scBERT as a large-scale pretrained deep language model for cell type annotation of single-cell RNA-seq data

Caleb Ellington

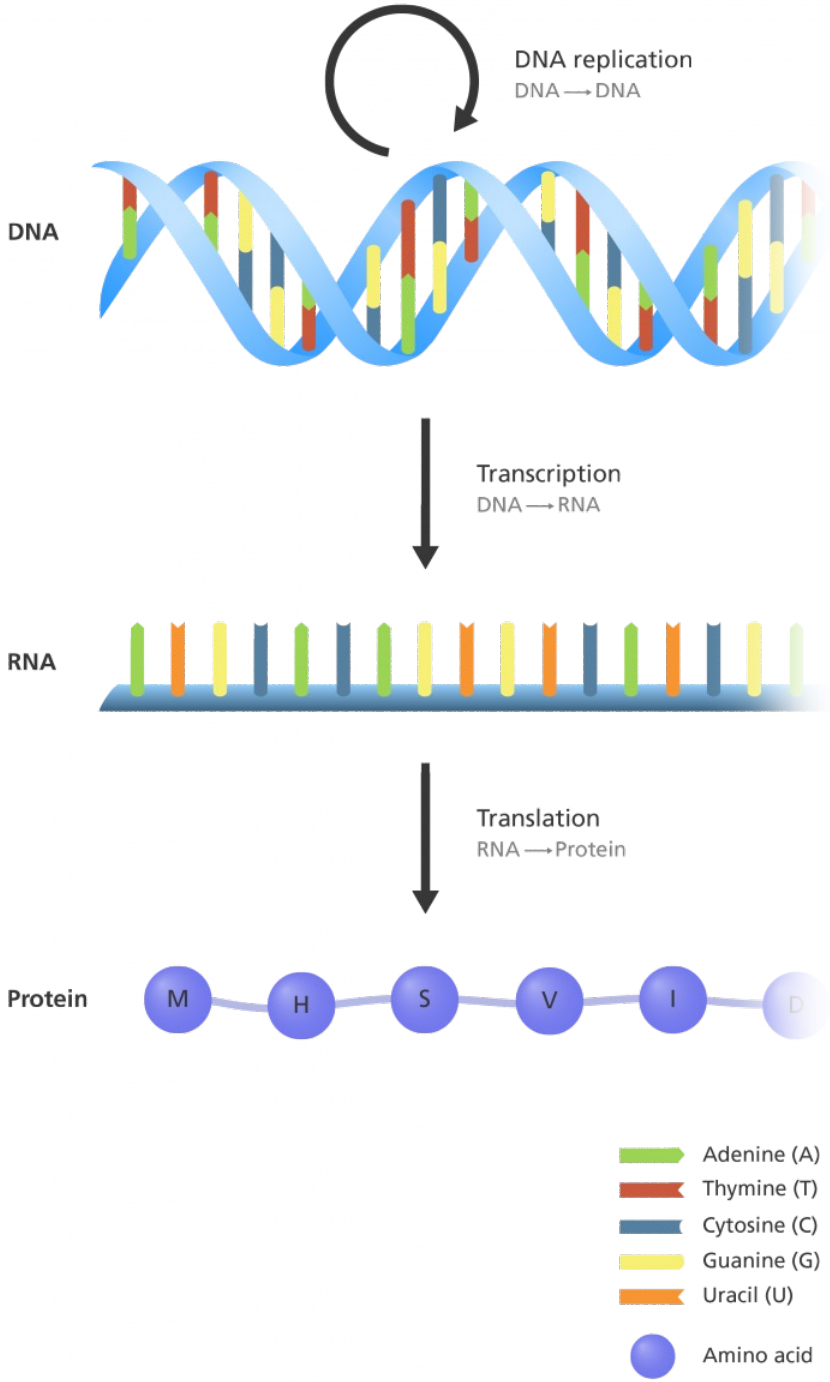
AI4Bio

Dec. 15th, 2022

Motivation: characterizing cell (pheno)types to study functions and relation to disease & treatment

Example: A high number of lymphocytes is seen as a symptom of a bacterial infection. However, a high neutrophil-to-lymphocyte ratio (low lymphocytes) in COVID patient indicates poor prognosis. How do we characterize these cells (e.g. define a lymphocyte) and understand their role (e.g. how do these cell types interact)?

Measuring Phenotypes

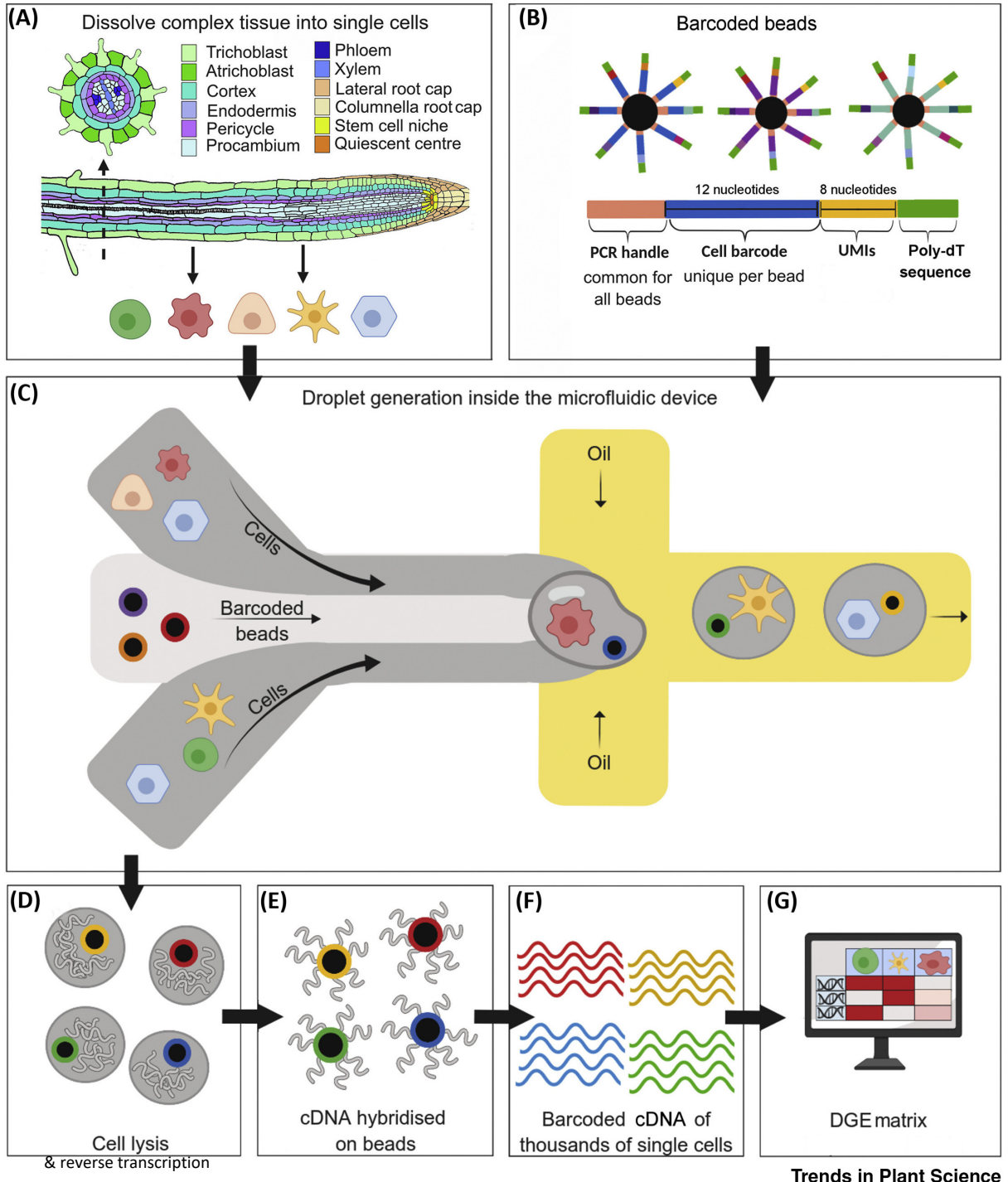


Easy to measure, but does not represent dynamic cellular systems

Dynamic & easy to measure!

