Cell Reports
Methods

Article

A mixture-of-experts deep generative model for integrated analysis of single-cell multiomics data

Kodai Minoura, Ko Abe, Hyunha Nam, Hiroyoshi Nishikawa, Teppei Shimamura

2021, Cell Reports Methods

2022.7.13

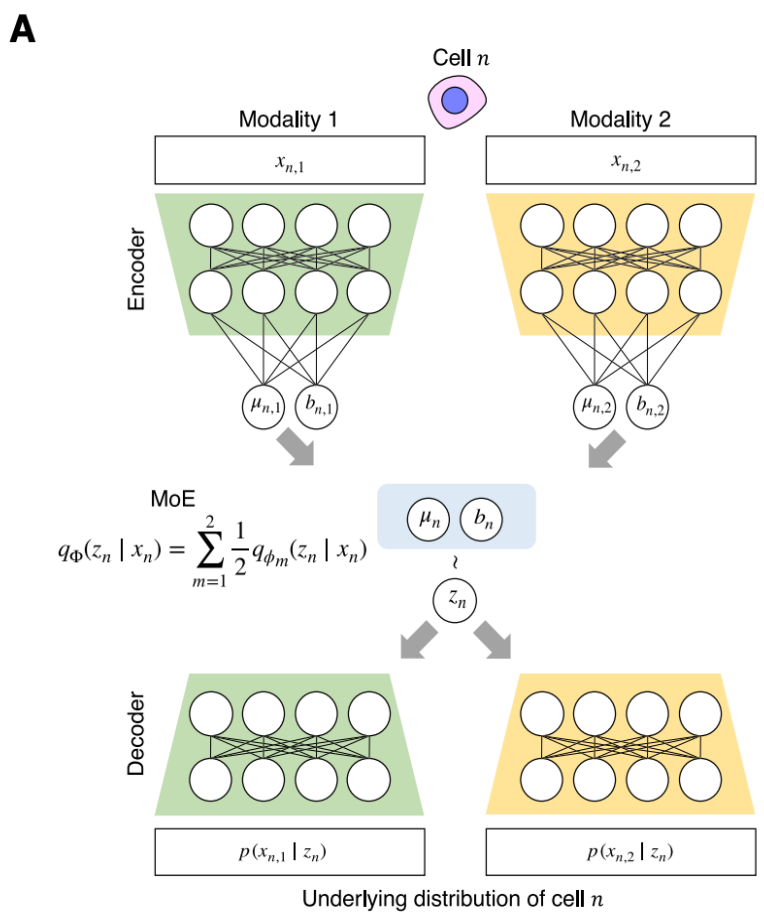
Motivation

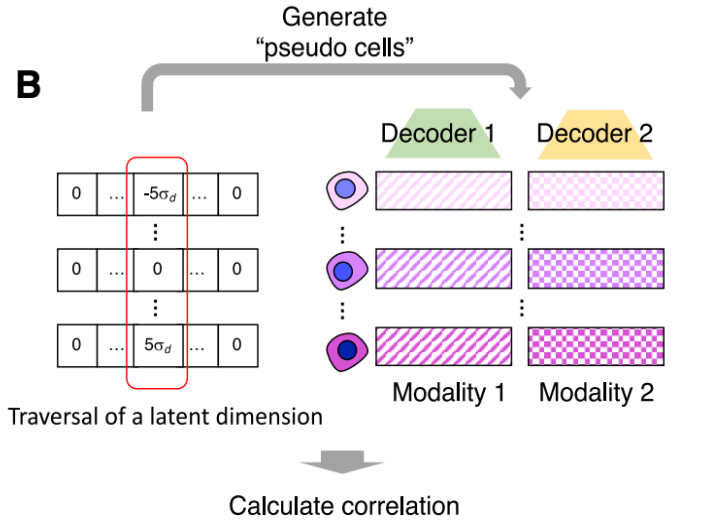
- ✓ Highly complex of single-cell multimodal data
- ✓ multimodal low-dimensional joint representations
- ✓ Cross-modal predictions is unsolved.
- ✓ “Black-box” nature of deep learning.

