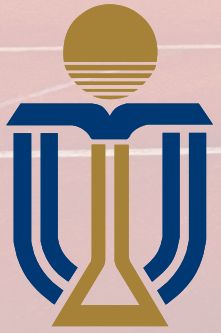
# Foundation Models on Single Cell RNA Sequencing Data

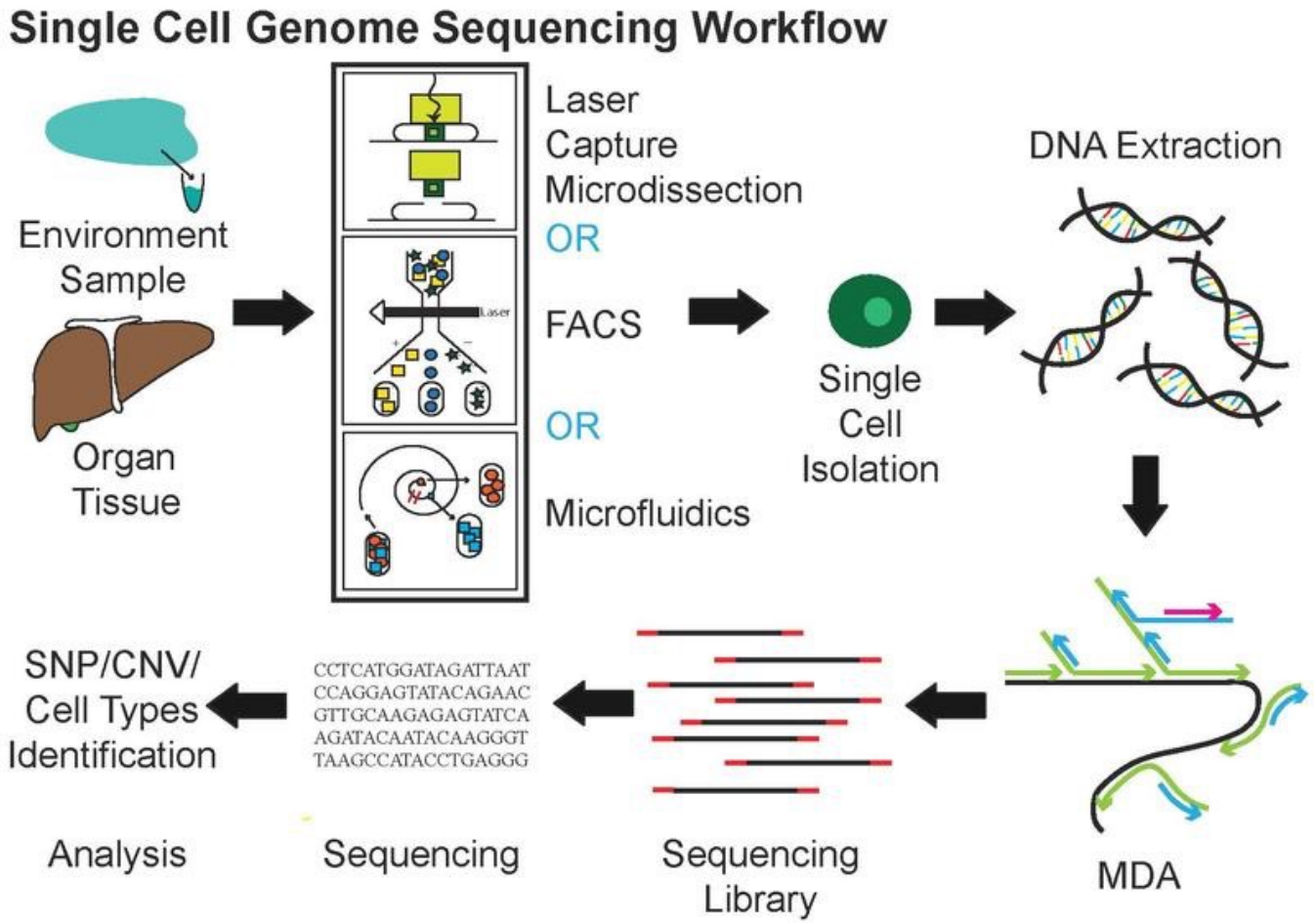
09/09/2023

Minghao WANG



# scRNA-seq

scRNA-seq **provides information on which genes are expressed** at the level of individual cells.




