source reads to target peaks

Our method: Cellcano

- Supervised celltyping for scATAC-seq, using scATAC-seq as reference.
- Fast and accurate.
- Provide pre-trained model for direct prediction.



Challenges

- Choice of input data
  - Bin counts, peak counts, or gene score.
- Choice of prediction method
  - Various off-the-shelf ML methods.
- How to deal with domain shift (batch effect).

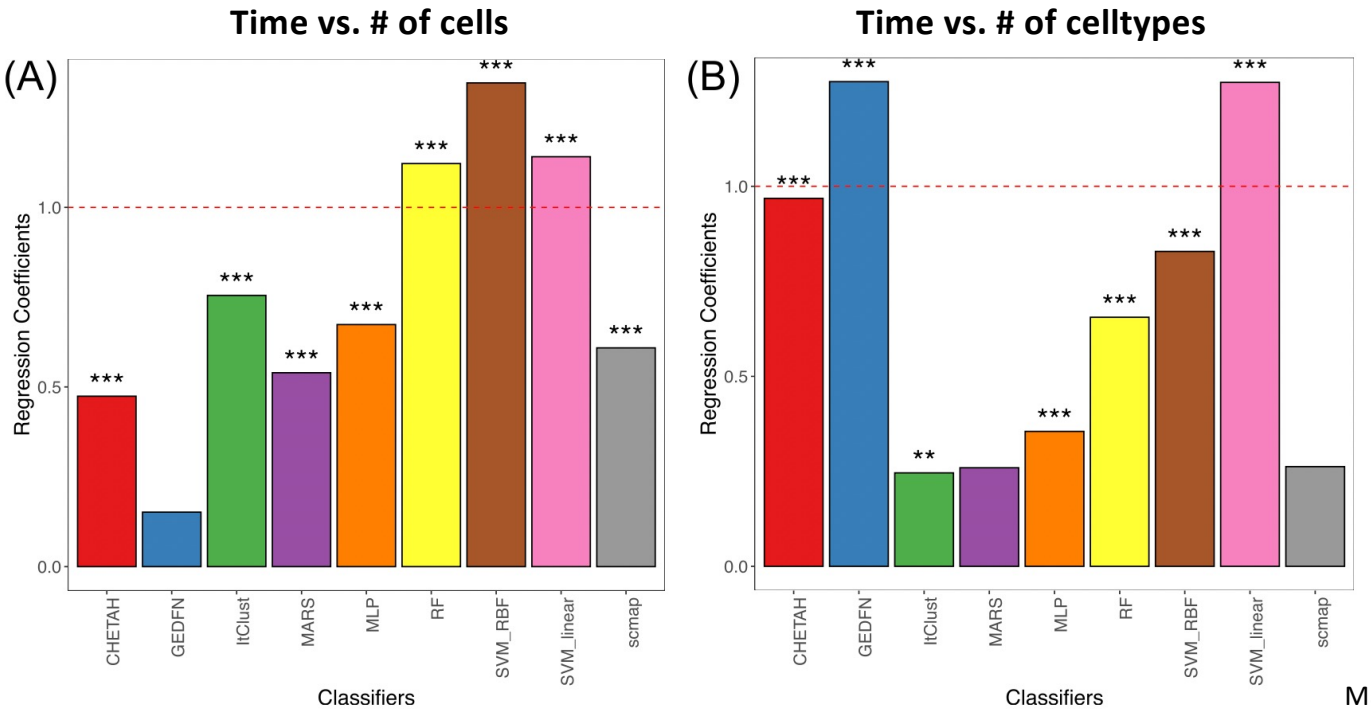
Choice of input: gene scores

- Bin counts: large
- Peak counts: not pre-defined, requires extra step of peak calling
- Gene scores:
  - Well-defined feature set
  - Small feature space
  - Easy to connect with other modality (such as scRNA-seq)

We use the "*GeneModel-GB-Exponential-Extend*" gene score provided by ArchR.

Choice prediction method: MLP

- Easy to implement
- Excellent computational performance

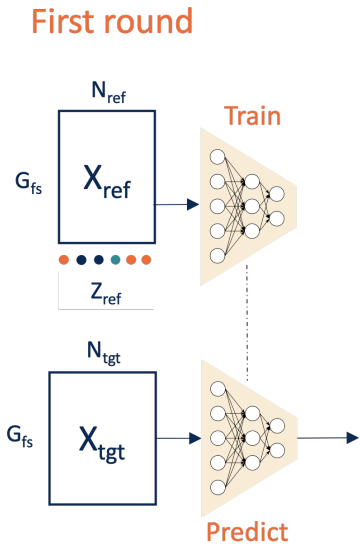


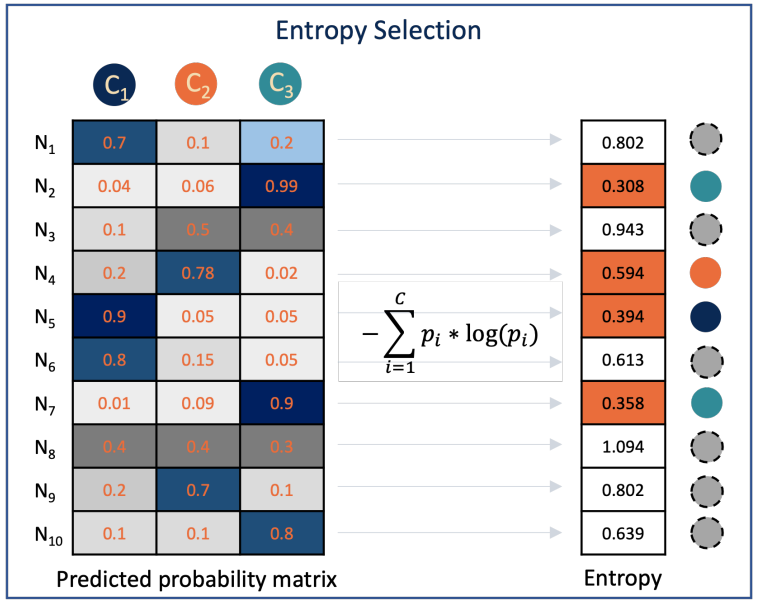
Ma and Wu (2021) Genome Biology

Deal with batch effect: a two-round procedure

Frist round:

Train an MLP from reference, predict in target, then select “anchor” cells based on prediction entropy.

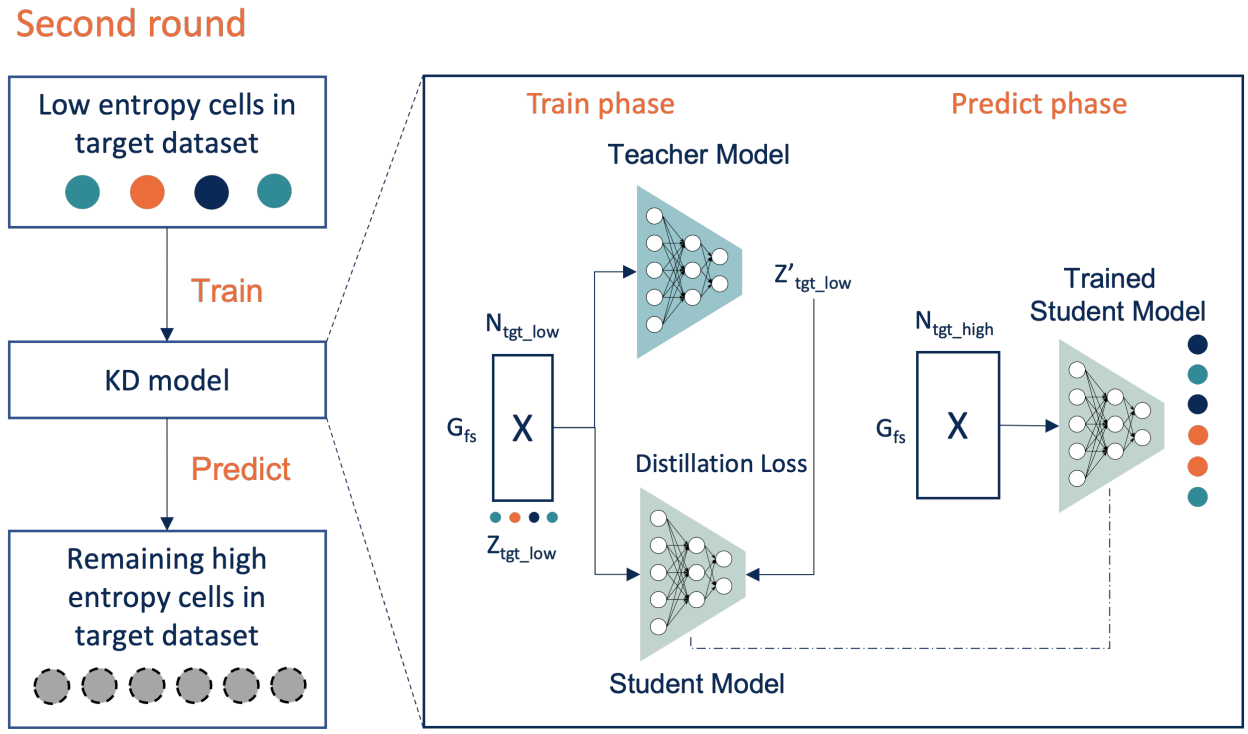




Deal with batch effect: a two-round procedure

Second round:

Train an MLP from anchor cells using a **knowledge-distillation** model, predict the remaining cells in target.