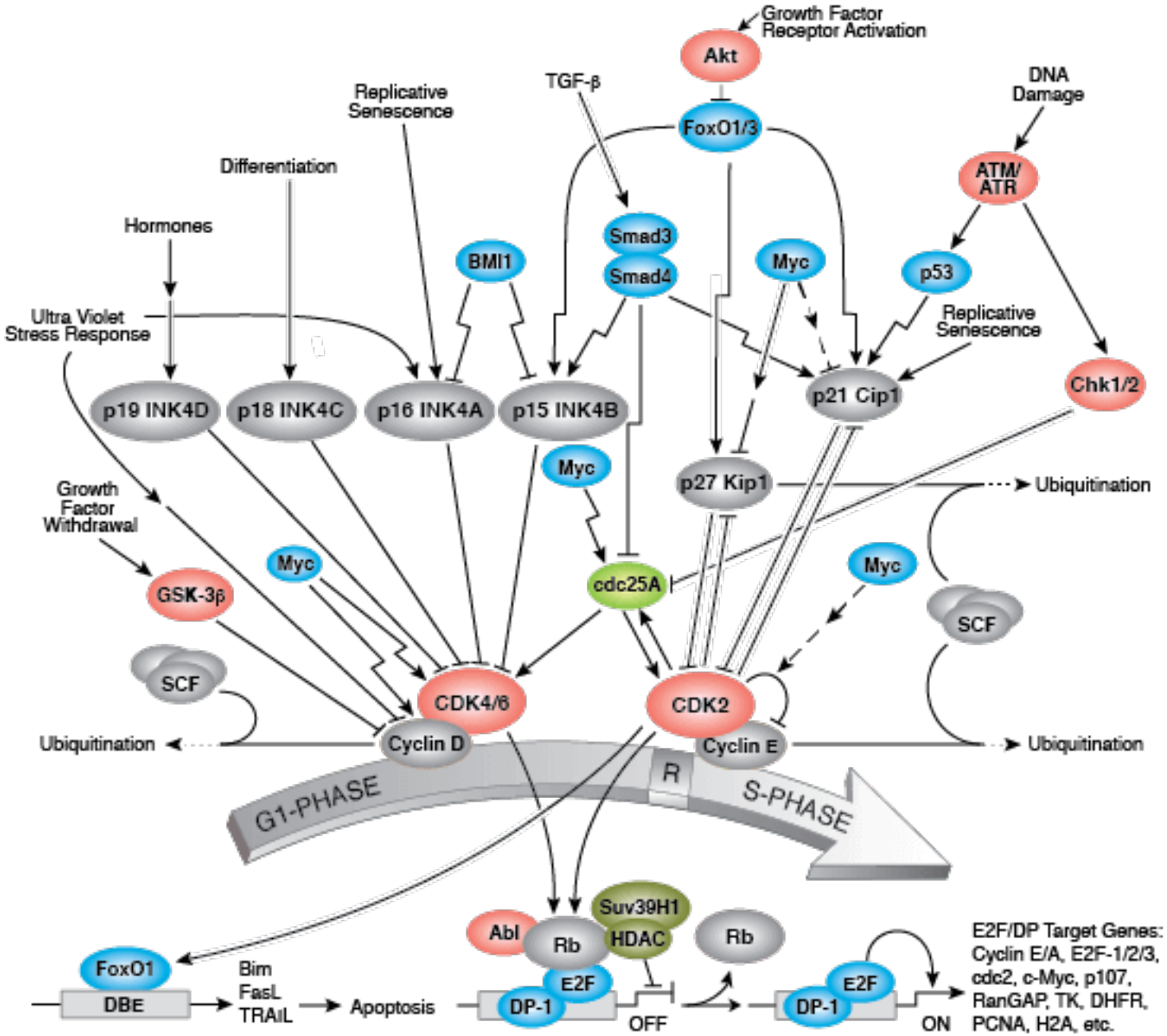
Dynamic and directly related to phenotypes, but incredibly difficult to measure

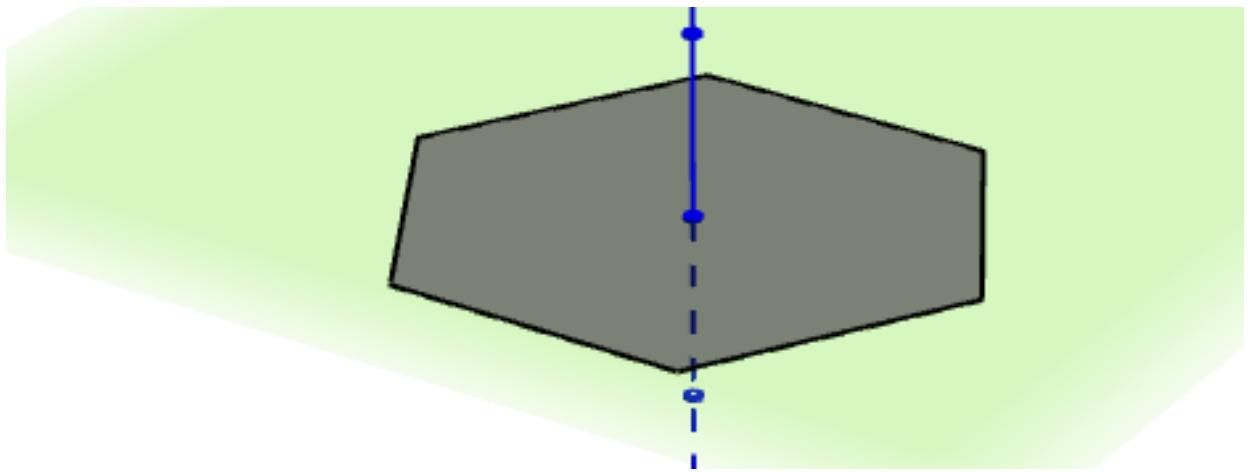
Single-cell RNA Sequencing gives an expression reading for each cell

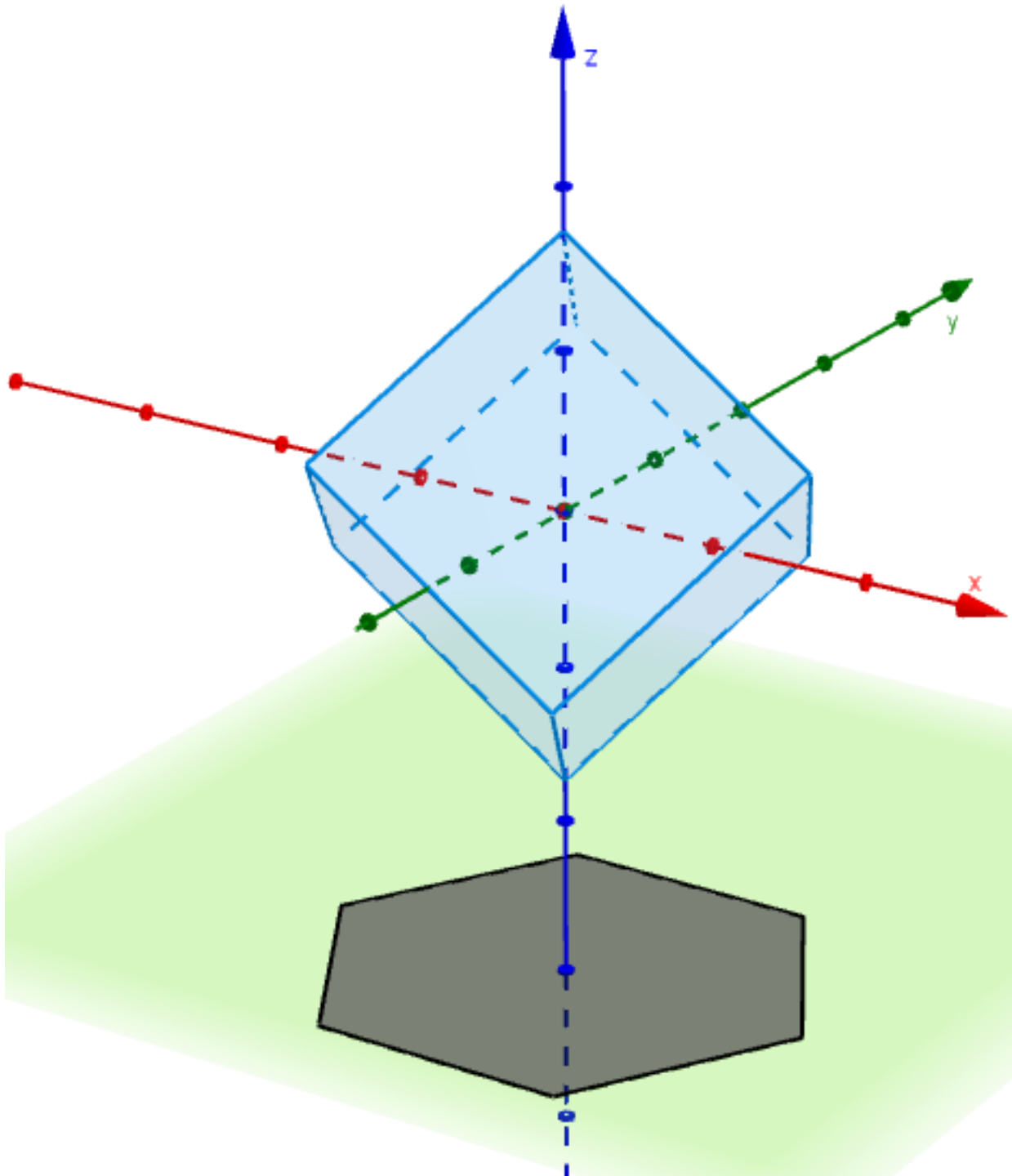


Trends in Plant Science

Functional biological systems are dynamic, stochastic, and heterogeneous. Gene expression is an incomplete, noisy picture of cell state.