- Explore which cell types are present in tissues
- Identify unknown/rare cell types or states
- Elucidate gene expression changes during differentiation or across time or across states
- …

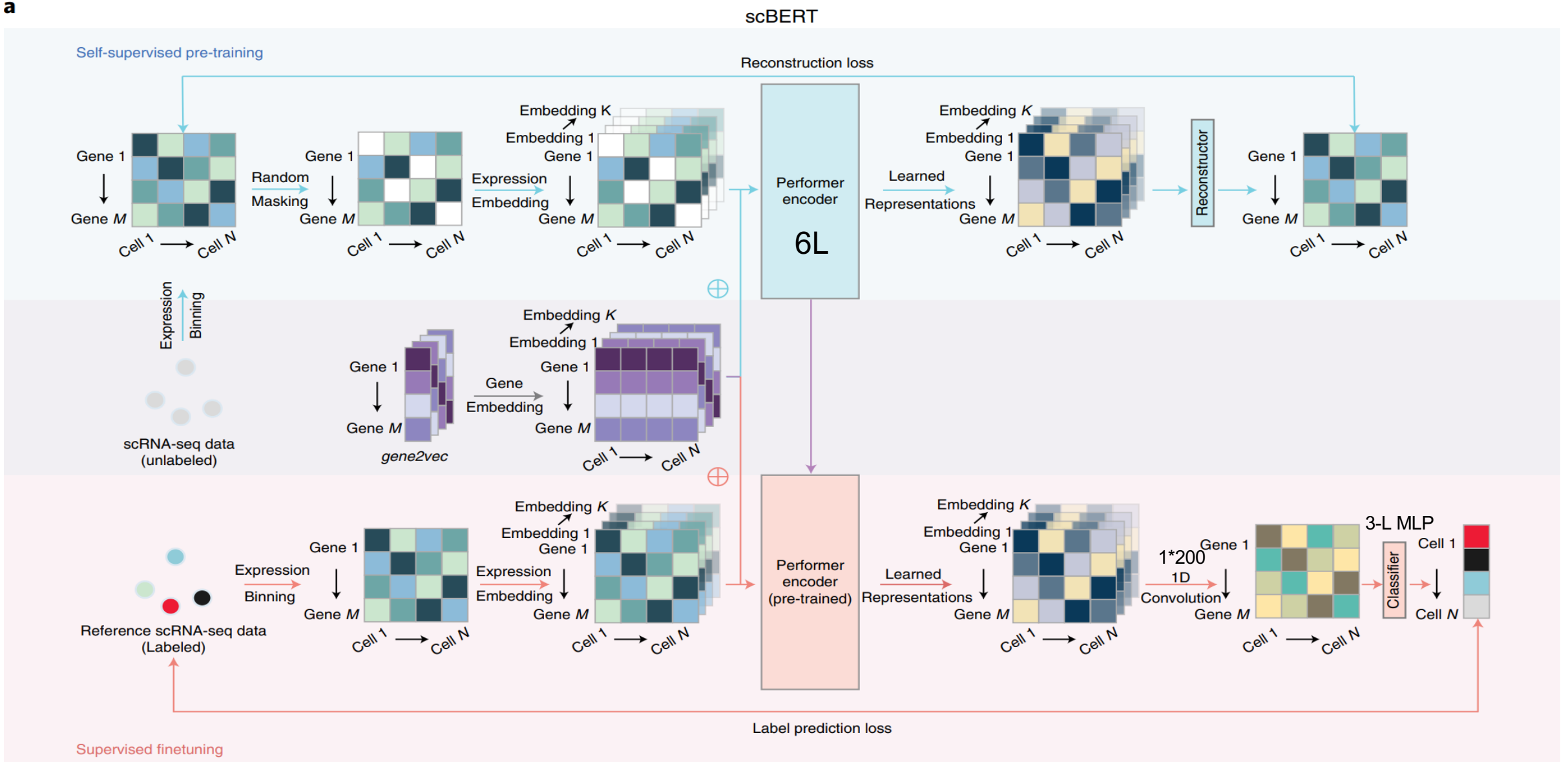
Data

| | AC145205.1 | ... | RP11-34P13.8 | RP11-34P13.7 | OR4F5 | FAM138A | MIR1302-10 |
|---|---|---|---|---|---|---|---|
| P1194.F_AAAGAACCAAGTTTGC-1 | | | | | | | |
| P1194.F_AAAGAACCACGAGGTA-1 | | | | | | | |
| P1194.F_AAAGGATAGTTACGTC-1 | | | | | | | |
| P1194.F_AAAGGATGTCATCGCG-1 | | | | | | | |
| P1194.F_AAAGGGCAGCTCCGAC-1 | | | | | | | |
| ... | | | | | | | |
| P969_TTTGTTGGTATGCGTT-1 | | | | | | | |
| P969_TTTGTTGGTGCATACT-1 | | | | | | | |
| P969_TTTGTTGGTTTAAGGA-1 | | | | | | | |
| P969_TTTGTTGTCAAGCTGT-1 | | | | | | | |
| P969_TTTGTTGTCCACAGGC-1 | | | | | | | |

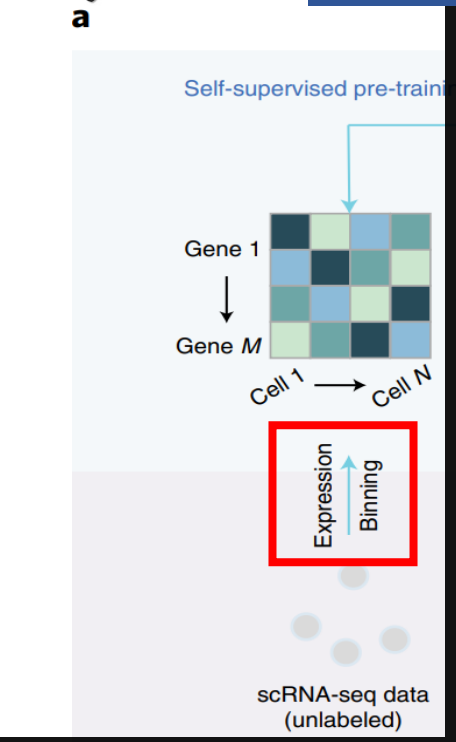
| | orig.ident | nCount_RNA | nFeature_RNA | percent.mt | rename_sample | type |
|---|---|---|---|---|---|---|
| P1194.F_AAAGAACCAAGTTTGC-1 | P1194.F | 10076.0 | 3081 | 7.423581 | P1194.F | WT |