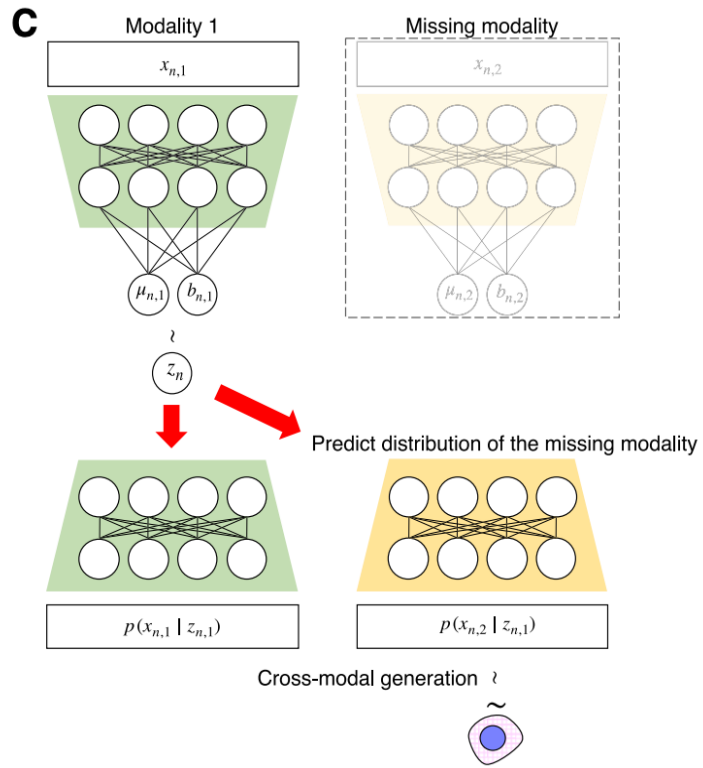
Highlights

- Learning low-dimensional joint representations from single-cell multi-omics data
- Detecting previously overlooked cell populations in single-cell multimodal data
- Accurately predicts missing modalities by crossmodal generation
- Pseudocell generation enables scMM to learn interpretable latent dimensions