z
y
x

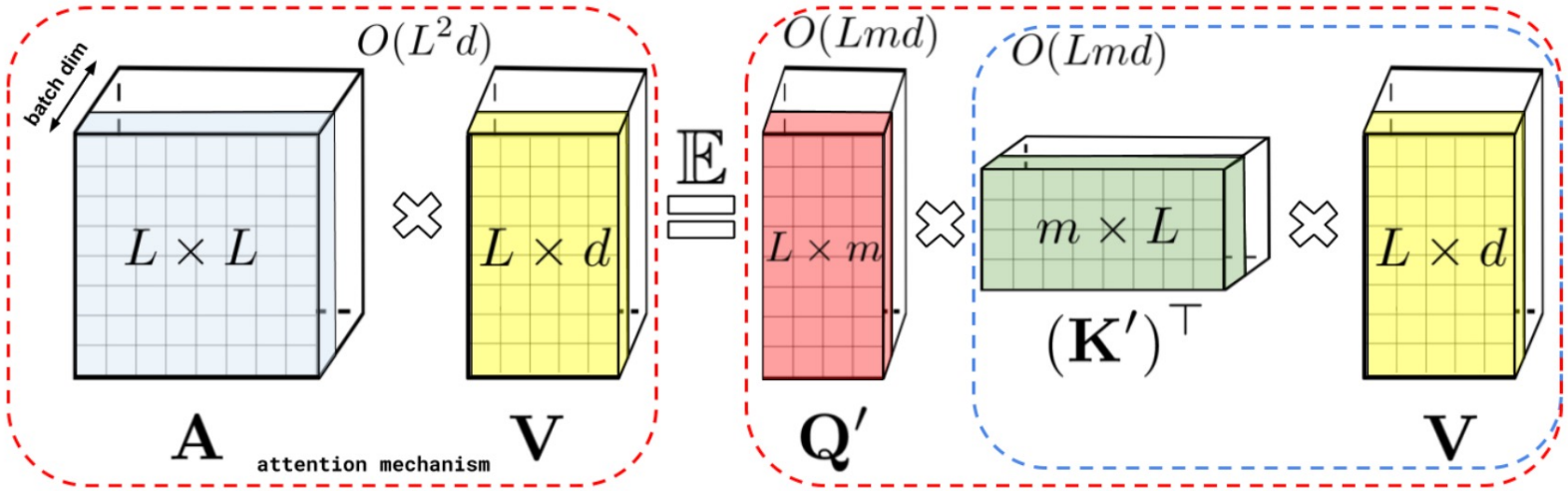
Task: Use noisy time-point expression data to characterize cell types with dynamic cell states