| P1194.F_AAAGAACCACGAGGTA-1 | P1194.F | 1772.0 | 805 | 1.523702 | P1194.F | WT |
| P1194.F_AAAGGATAGTTACGTC-1 | P1194.F | 4984.0 | 2143 | 4.133226 | P1194.F | WT |
| P1194.F_AAAGGATGTCATCGCG-1 | P1194.F | 3292.0 | 1312 | 9.538275 | P1194.F | WT |
| P1194.F_AAAGGGCAGCTCCGAC-1 | P1194.F | 6060.0 | 2278 | 3.036304 | P1194.F | WT |
| ... | ... | ... | ... | ... | ... | ... |
| P969_TTTGTTGGTATGCGTT-1 | P969 | 9172.0 | 2609 | 7.064980 | P969 | WT |
| P969_TTTGTTGGTGCATACT-1 | P969 | 29911.0 | 5913 | 5.613320 | P969 | WT |
| P969_TTTGTTGGTTTAAGGA-1 | P969 | 16849.0 | 4393 | 8.837320 | P969 | WT |
| P969_TTTGTTGTCAAGCTGT-1 | P969 | 10355.0 | 3377 | 9.570256 | P969 | WT |
| P969_TTTGTTGTCCACAGGC-1 | P969 | 4188.0 | 2064 | 7.139446 | P969 | WT |

# scBERT

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
6L
Learned Representations
Reconstructor
Expression Binning
scRNA-seq data (unlabeled)
gene2vec
Gene Embedding
Reference scRNA-seq data (Labeled)
Performer encoder (pre-trained)
1*200 1D Convolution
3-L MLP
Classifier
Label prediction loss
Supervised finetuning