Knowledge Distillation (KD) model

- Learn a label smoothing regularization
- Works better with similar classes for preventing MLP being overconfident

Soft labels:

$$p_i = \frac{exp\left(\frac{z_i}{T}\right)}{\sum_j \exp\left(\frac{z_j}{T}\right)}$$

Z: logits
i: class i
p: probability
T: temperature

When T=1, it is a standard softmax function
If T grows, the probability becomes "softer"

Distillation Loss:

$$\mathcal{L}(x; W) = \alpha * \mathcal{H}(y, \sigma(z_s; T = 1)) + \beta * \mathcal{H}(\sigma(z_t; T = \tau), \sigma(z_s, T = \tau))$$

Weighted sum of "Student vs True label" and "Teacher vs Student", and they found best results when α is much smaller than β

x: input; y: ground truth
W: student model parameters
H: cross-entropy loss
σ: soft max function parametrized by T
α, β: hyperparameters, $\beta = 1 - \alpha$

Hinton et al. NIPS (2014). Yuan et al. CVPR (2020). https://intellabs.github.io/distiller/knowledge_distillation.html

A high level summary of Cellcano

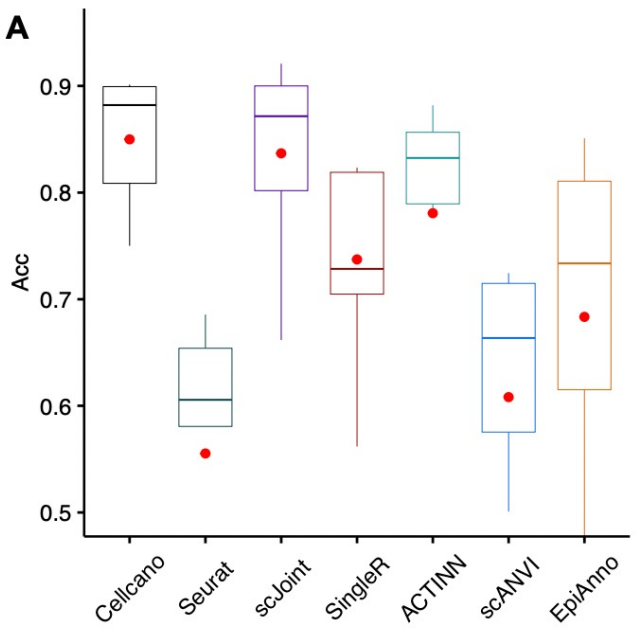
- Use gene score as input
- Use MLP as classifier
- Use a two-round procedure
  - The “self-supervised” training in the second step alleviates batch effect.
    - Seurat has similar approach, but its way to select and use anchor is not very good.
  - The KD model alleviates the imperfect label problem in the second round.

Benchmarking

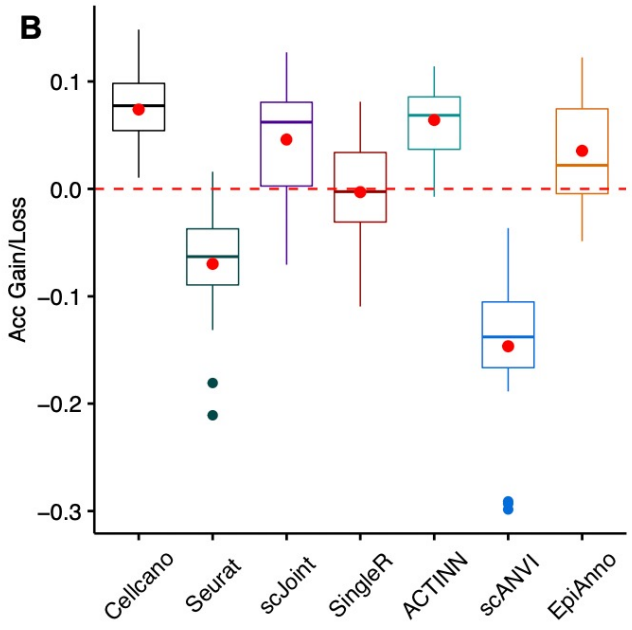
- Datasets
  - Human PBMC (4 datasets, 29 experiments)
  - Mouse Brain (2 datasets, 21 experiments)
- Methods under comparison
  - scATAC-seq methods: Seurat (changed to using scATAC-seq as reference), scJoint, EpiAnno.
  - scRNA-seq methods: SingleR, ACTINN, scANVI
- Metrics
  - Overall Accuracy (Acc)
  - Adjusted Rand Index (ARI)
  - Macro F1 score (macroF1)