Architecture



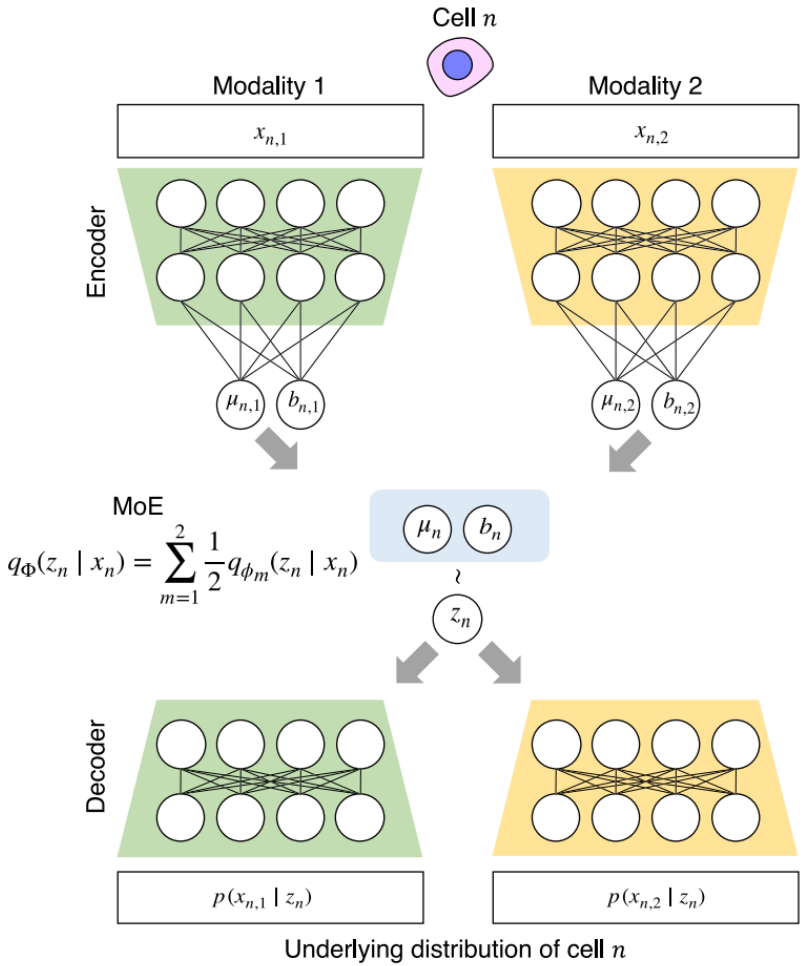




Multimodal latent variables
- Clustering
- Visualization

Data generation
- Prediction of missing modality
- Integration of datasets

A



Multimodal latent variables
- Clustering
- Visualization

Data generation
- Prediction of missing modality
- Integration of datasets

Encoder: $q_{\varphi_m}(\boldsymbol{z}|\boldsymbol{x}_m)$

Sampling:

1. Transcriptome and surface protein

Negative Binomial Distribution

X~NB(r ; P)

Non-negative counts with overdispersion(Gayoso et ，2021)

2. Chromatin accessibility

zero-inflated negative binomial (ZINB) distribution

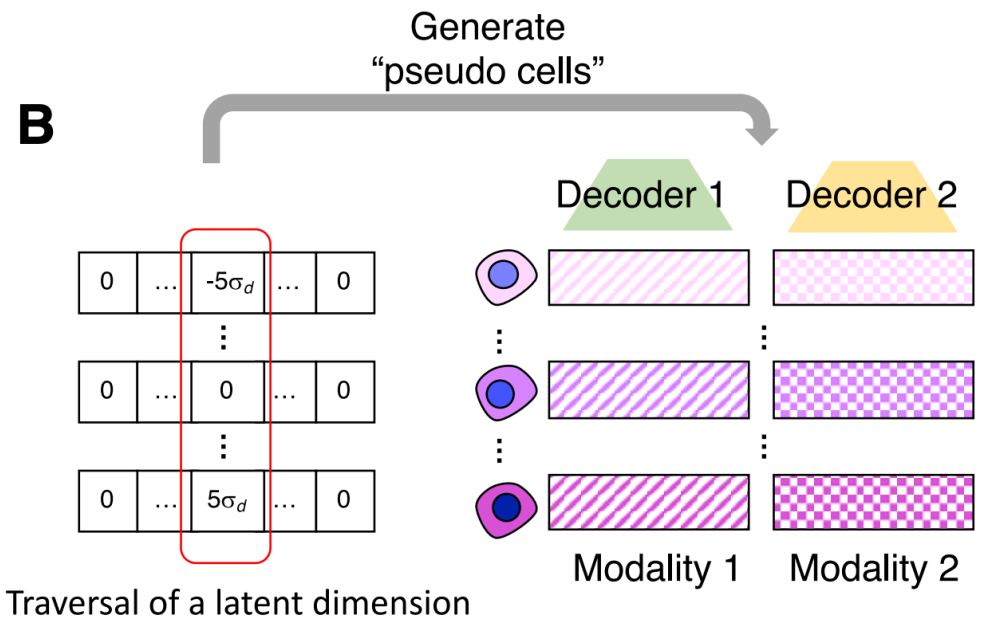
Extreme sparsity

MOE(mixture-of-experts) $q_\Phi(z_n \mid x_n) = \sum_{m=1}^{2} \frac{1}{2} q_{\phi_m}(z_n \mid x_n)$