# scBERT



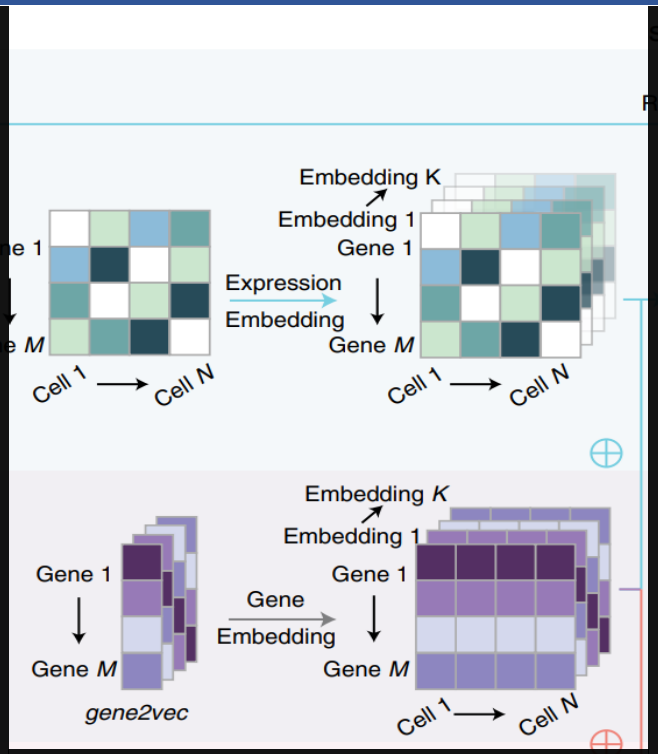

1. Expression Binning

$$x_j^{(i)} = \begin{cases} k, & if\ X_{i,j} > 0\ and\ X_{i,j} \in [b_k, b_{k+1}] \\ 0, & if\ X_{i,j} = 0 \end{cases}$$

2. Cell by Gene Matrix

What does each block mean? What if shuffling columns? What about rows?

# scBERT




3. Expression embedding

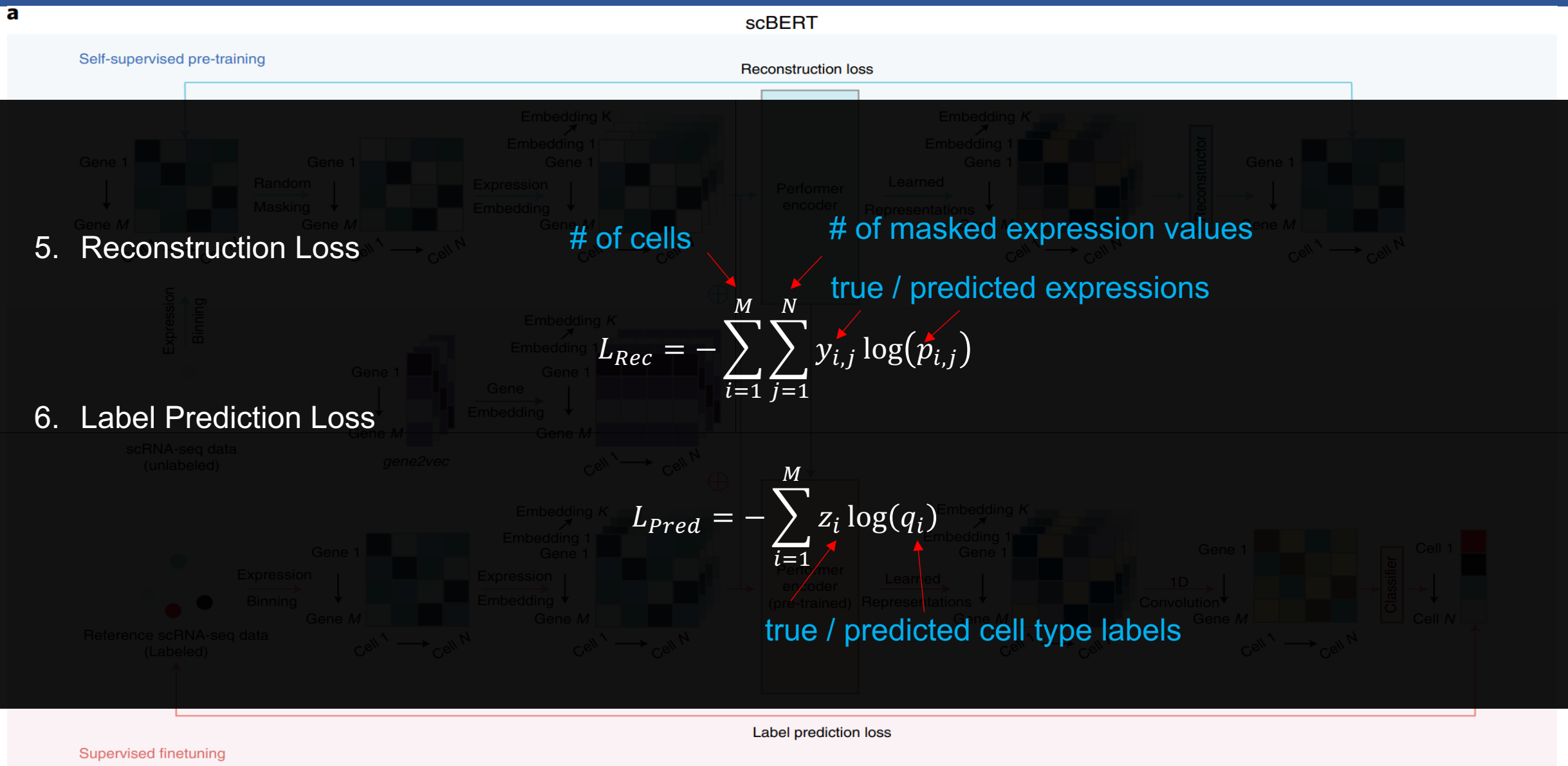
Word Frequency --> Bag of Words, term-frequency stat analysis