Overall results

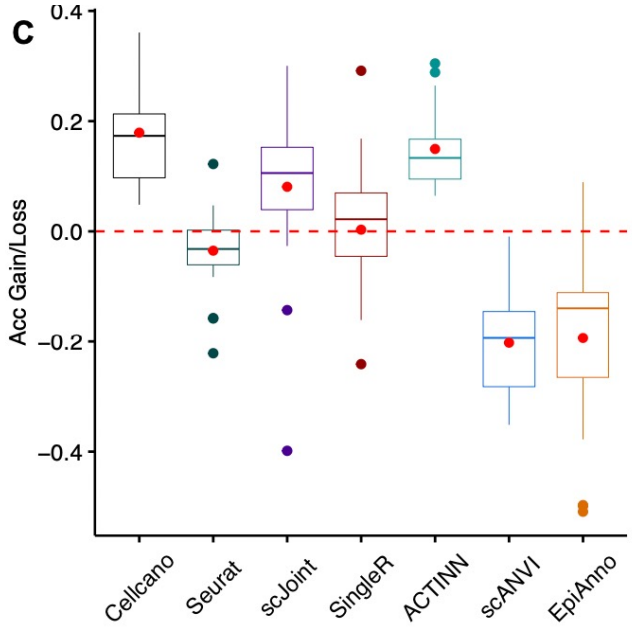
On gold standard data

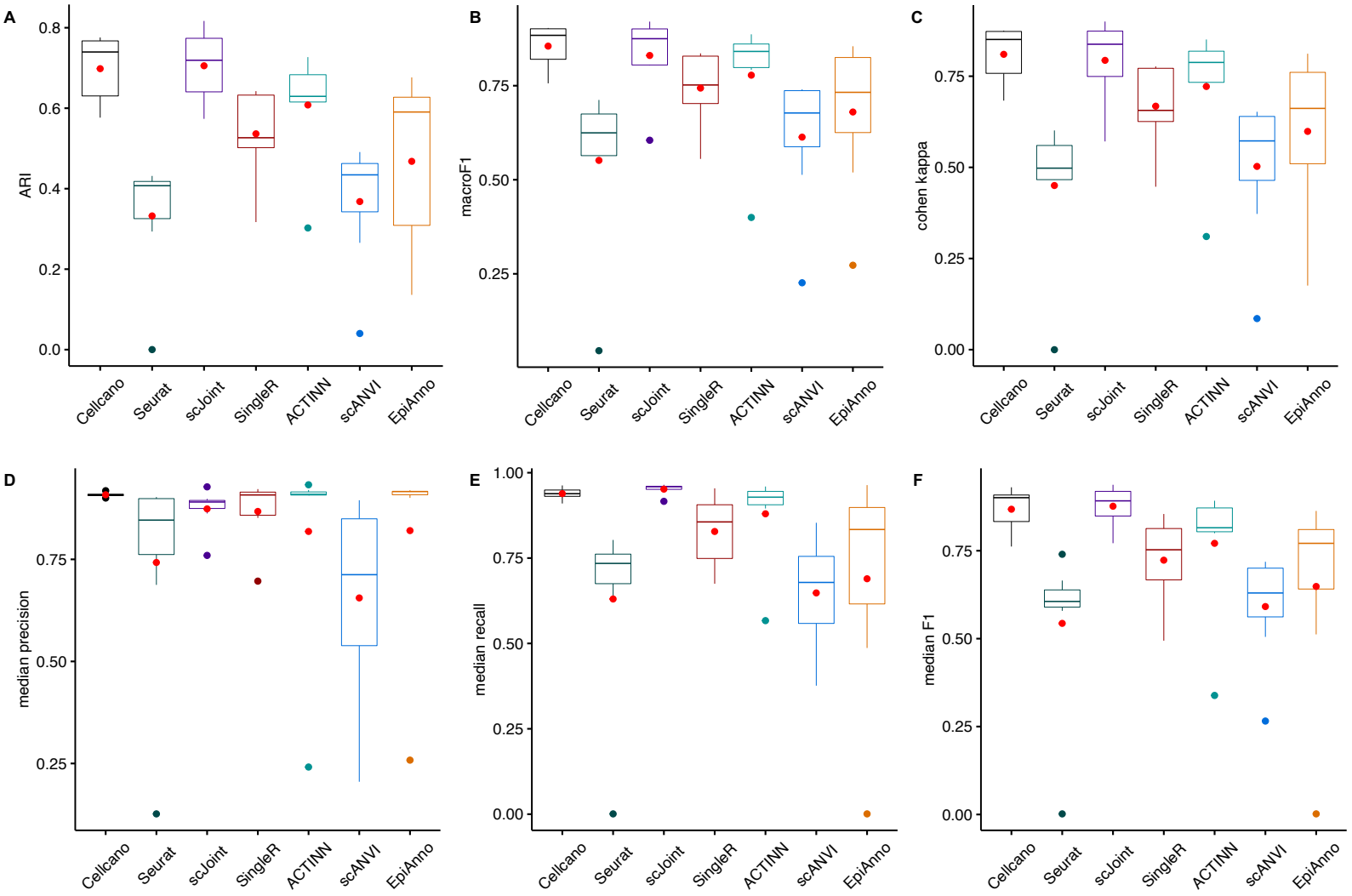


Human PBMC

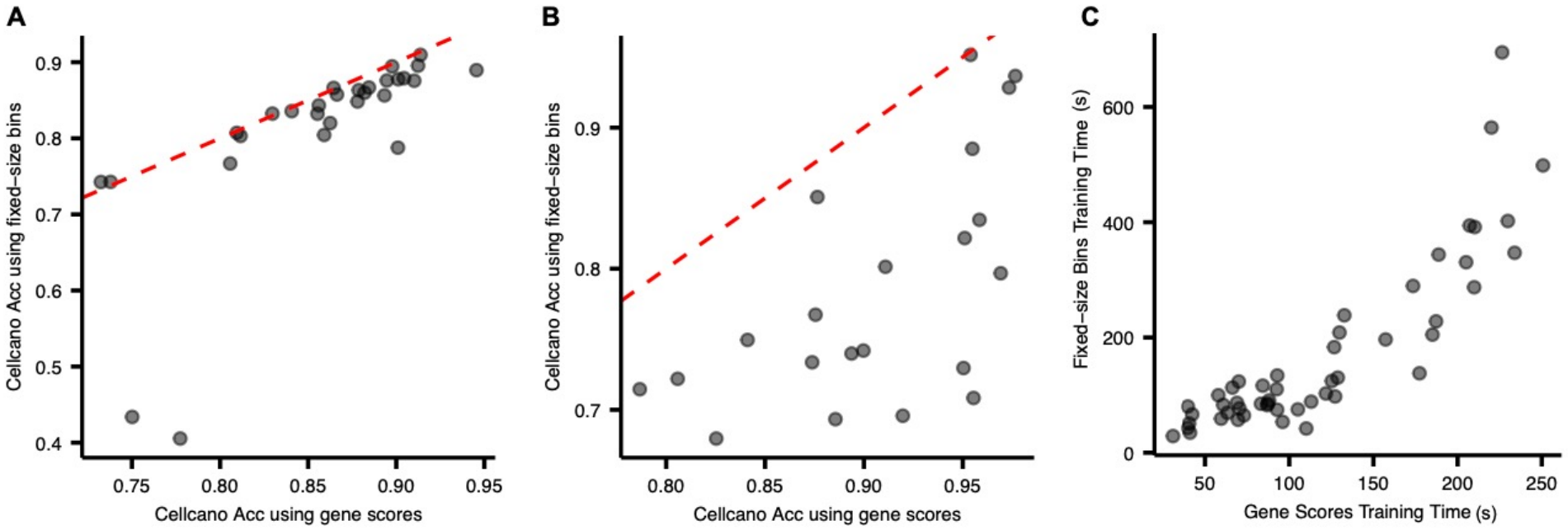


Mouse brain

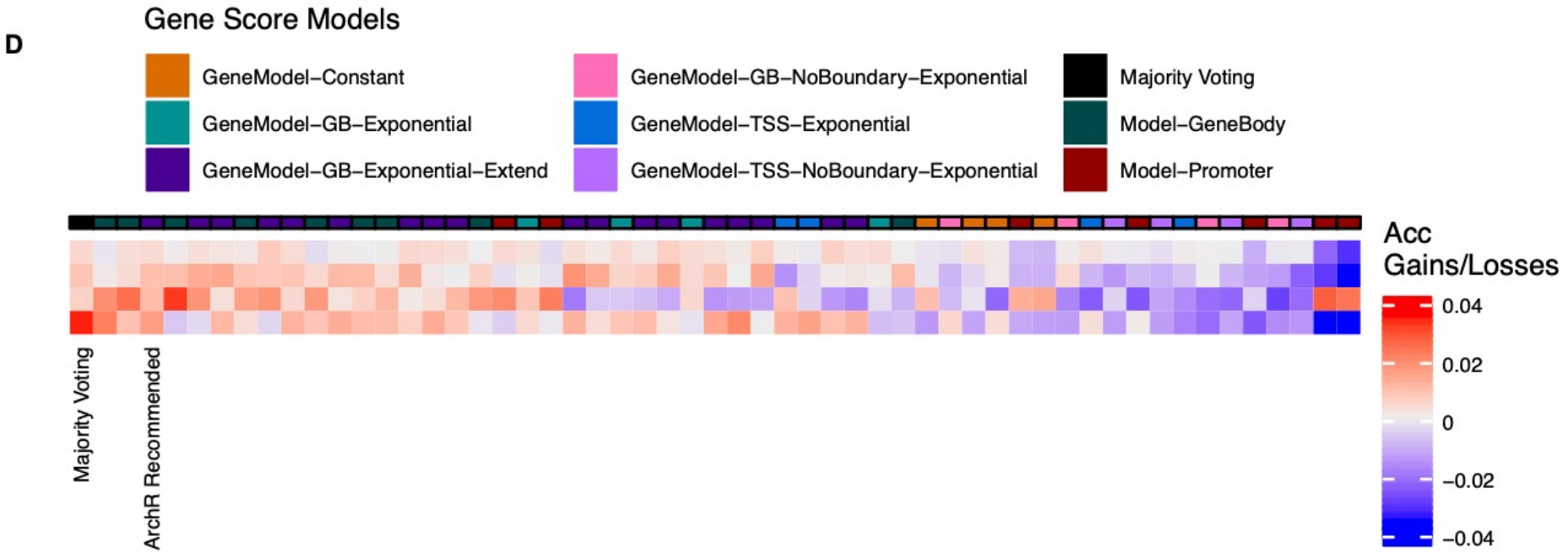



A
ARI
0.0
0.2
0.4
0.6
0.8
B
macroF1
0.25
0.50
0.75
C
cohen kappa
0.00
0.25
0.50
0.75
D
median precision
0.25
0.50
0.75
E
median recall
0.00
0.25
0.50
0.75
1.00
F
median F1
0.00
0.25
0.50
0.75
Cellcano
Seurat
scJoint
SingleR
ACTINN
scANVI
EpiAnno