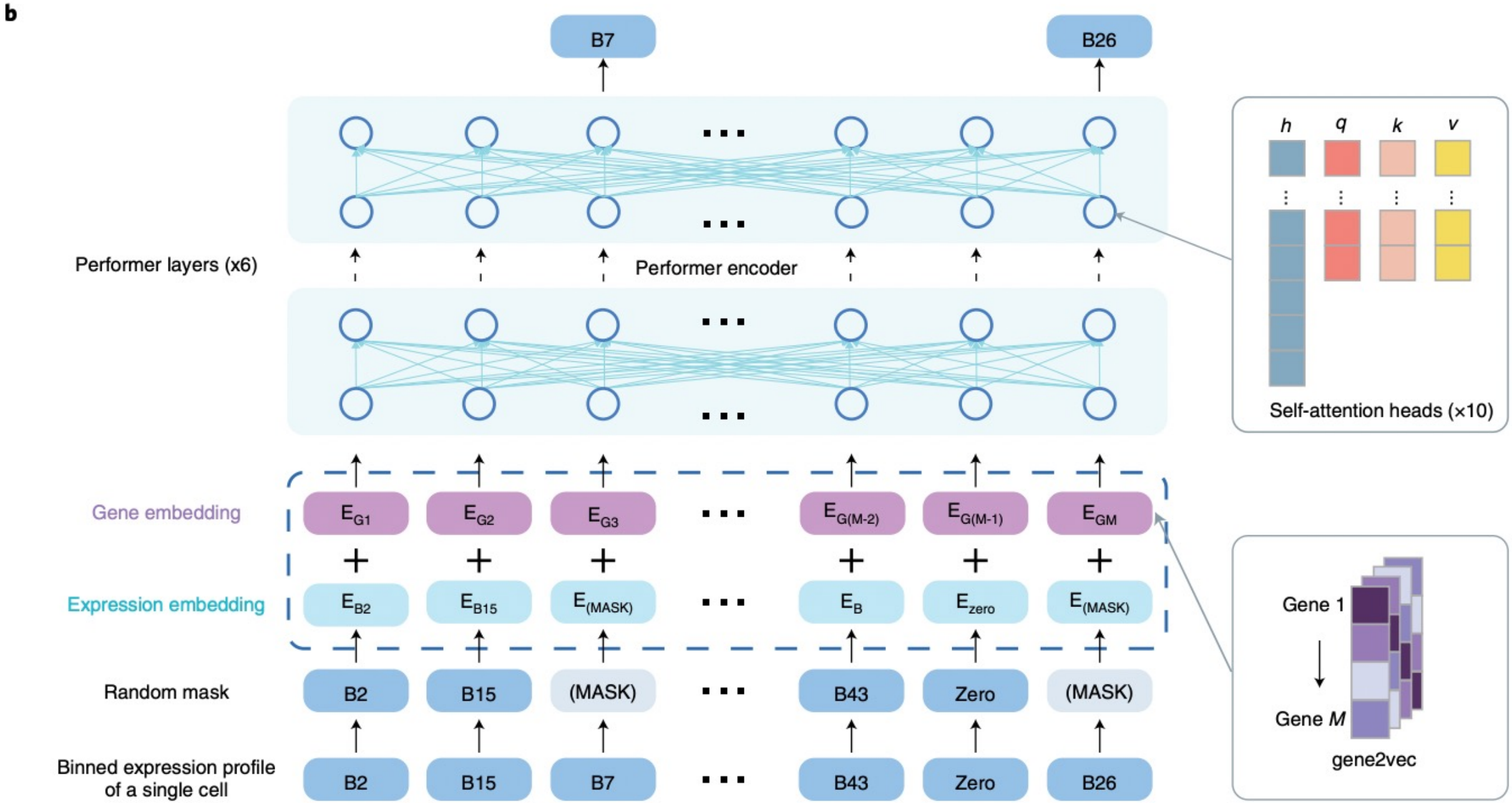
Baselines

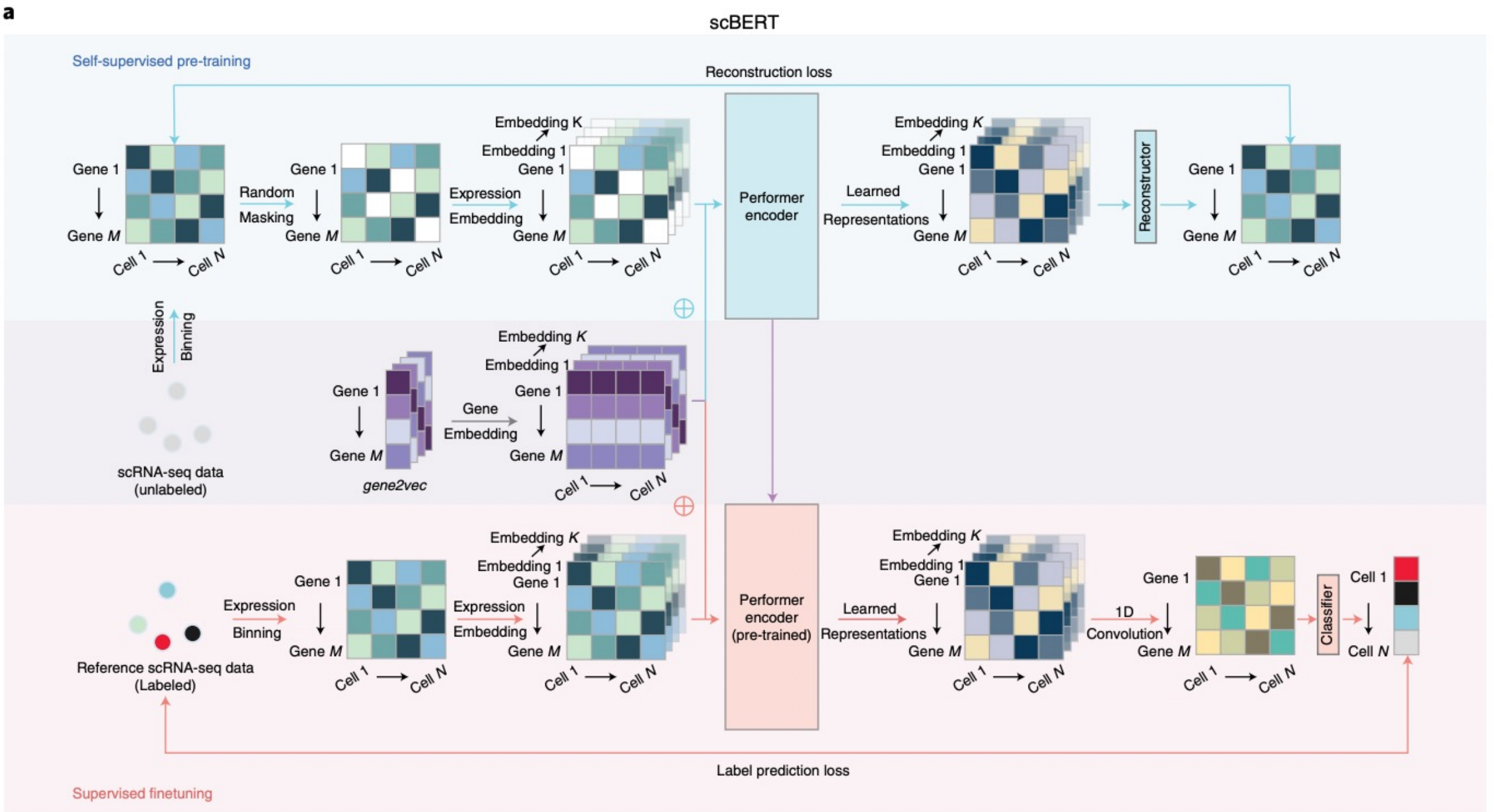
- Marker-based
  - Checks for marker genes associated with high expression in each cell type (like checking a fact bank)
  - SCINA, Garnett, scSorter
- Correlation-based
  - Check a sample against reference data, assign the cell type with the highest correlation
  - Seurat, SingleR, CellID, scmap
- Supervised classification
  - Build model on labeled data, apply to label a test set
  - scNym, SciBet (SOTA)

scBERT Preliminaries

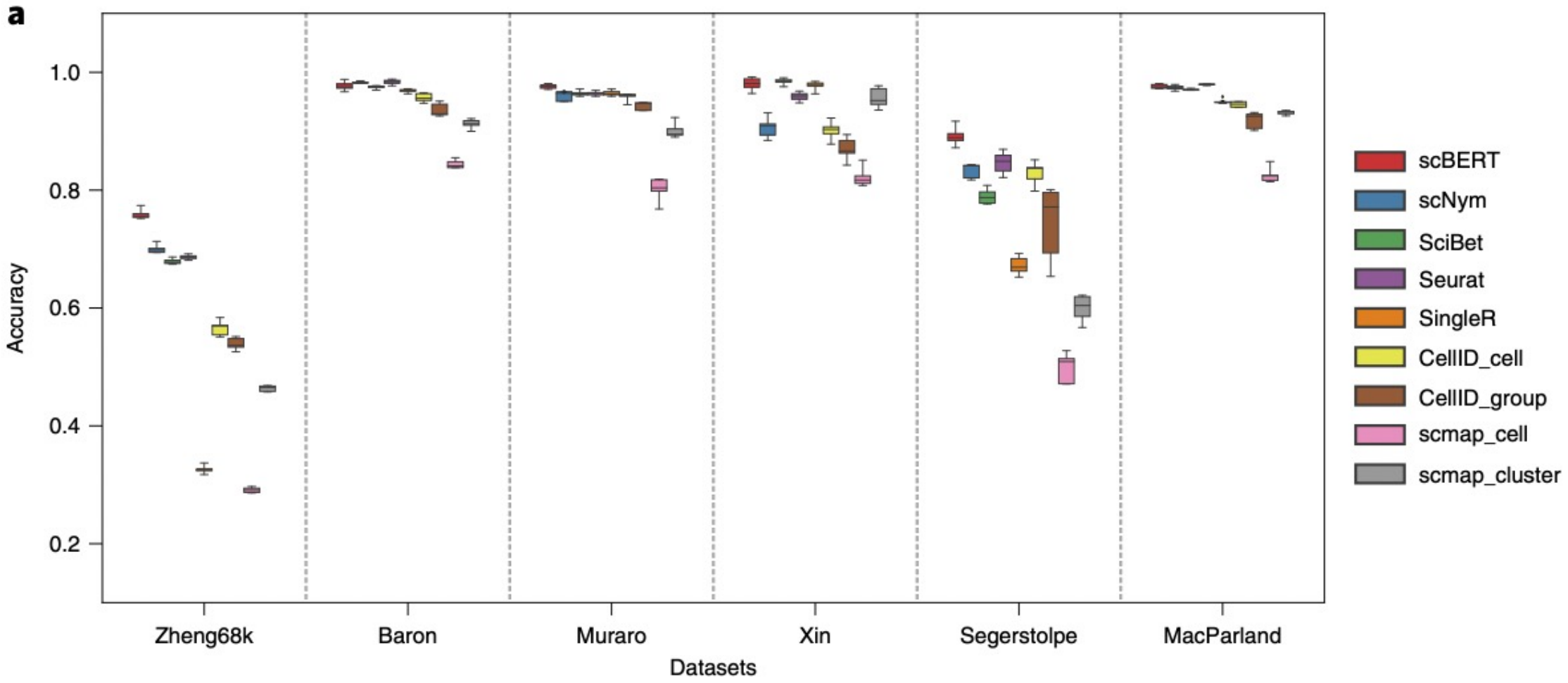
- Gene2vec:
  - Word2vec: co-occurring words {0, 1} have similar embeddings
  - Gene2vec: co-expressed genes [0, 1] have similar embeddings
- Performer
  - Transformers with low-rank attention for linear sequence-length scaling