Gene Expression --> Occurrence of each gene, binning and converting to 200-d --> Embedding Token

4. Gene embedding

Word2vec --> Gene2vec

# scBERT




5. Reconstruction Loss

# of cells

# of masked expression values

true / predicted expressions

$$L_{Rec} = -\sum_{i=1}^{M}\sum_{j=1}^{N} y_{i,j}\log(p_{i,j})$$

6. Label Prediction Loss

$$L_{Pred} = -\sum_{i=1}^{M} z_i \log(q_i)$$

true / predicted cell type labels

# scGPT

Input Embedding

$$\begin{bmatrix} C_1 \\ C_2 \\ \dots \\ C_N \end{bmatrix} \rightarrow \begin{bmatrix} g_1^{(2)} \\ g_2^{(2)} \\ \dots \\ g_M^{(2)} \end{bmatrix}$$

$$\begin{bmatrix} x_1 \\ x_2 \\ \dots \\ x_N \end{bmatrix} \quad \begin{bmatrix} id\left(g_1^{(2)}\right) \\ id\left(g_2^{(2)}\right) \\ \dots \\ id\left(g_M^{(2)}\right) \end{bmatrix} \quad \begin{bmatrix} t_{c,1} \\ t_{c,2} \\ \dots \\ t_{c,M} \end{bmatrix}$$

$$t_g^{(i)} = \left[id\left(g_1^{(i)}\right), id\left(g_2^{(i)}\right), \dots, id\left(g_M^{(i)}\right)\right]$$

Gene Tokens

Assign each gene a unique ID

Input gene ($g$) tokens of each cell $i$

Maximum input length

$$t_c^{(i)} = \left[t_{c,1}^{(i)}, t_{c,2}^{(i)}, \dots, t_{c,M}^{(i)}\right]$$

Condition Tokens (perturbations, …)

scalar

$$x^{(i)} = \left[x_1^{(i)}, x_2^{(i)}, \dots x_j^{(i)}, \dots, x_M^{(i)}\right]$$

Expression Values

Value binning technique:

$$x_j^{(i)} = \begin{cases} k, & if\ X_{i,j} > 0\ and\ X_{i,j} \in [b_k, b_{k+1}] \\ 0, & if\ X_{i,j} = 0 \end{cases}$$

$$h^{(i)} = emb_g\left(t_g^{(i)}\right) + emb_x\left(x^{(i)}\right) + emb_c\left(t_c^{(i)}\right)$$