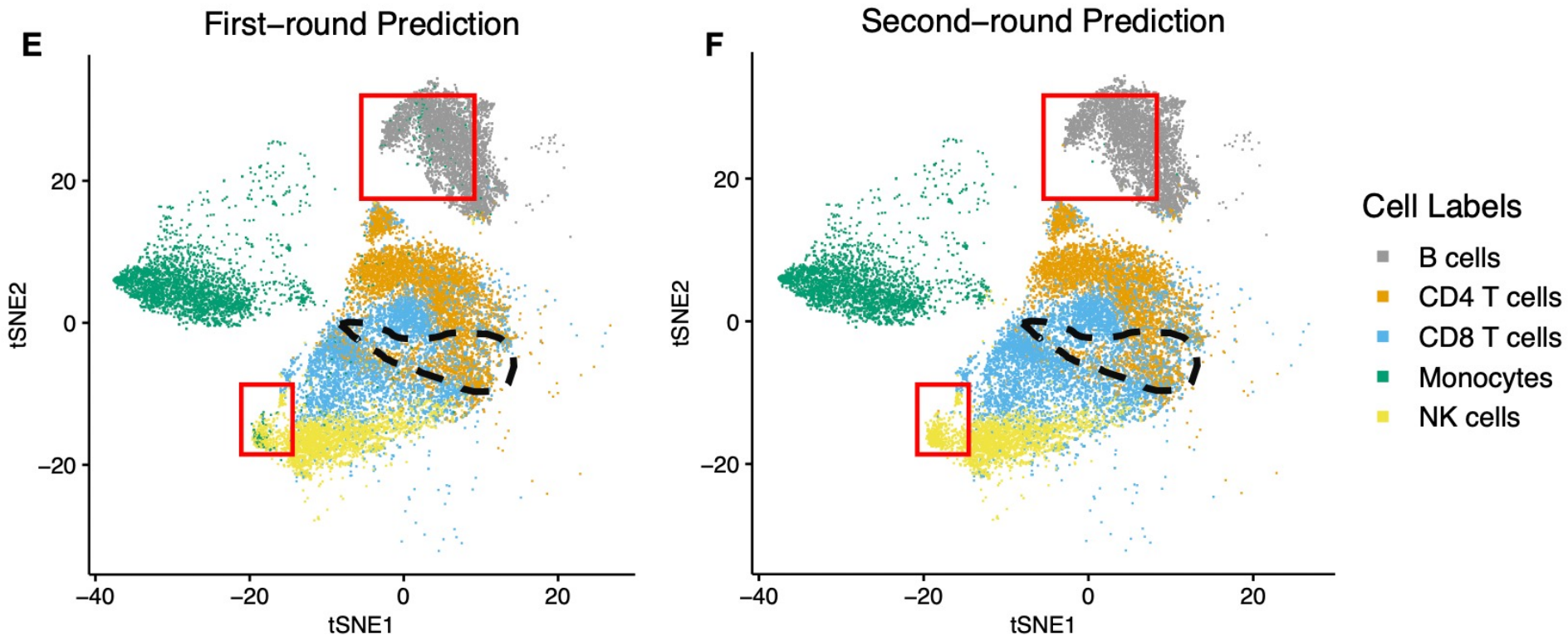
Gene score vs bin counts as input



Different gene score models

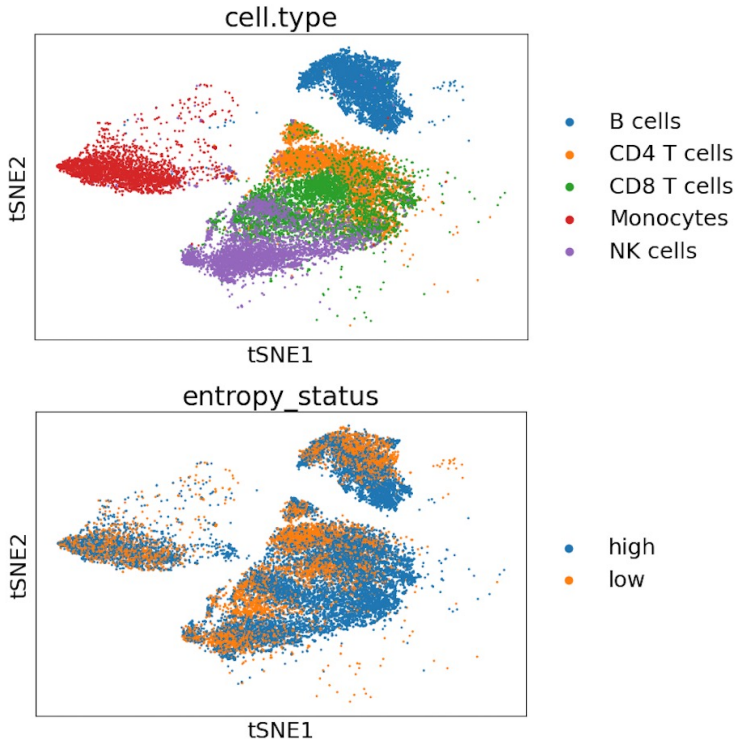


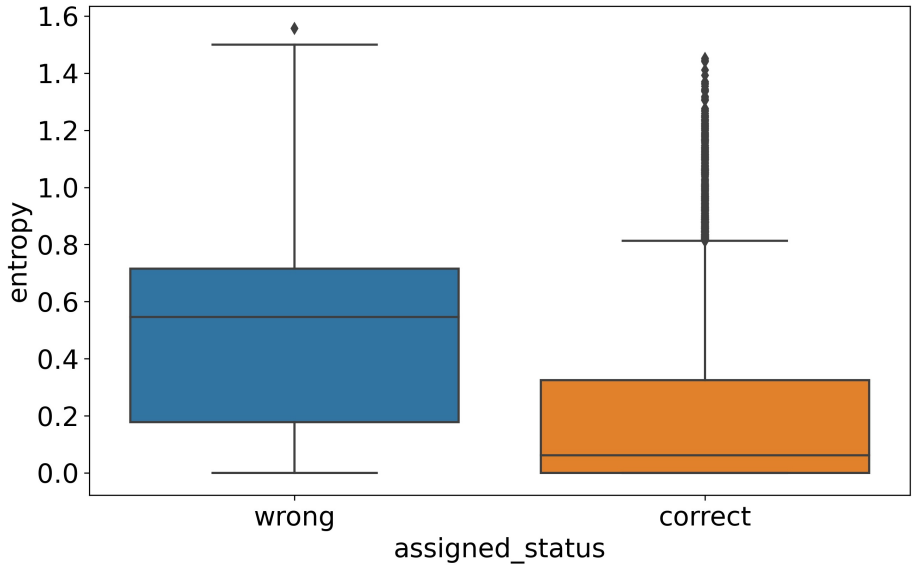
Effect of the two-round procedure



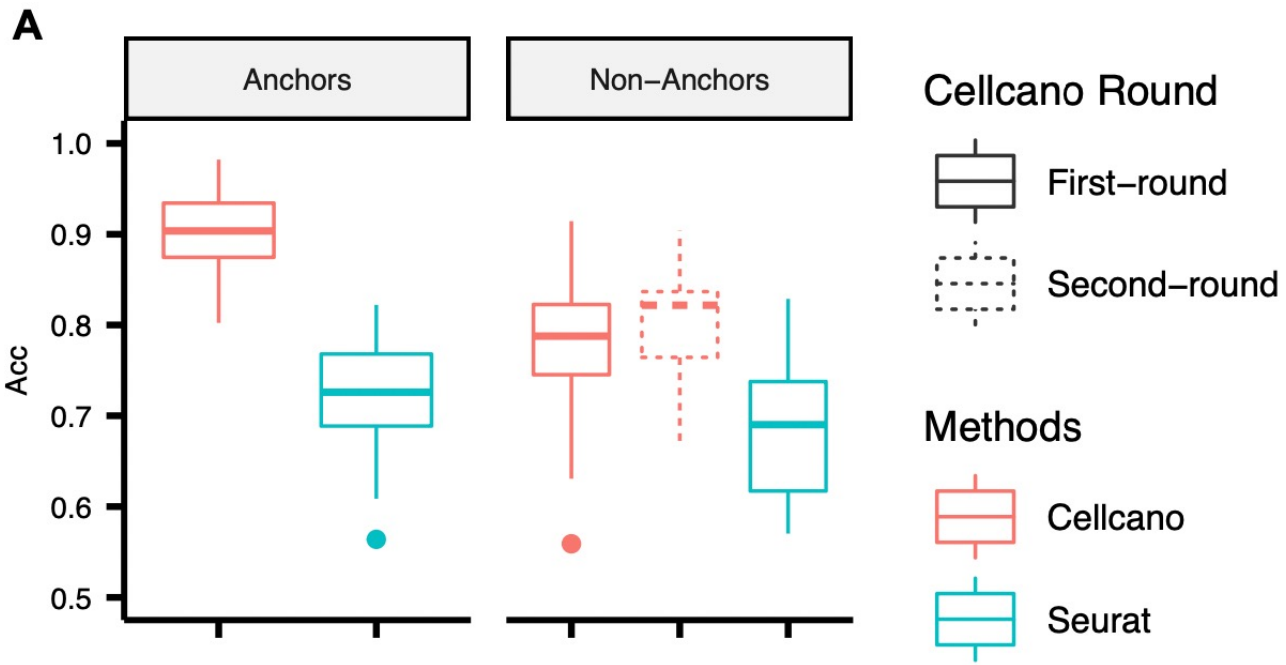
Anchor cell properties

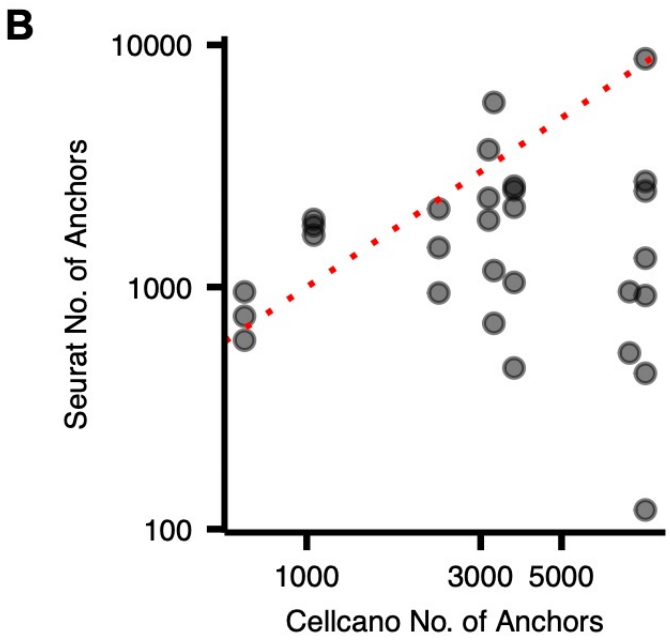
- Cells with low entropy (anchor cells) have higher accuracies.
- Cells from target can better capture target distribution.