Decoder: $p_{\theta_m}(\boldsymbol{x}_m|\boldsymbol{z})$

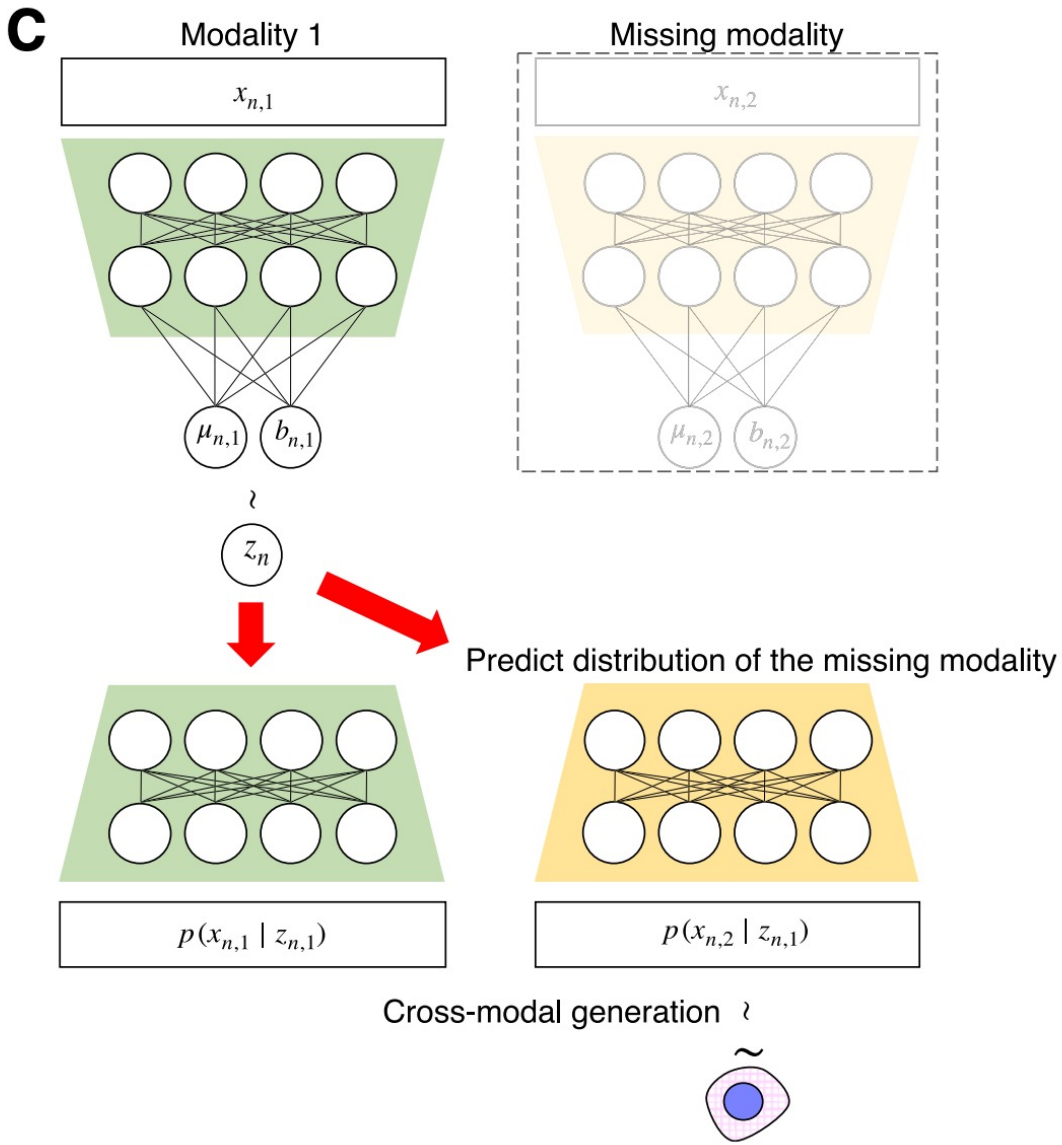
Loss $= \mathbb{E}_{\boldsymbol{z}\sim q_\Phi(\boldsymbol{z}|\boldsymbol{x})}[\log p_\Theta(\boldsymbol{x}|\boldsymbol{z})] - \mathrm{KL}[q_\Phi(\boldsymbol{z}|\boldsymbol{x}) \,\|\, p(\boldsymbol{z})],$

Interpretability



$-5\sigma_d$ to $5\sigma_d$ at a rate of $0.5\sigma_d$.