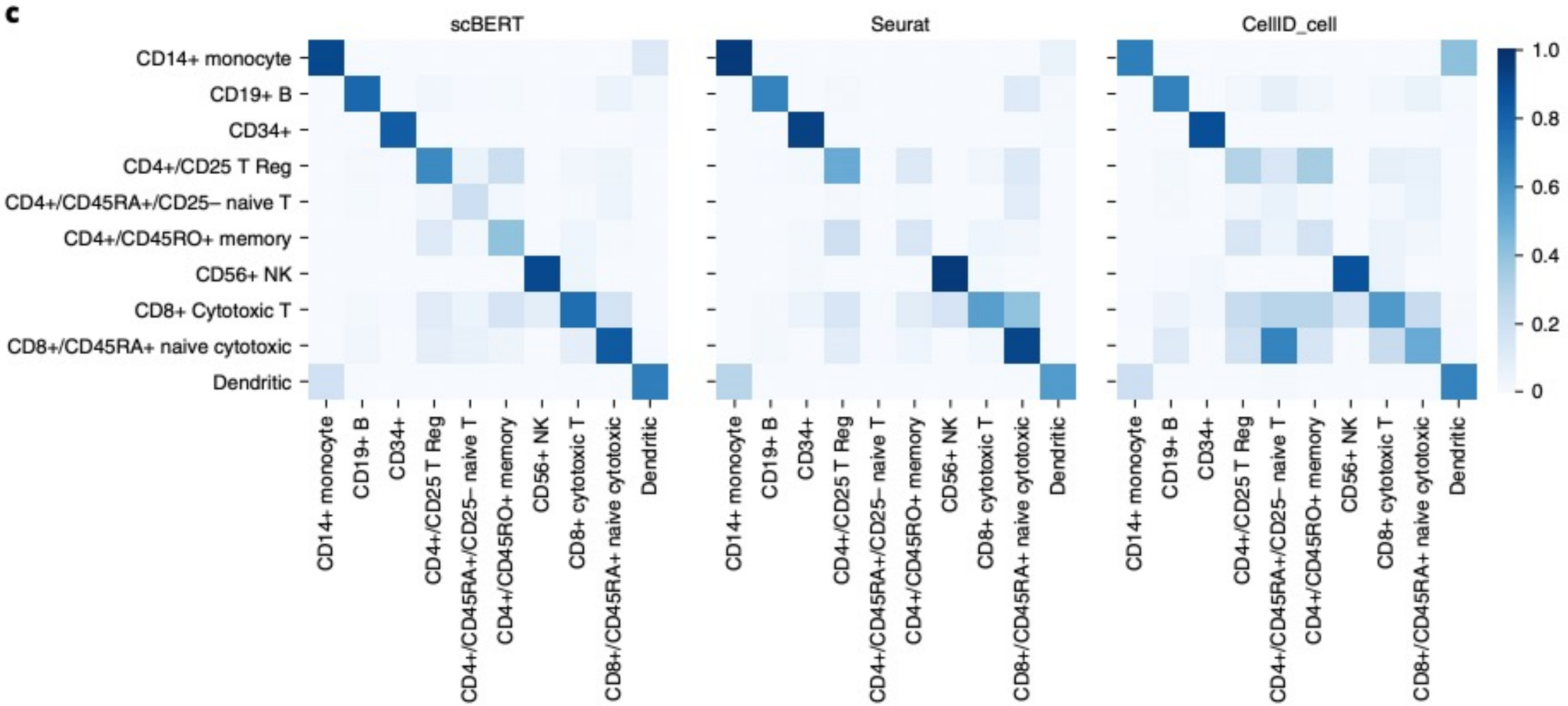

b
B7
B26
Performer layers (x6)
Performer encoder
h
q
k
v
Self-attention heads (×10)
Gene embedding
Expression embedding
Random mask
Binned expression profile of a single cell
E_G1
E_G2
E_G3
E_G(M-2)
E_G(M-1)
E_GM
E_B2
E_B15
E_(MASK)
E_B
E_zero
E_(MASK)
B2
B15
(MASK)
B43
Zero
(MASK)
B2
B15
B7
B43
Zero
B26
Gene 1
Gene M
gene2vec


a
scBERT
Self-supervised pre-training
Reconstruction loss
Gene 1
Gene M
Cell 1
Cell N
Random
Masking
Expression
Embedding
Embedding K
Embedding 1
Performer
encoder
Learned
Representations
Reconstructor
Expression
Binning
scRNA-seq data
(unlabeled)
gene2vec
Gene
Embedding
Reference scRNA-seq data
(Labeled)
Performer
encoder
(pre-trained)
1D
Convolution
Classifier
Label prediction loss
Supervised finetuning

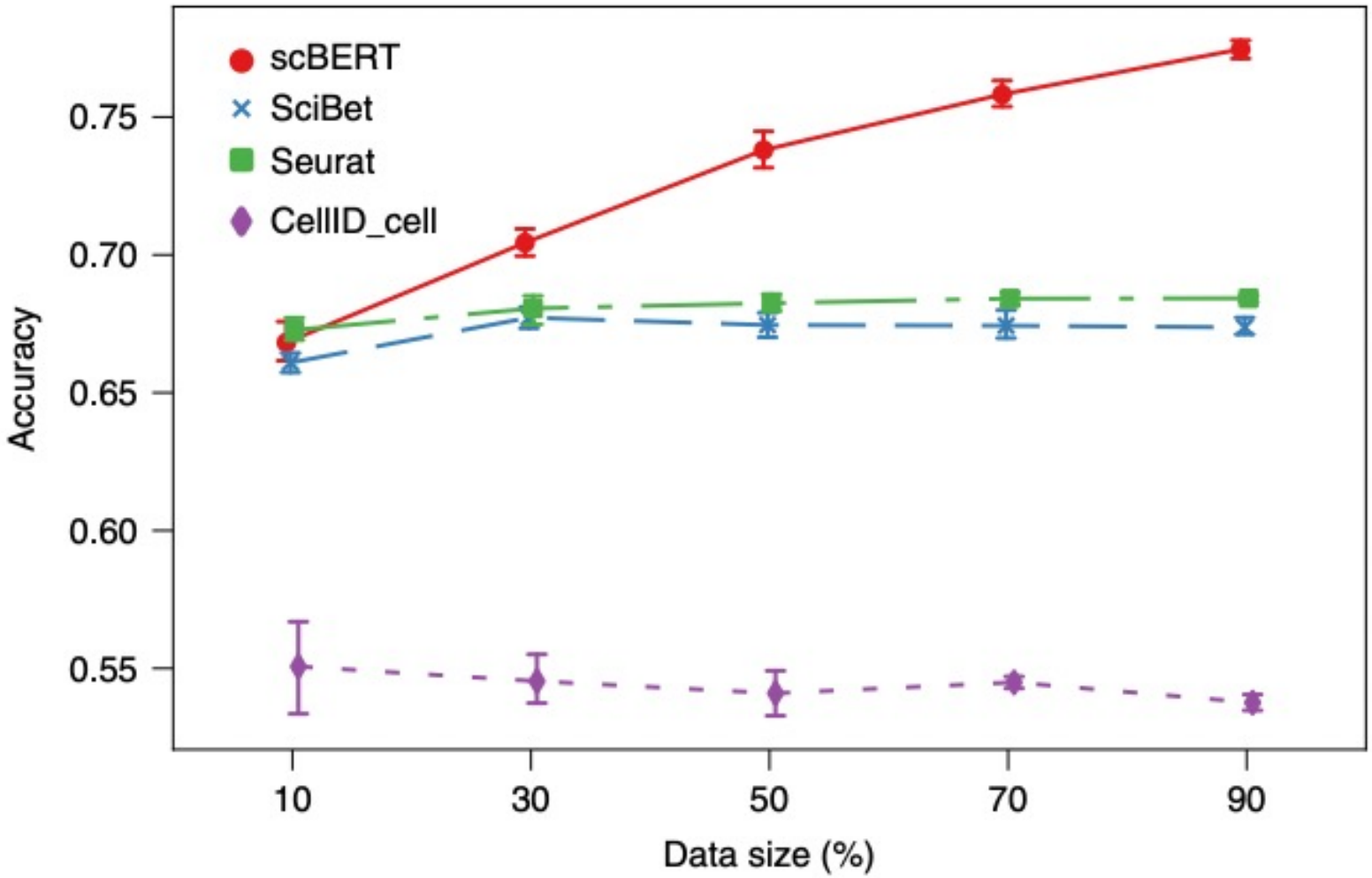
scBERT improves recovery of expert celltype annotations



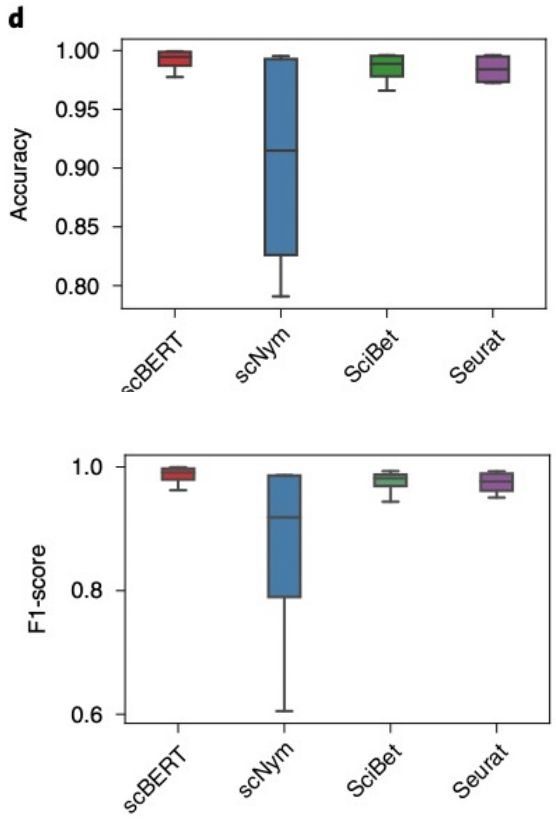
scBERT is less confused about cell types