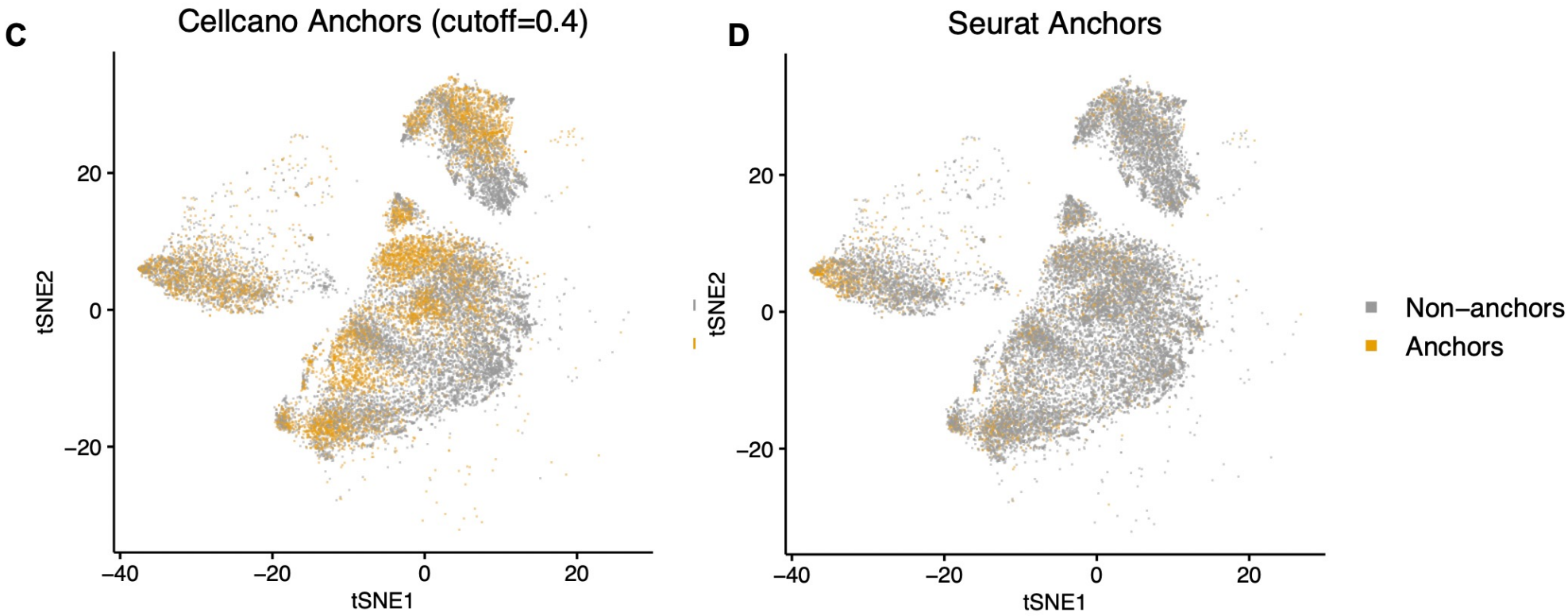


Anchor comparisons: Cellcano and Seurat



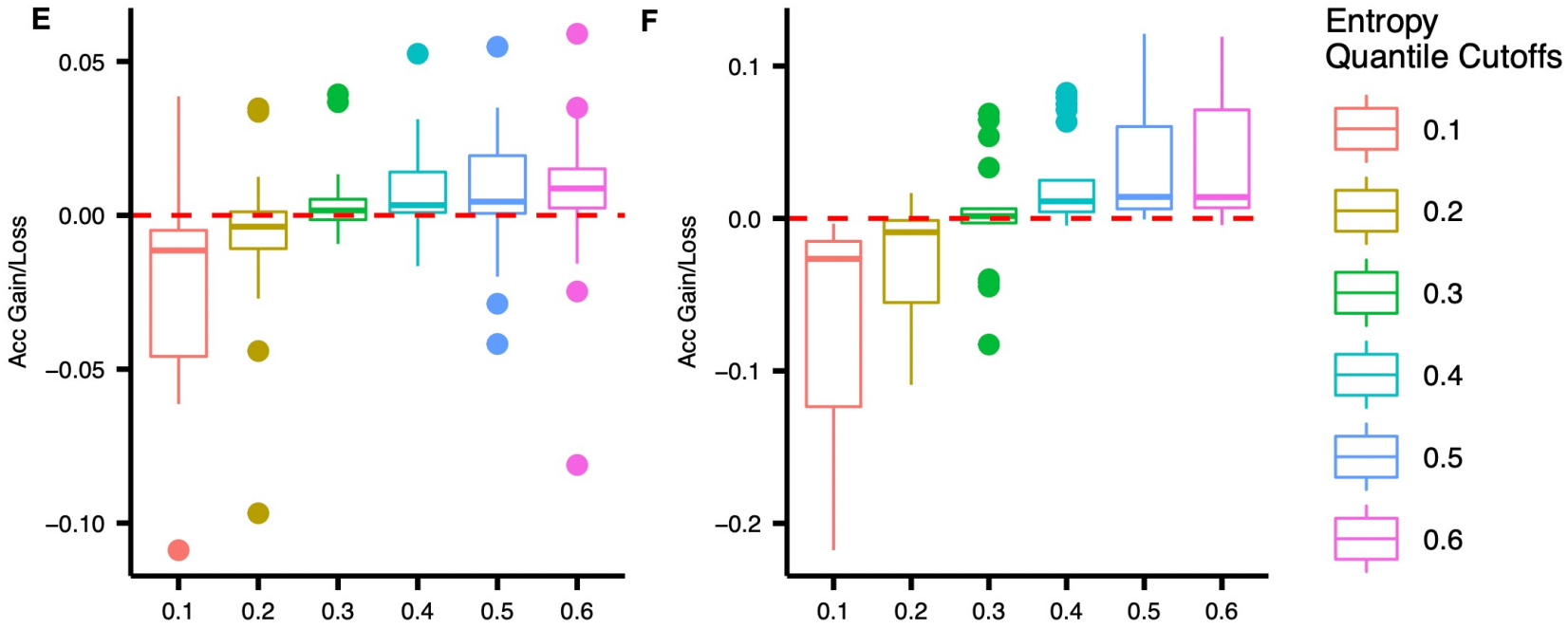


Anchor comparisons: Cellcano and Seurat

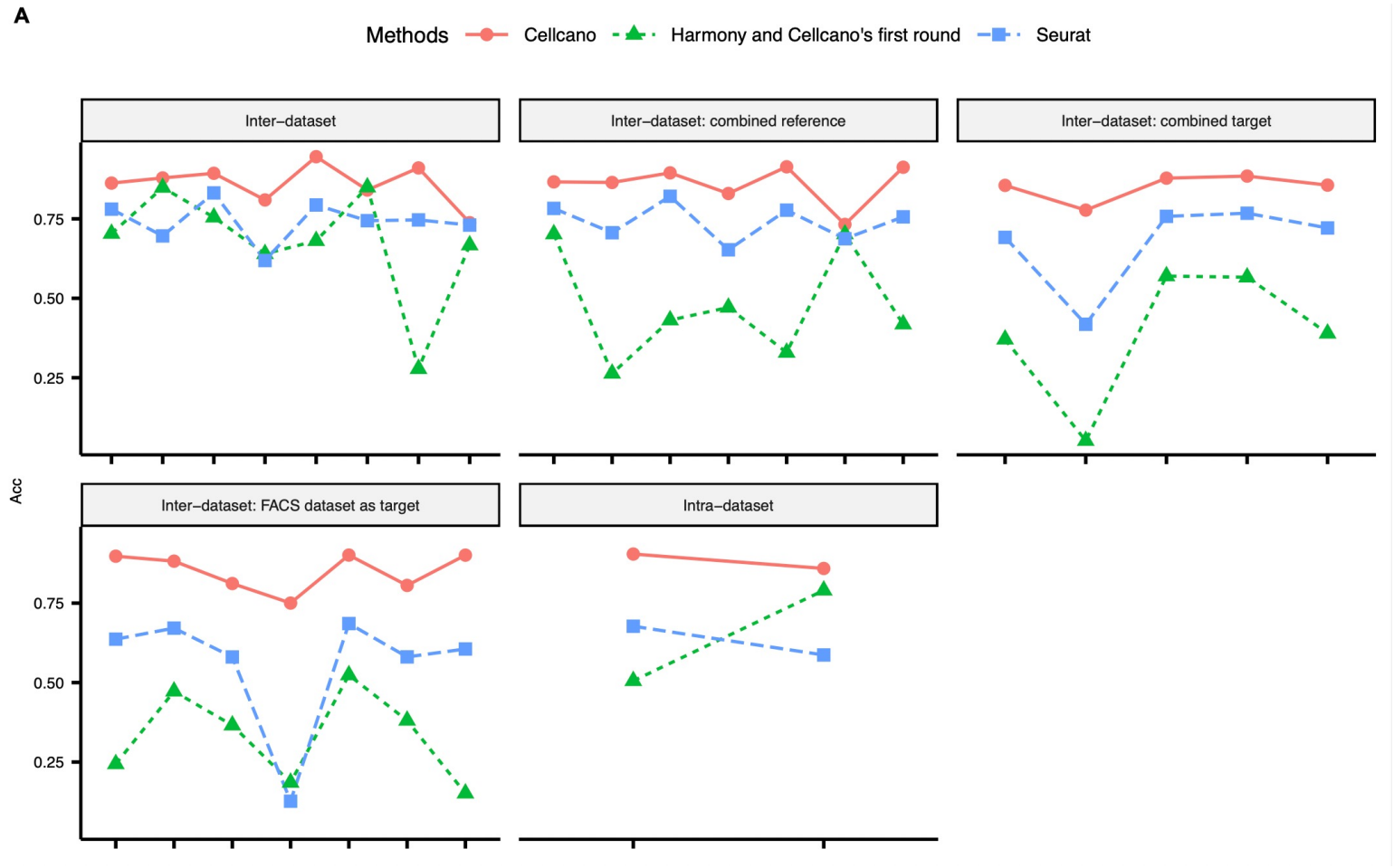


Number of anchor cells

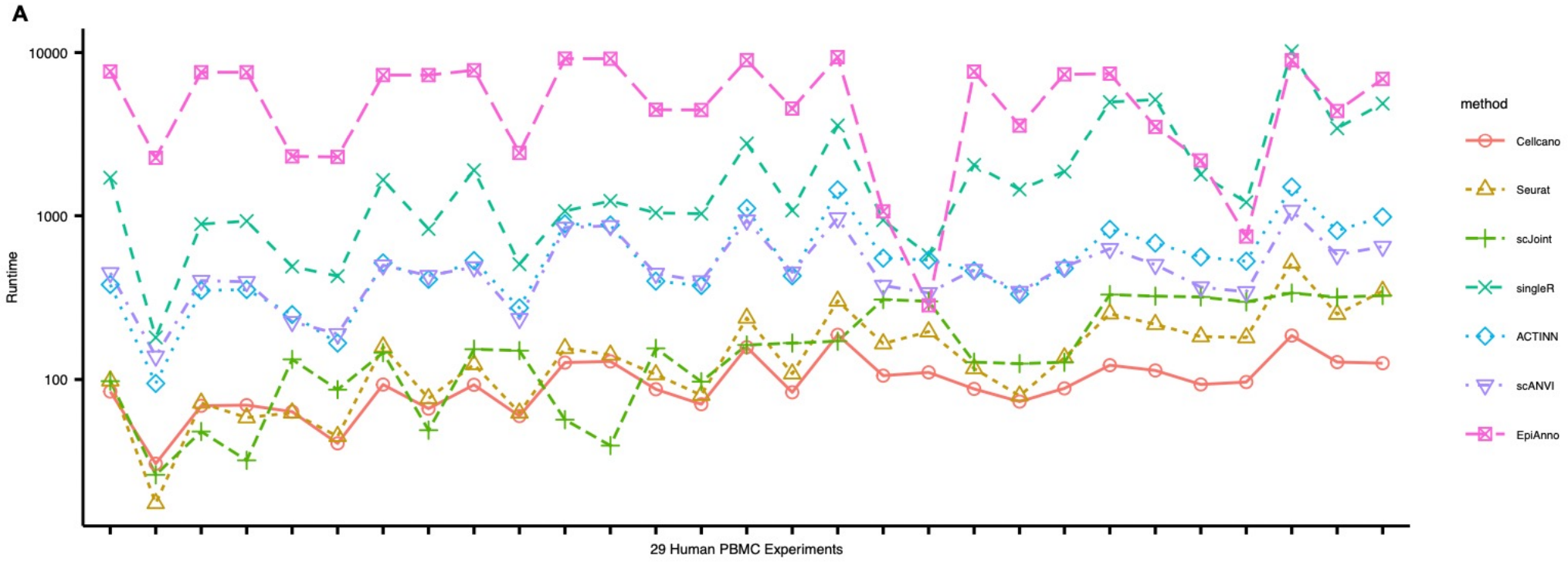
- Need to achieve a balance of accuracy and good training size.
- Default: 40% of the target cells.



Batch effect removal (Harmony) + MLP



Computational time



Takeaways

- Using gene score as input achieves similar result as using fixed-size bins as input.
- One gene score model as input for scATAC-seq celltyping is sufficient and can be further connected with scRNA-seq.
- The two-round strategy improves performance.
  - Better than using Harmony + one-step prediction
- The KD model helps to stabilize the results and deals better with similar cell types.

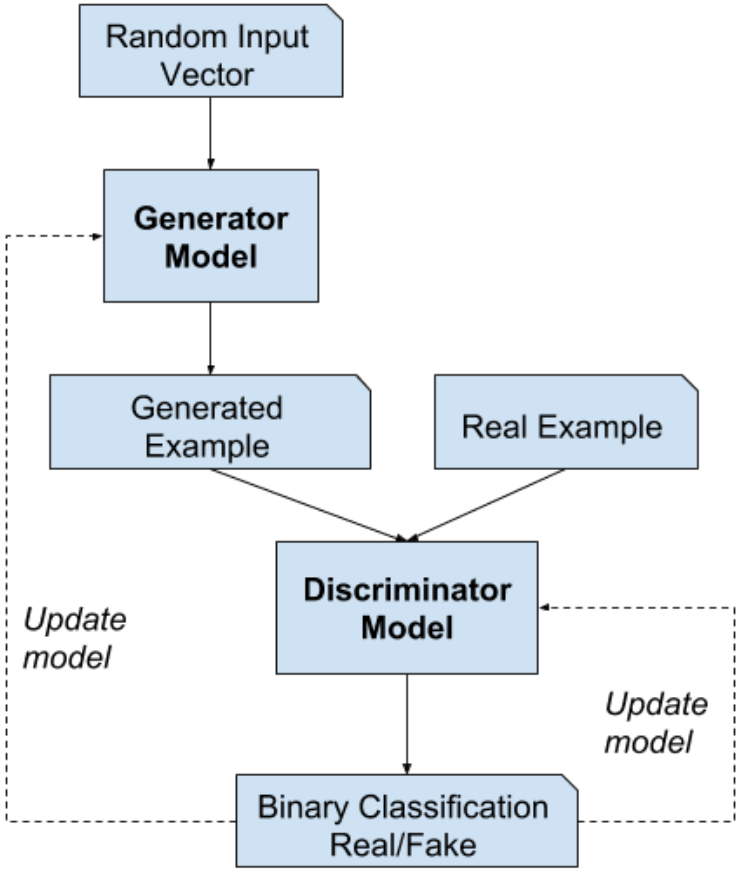
One step further

- Use scRNA-seq as reference for scATAC-seq celltyping
  - There are much more scRNA-seq data.
  - Method can potentially be extended to other single cell assays.
- Challenges
  - Domain shift: data distributions are very different.
- Existing methods: Seurat, scJoint.

Our idea

- Use Generative Adversarial Networks (GAN) for data harmonization.
- Transform scATAC-seq data to match scRNA-seq data.
- Use (pre-trained) scRNA-seq classifier on transformed scATAC-seq data to predict cell types.