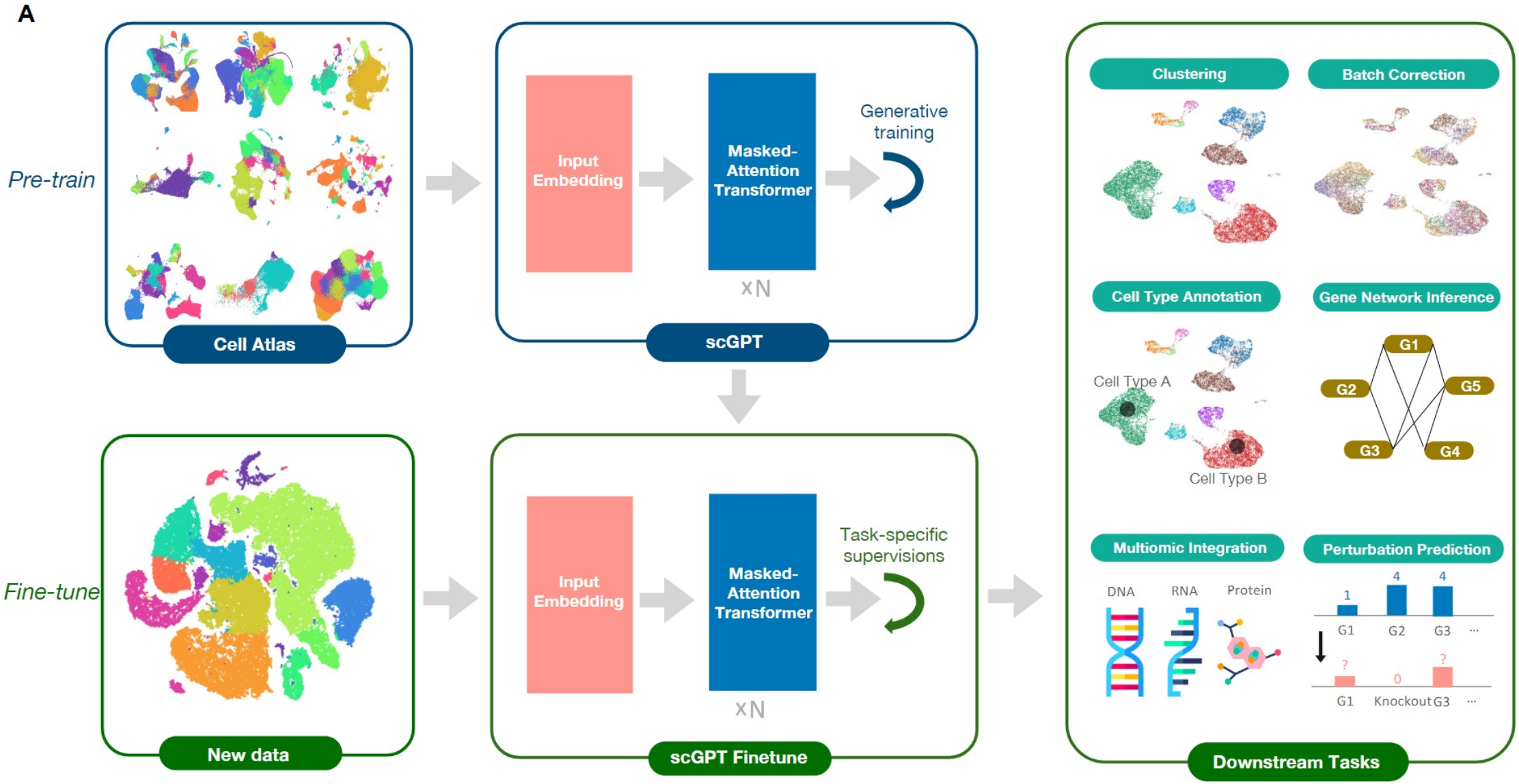
Final Embedding

Fully Connected



Cui, Haotian, et al. "scGPT: Towards Building a Foundation Model for Single-Cell Multi-omics Using Generative AI." bioRxiv (2023): 2023-04.

# scGPT




Cui, Haotian, et al. "scGPT: Towards Building a Foundation Model for Single-Cell Multi-omics Using Generative AI." bioRxiv (2023): 2023-04.

# Thank you.