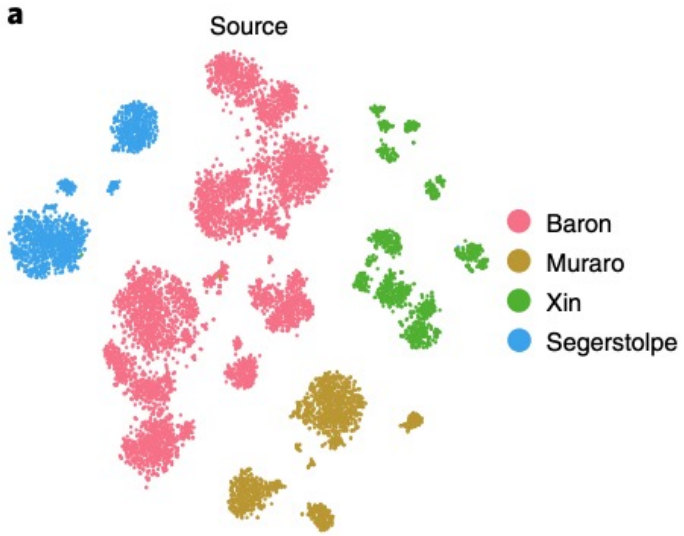


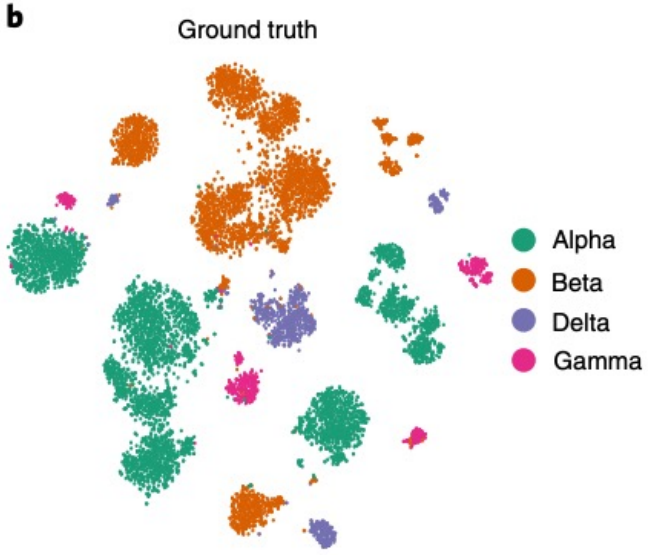
scBERT accuracy scales with dataset size

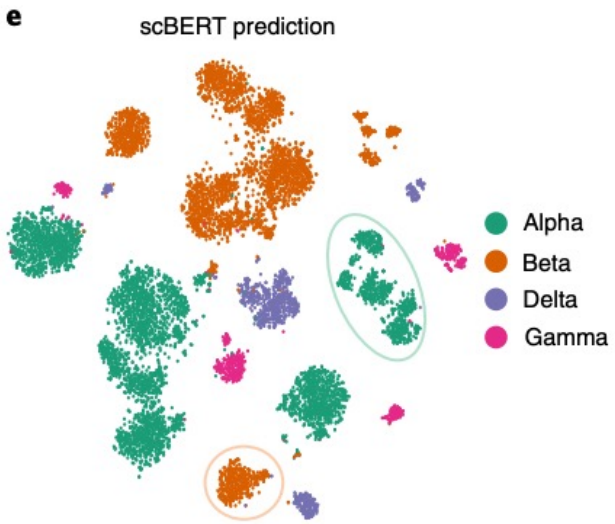


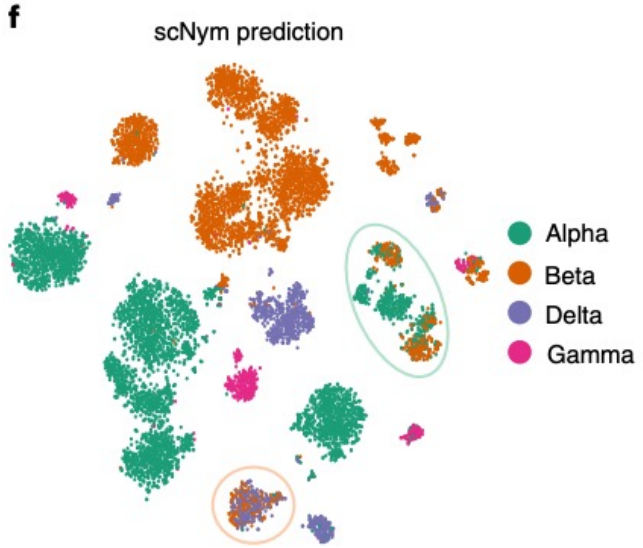
Batch effects do not affect cell type prediction in scBERT