Generative Adversarial Network (GAN)



- Generator G: learn the real data distribution
- Discriminator D: try to discriminate between generated and real samples

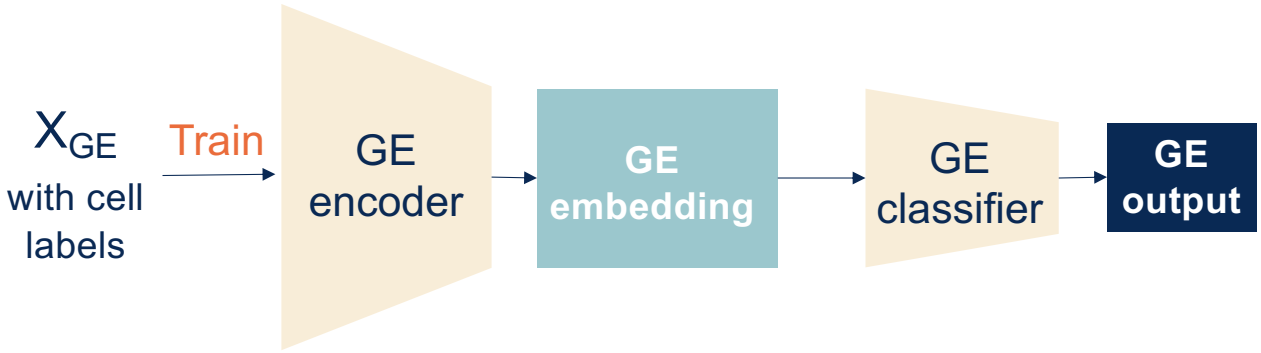
$$\min_G \max_D V(D,G) = \mathbb{E}_{\boldsymbol{x}\sim p_{\text{data}}(\boldsymbol{x})}[\log D(\boldsymbol{x})] + \mathbb{E}_{\boldsymbol{z}\sim p_{\boldsymbol{z}}(\boldsymbol{z})}[\log(1 - D(G(\boldsymbol{z})))].$$

- Both D and G can be neural networks

Goodfellow et al. 2014; https://machinelearningmastery.com/what-are-generative-adversarial-networks-gans/

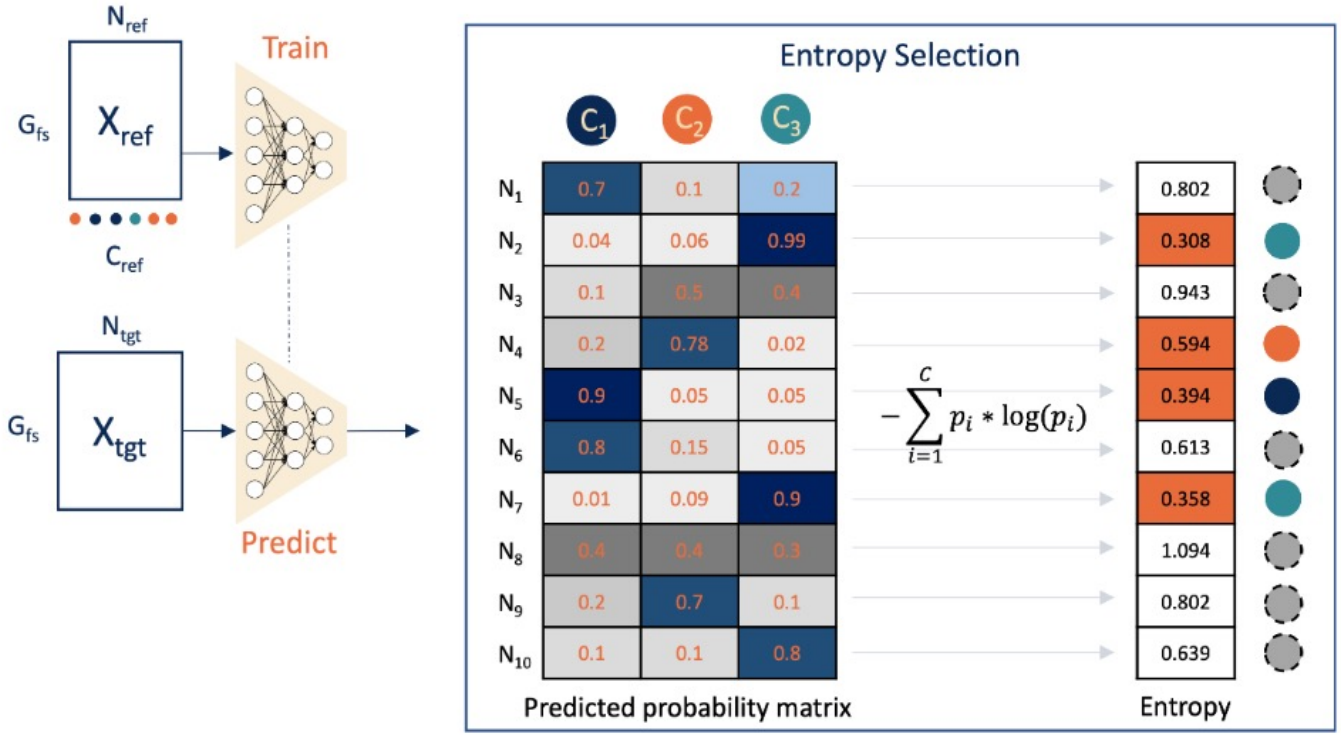
CellGAN

Step 1: Pre-training (skip if a pre-trained model on scRNA-seq is provided)

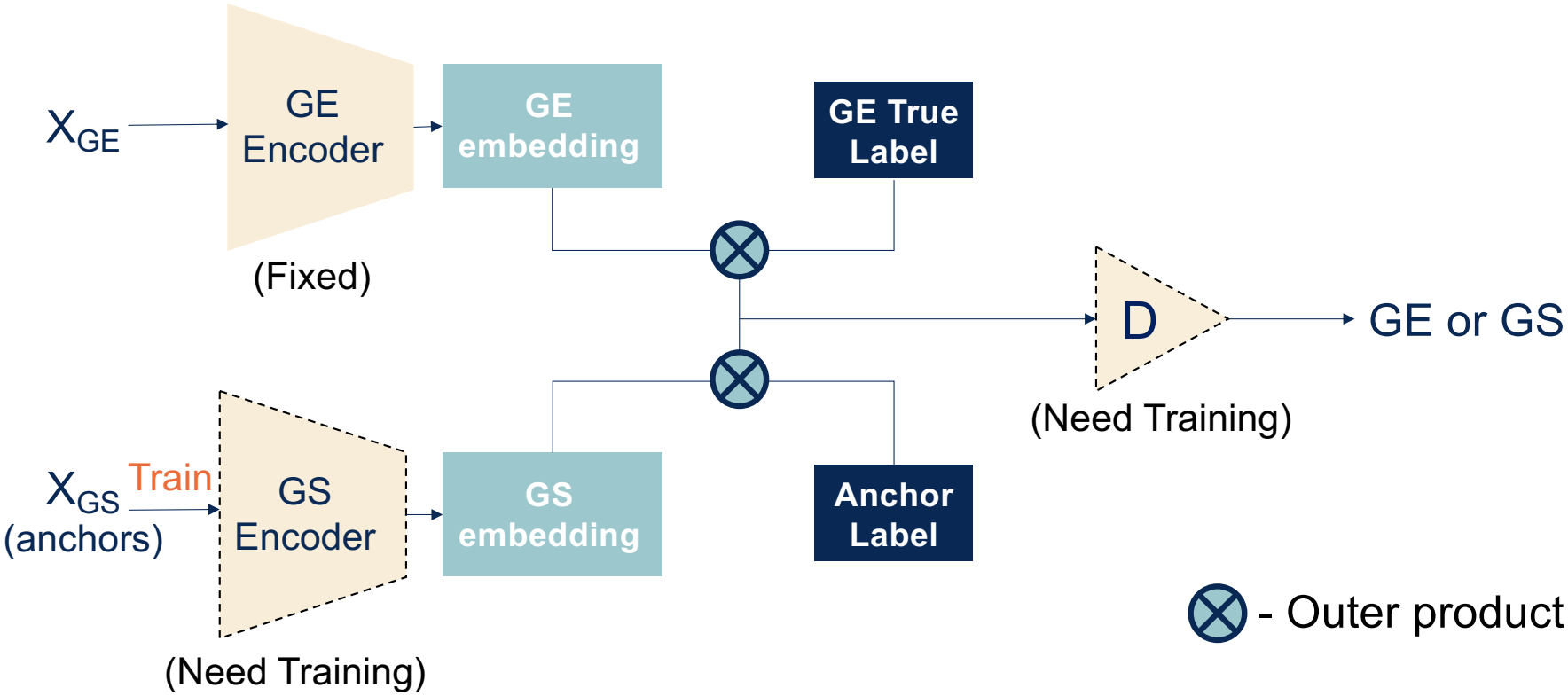


GE: gene expression; GS: gene scores from scATAC-seq

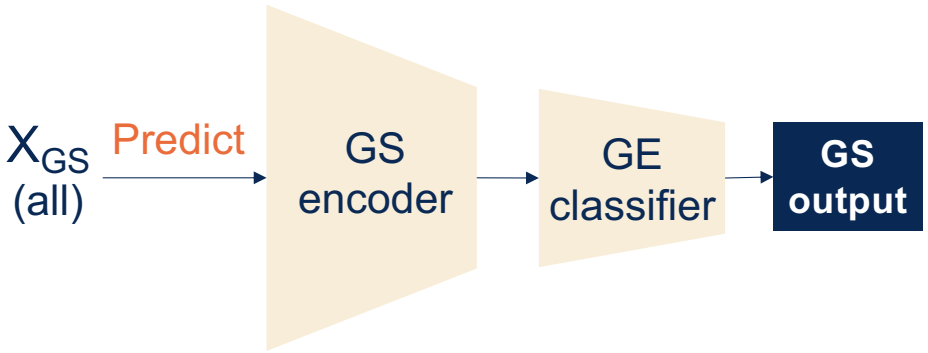
Step 2: Use Cellcano to predict on scATAC-seq and select anchors



Step 3: Use CellGAN to align embedding space on anchor cells

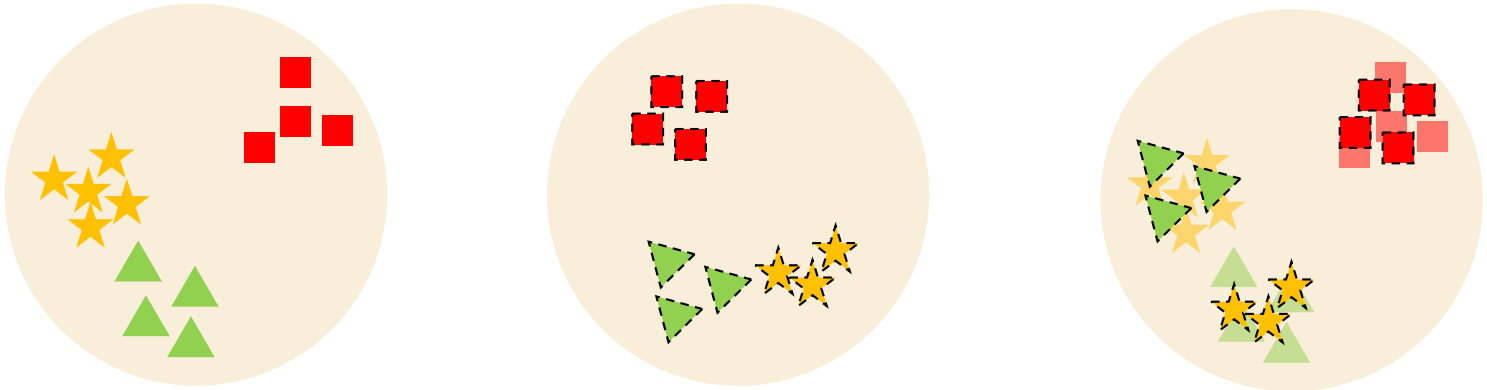


Step 4: Use CellGAN to predict all cells



Challenges

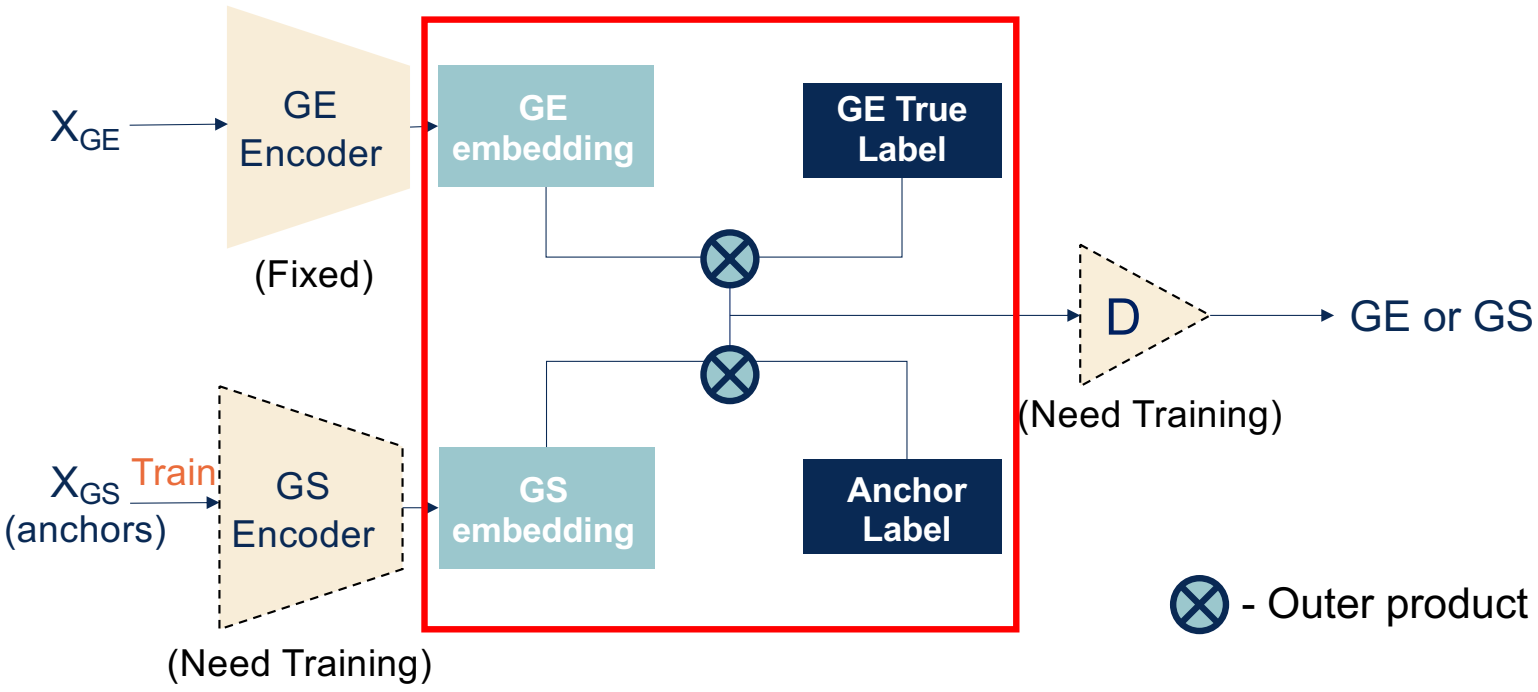
- GAN only matches the *marginal* distributions, not the *conditional* distribution.
  - Cell types can be misaligned, e.g., CD4 in scATAC-seq is aligned with CD8 in scRNA-seq.



Reference $P(X_R)$ Target $P(X_T)$ $P(X_R) \approx P(X_T)$

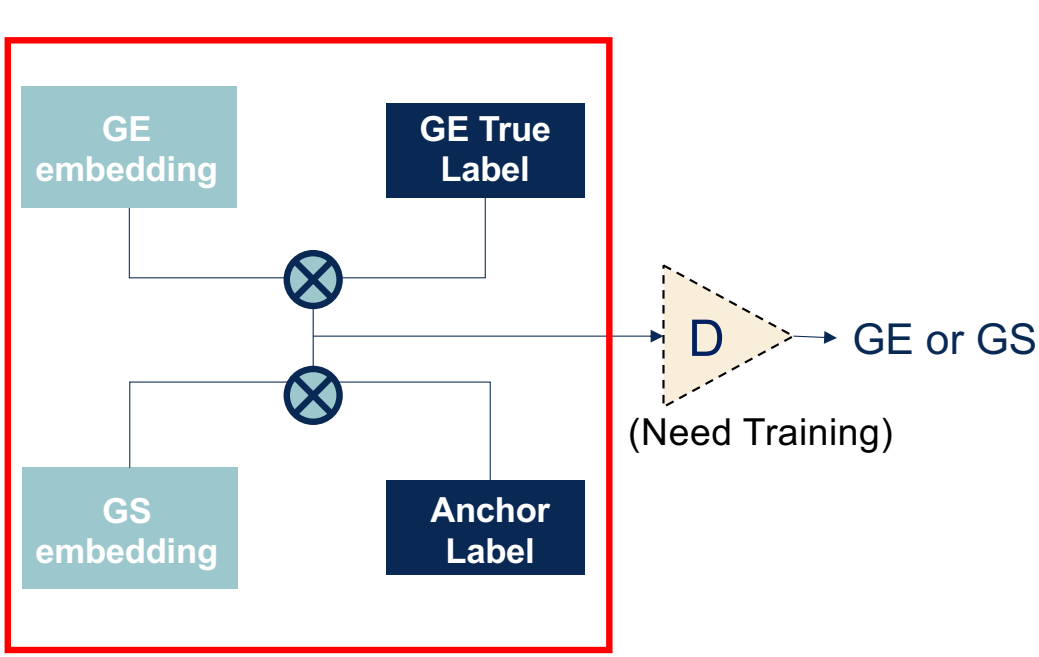
Conditional GAN

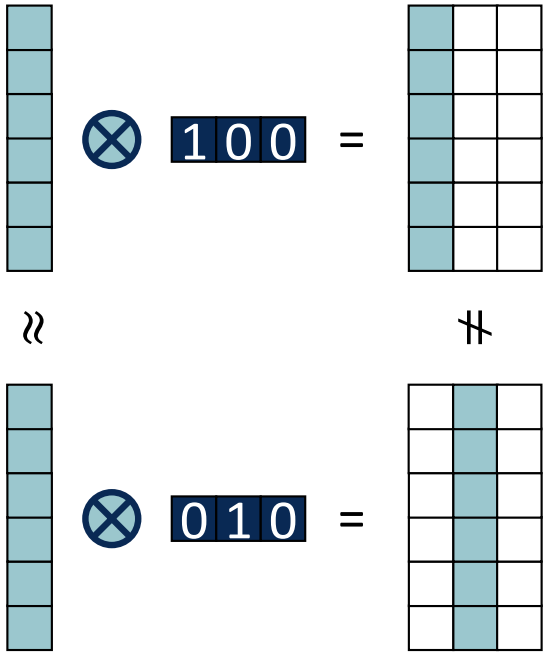
- We use outer product to let GAN match the conditional distribution instead of the marginal distribution



Conditional GAN

The embedding is a N vector and the label vector is a one-hot encoded vector

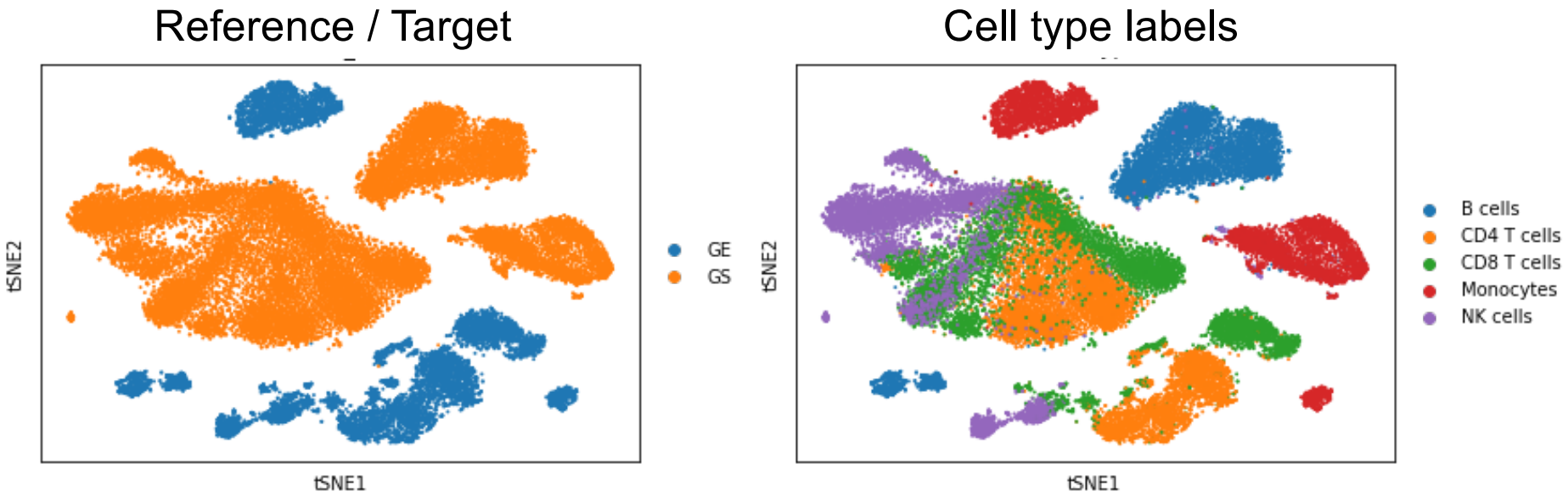




By doing the outer product, we can accurately match the conditional distribution.

GAN vs Conditional GAN

- Reference: 10X human PBMCs scRNA-seq
- Target: human PBMCs FACS-sorted scATAC-seq (Lareau et al.)