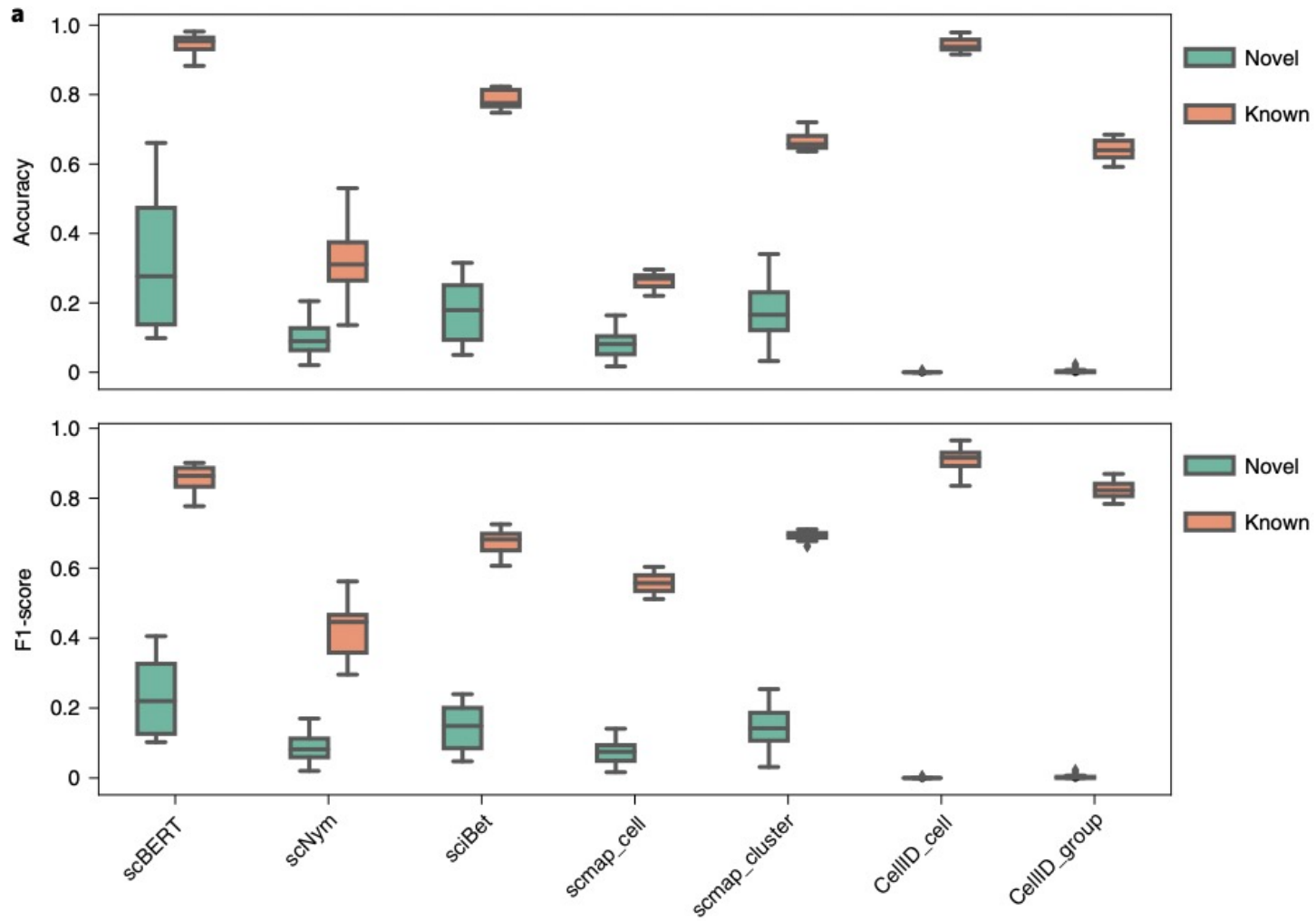




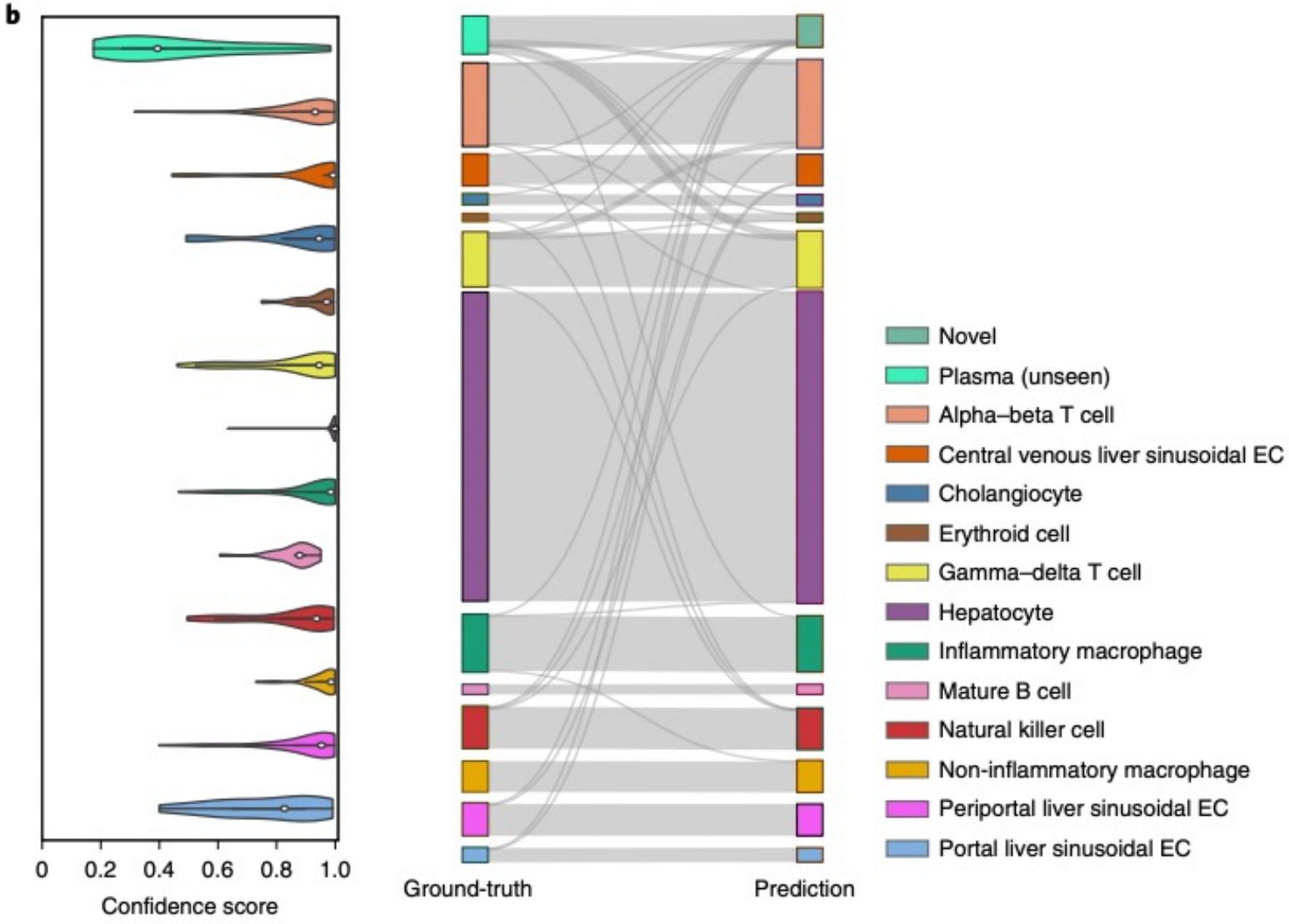




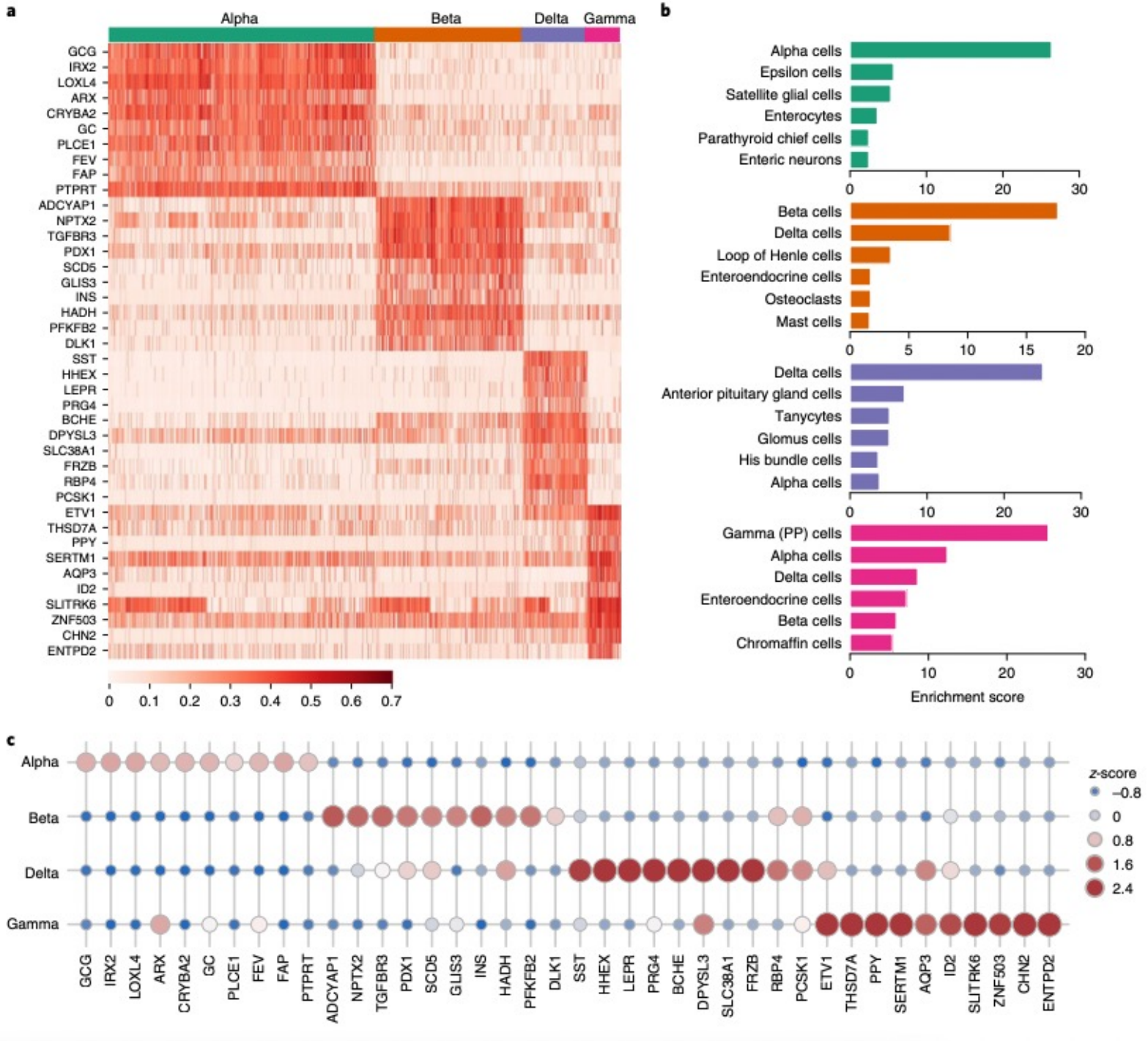
Improved cell state representation leads to improved discovery of novel (held-out) cell types