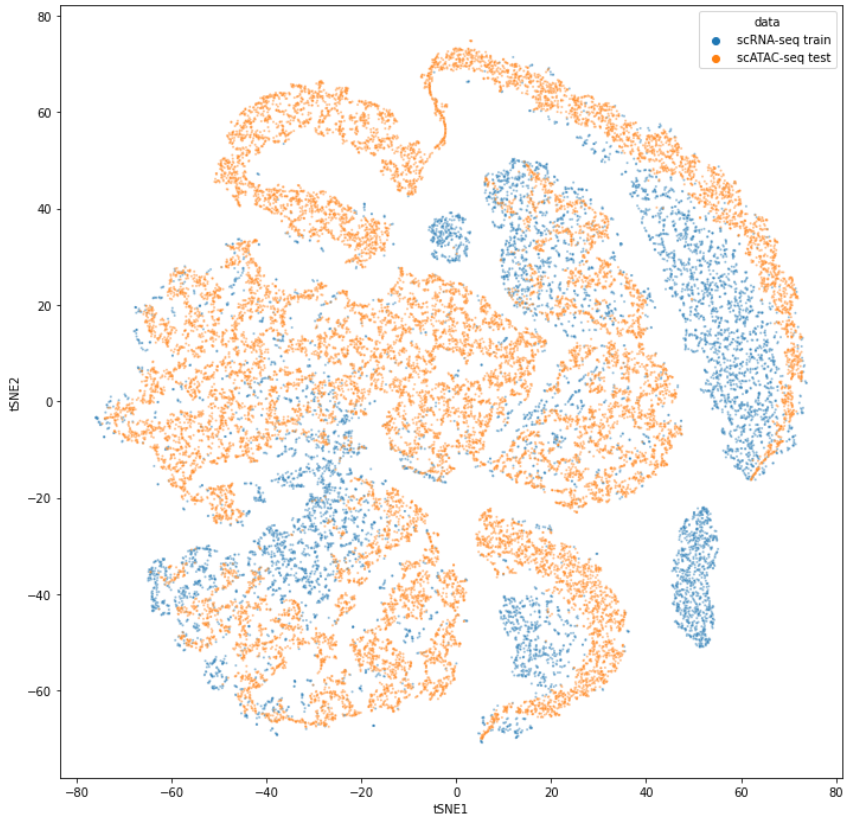


Lareau *et al*. *Nat Biotechnol* (2019).

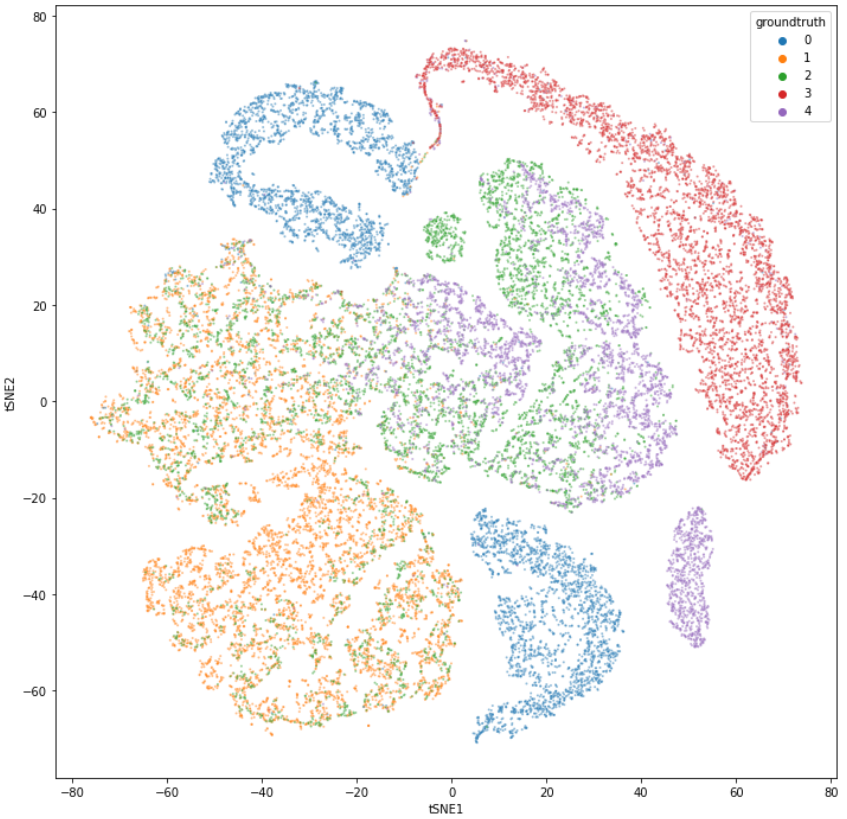
GAN vs Conditional GAN

Acc: 0.638
macroF1: 0.593

Reference / Target



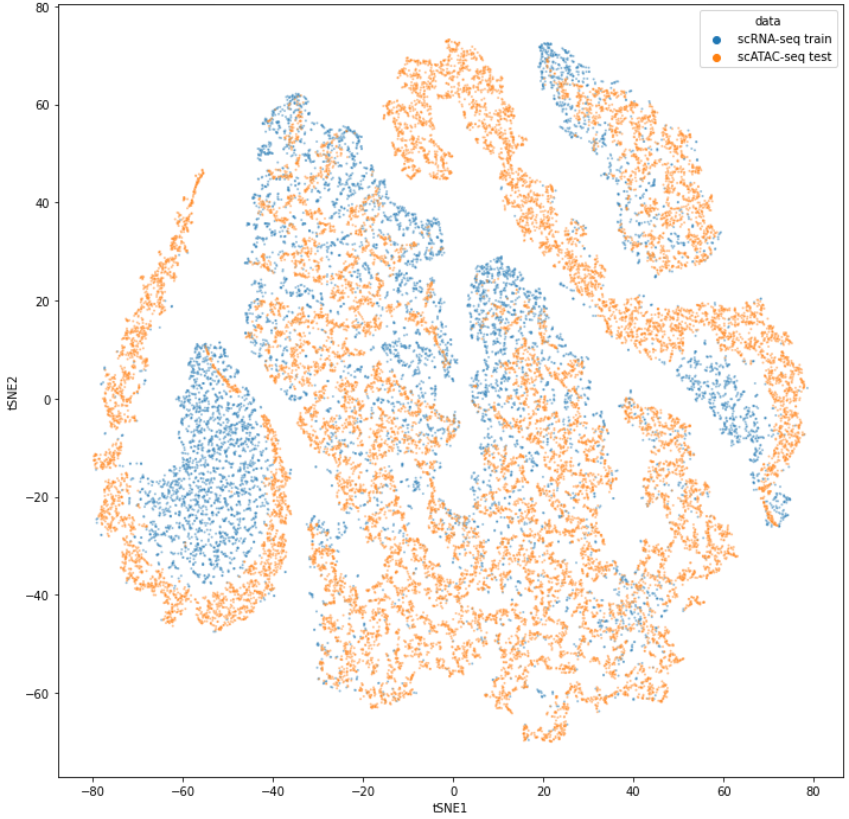
Cell type labels



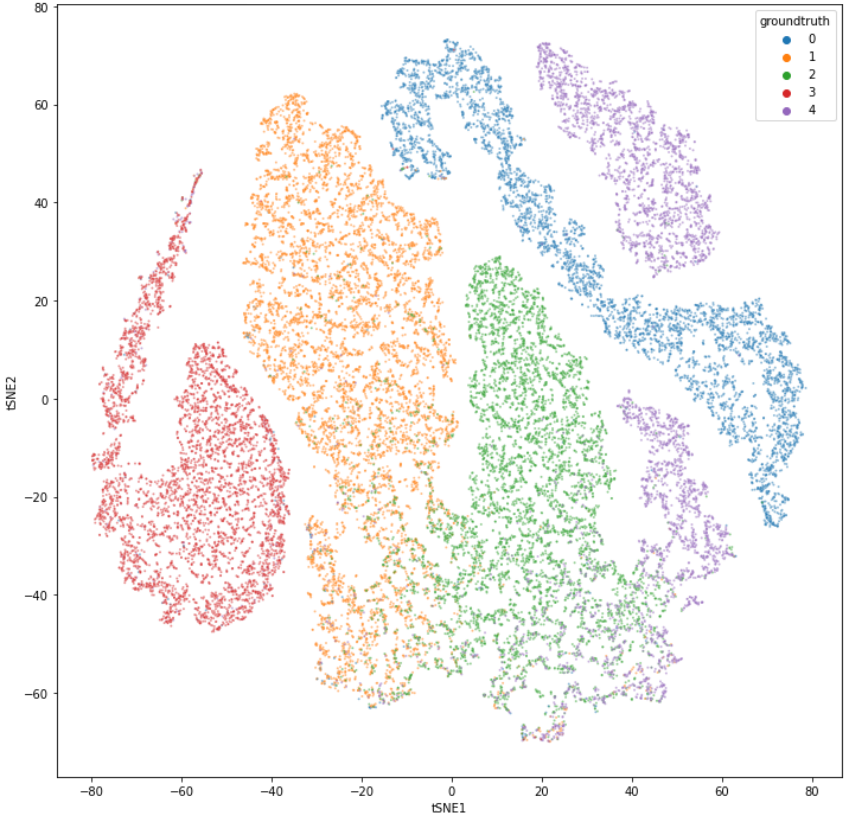
GAN vs Conditional GAN

Acc: 0.871
macroF1: 0.876

Reference / Target



Cell type labels



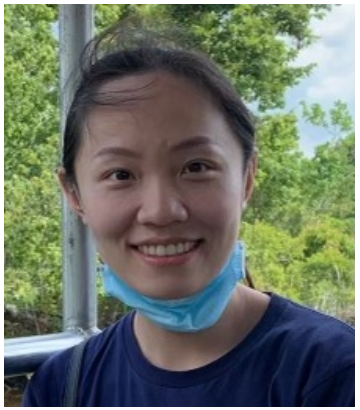
More results

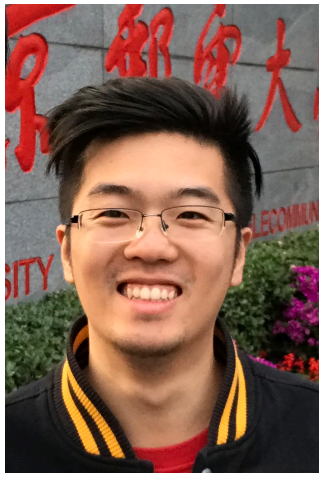
Coming soon …

Summary

- scATAC-seq data is more challenging to analyze than scRNA-seq
  - Weaker signals
  - No well-defined feature
- We develop method for scATAC-seq celltyping
- To overcome domain shift, we develop
  - Two-round procedure
  - GAN for data harmonization
- Software Cellcano: https://marvinquiet.github.io/Cellcano/

Acknowledgement





Wenjing Ma
Emory CS

Jiaying Lu
Emory CS

NIH
National Institutes of Health

R01GM122083, P01NS097206