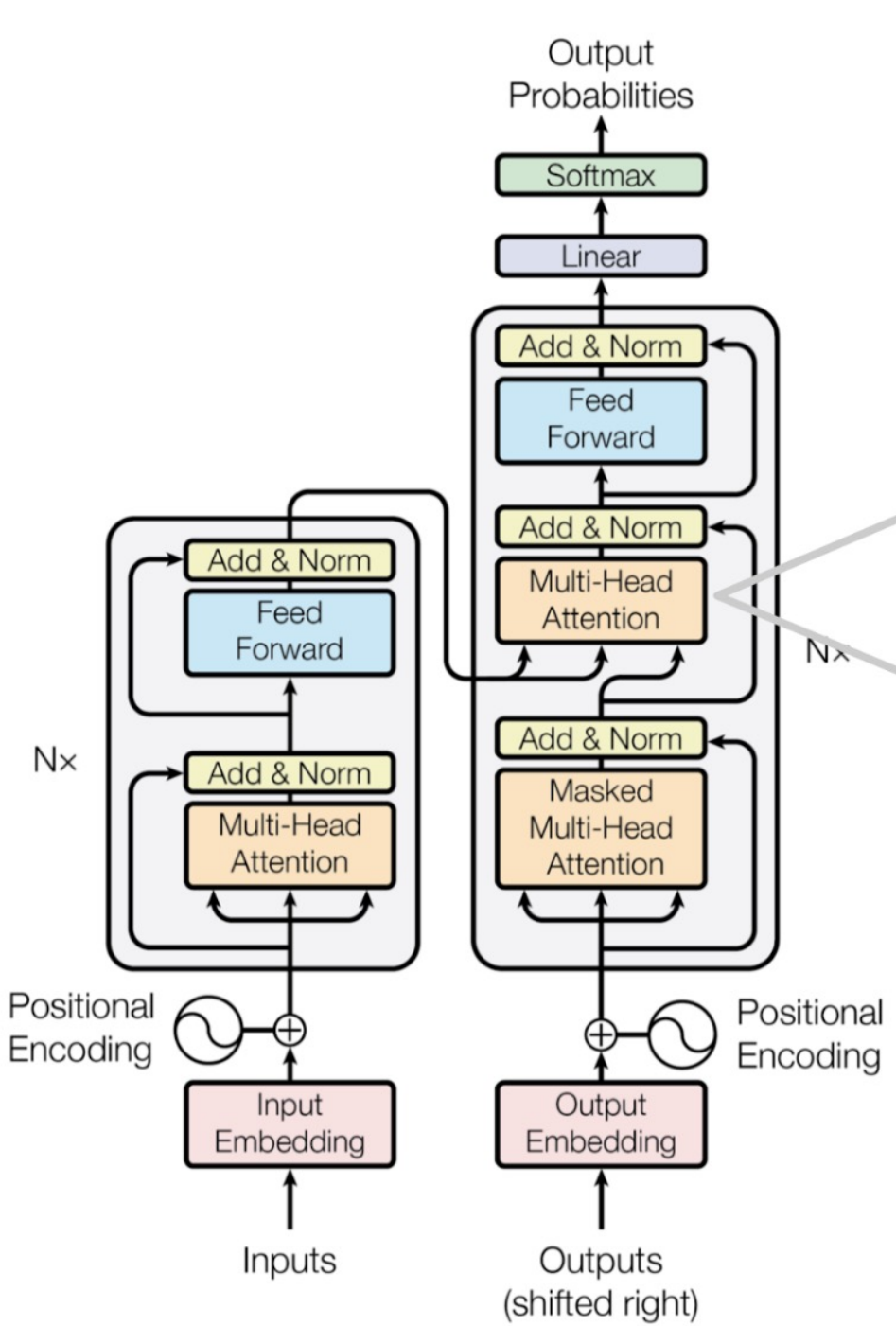
scBERT is less confused about novel cell types

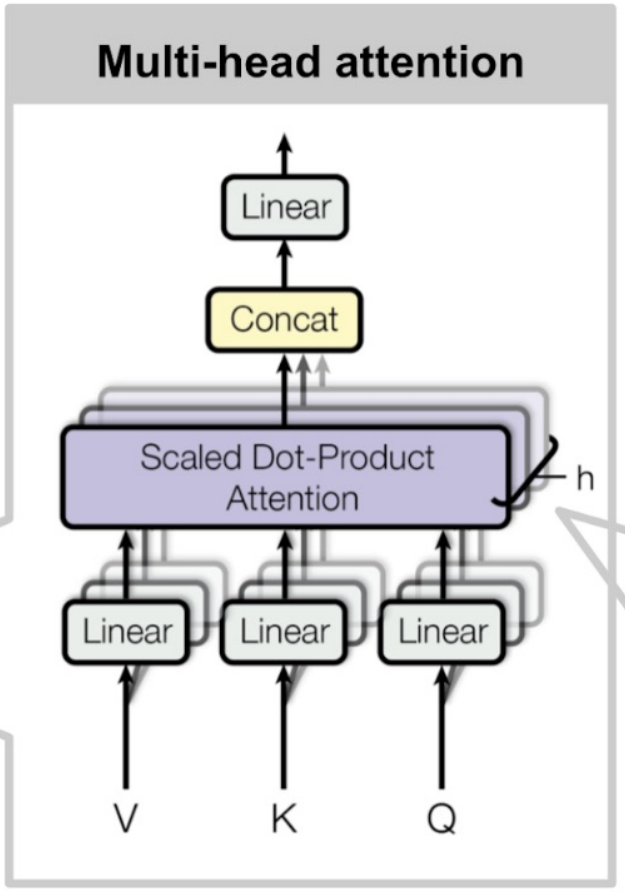


Gene-averaged attention weights parallel the specificity of genes to each cell-type



Extra Info

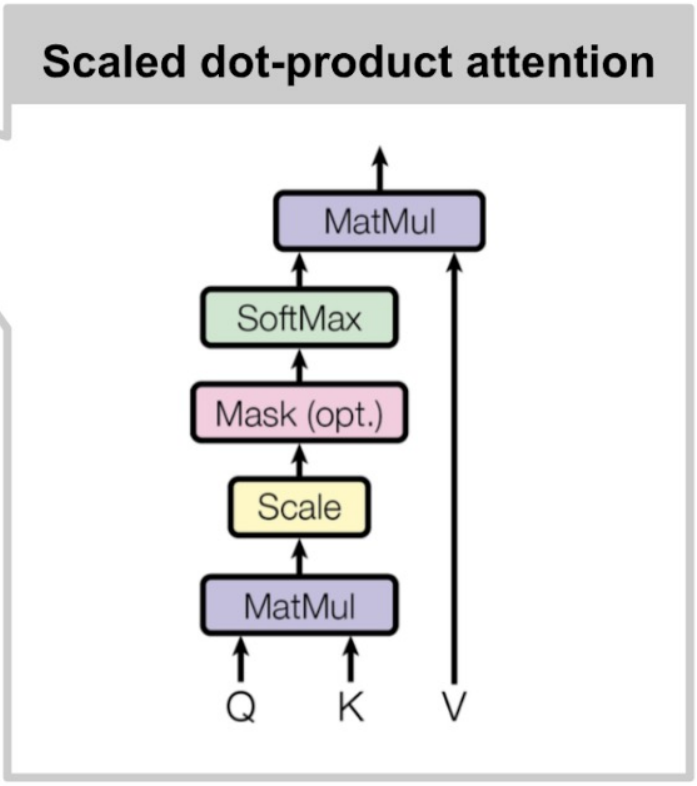




Zoom-In!

10 attention heads

Performer replaces this



Zoom-In!

Model interpretability

We conducted a comprehensive interpretability analysis to explore the key genes for decision-making, as scBERT models were built on the self-attention mechanism and all of the genes' representations remained at the end of our workflow. The attention weights reflect the contribution of each gene and the interaction of gene pairs. The attention weights can be obtained from equation (1), modified by replacing V with V^0, where V^0 contains one-hot indicators for each position index. We integrated all the attention matrices into one matrix by taking an element-wise average across all attention matrices in multi-head multi-layer Performers. In this average attention matrix, each value $A(i,j)$ represented how much attention from gene i was paid to gene j. To focus on the importance of genes to each cell, we summed the attention matrix along with columns into an attention-sum vector, and its length is equal to the number of genes. In this way, we could obtain the top attention genes corresponding to a specific cell type compared to other cell types. The attention weights were visualized and the top genes were sent to Enrichr[32] for enrichment analysis.