Cross-model Generation



Experiment

Transcriptome and surface protein

PMBC of CITE-seq dataset（接种疫苗患者的外周血单核细胞数据）

the transcriptome and 224 surface protein, and 54 cell populations

measurements for over 160,000 cells

BMNC if CITE-seq（接种疫苗患者的骨髓单核细胞数据）

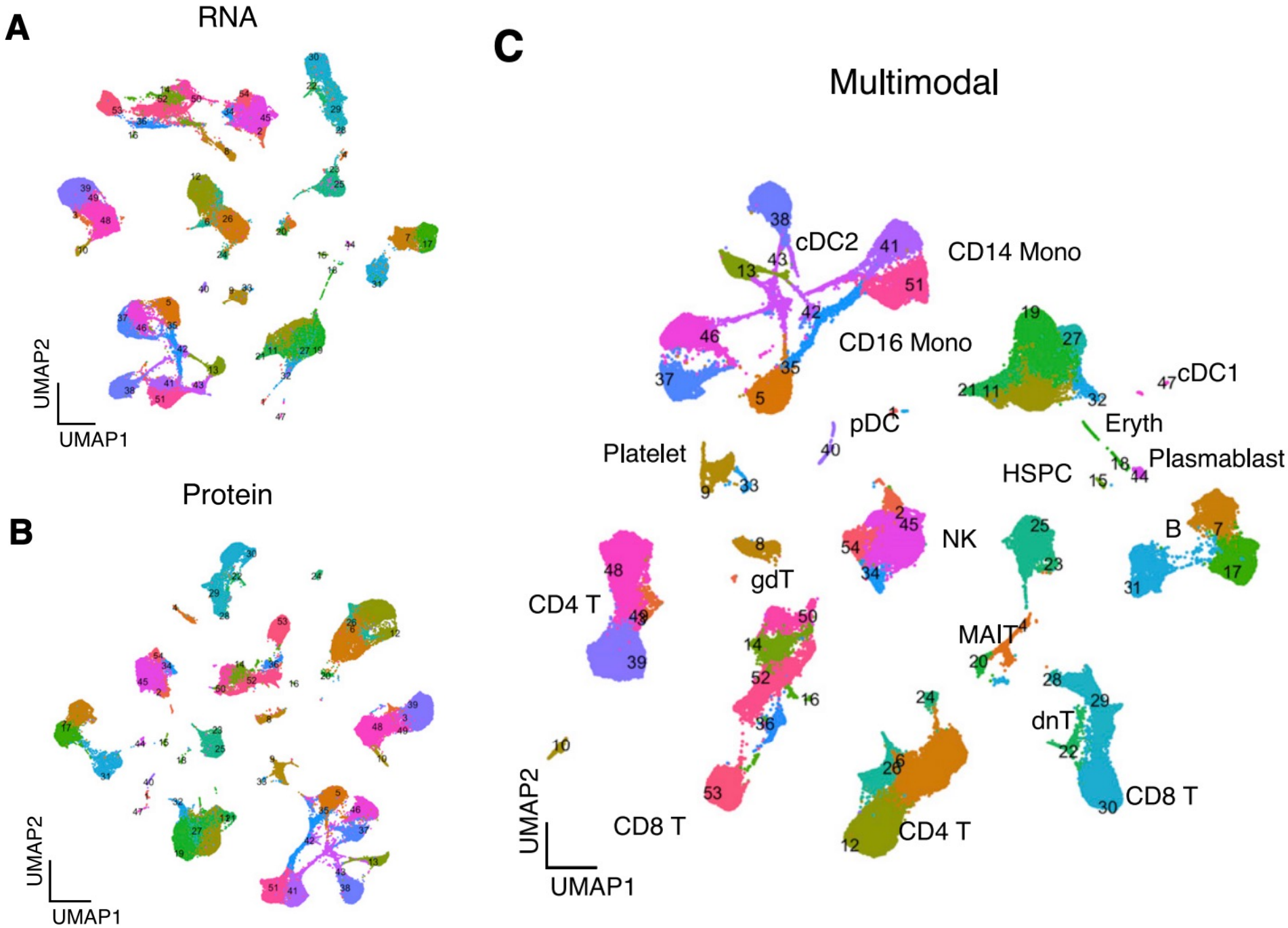
30,000 cells with transcriptome and 25 surface protein

Used for crossmodal estimation

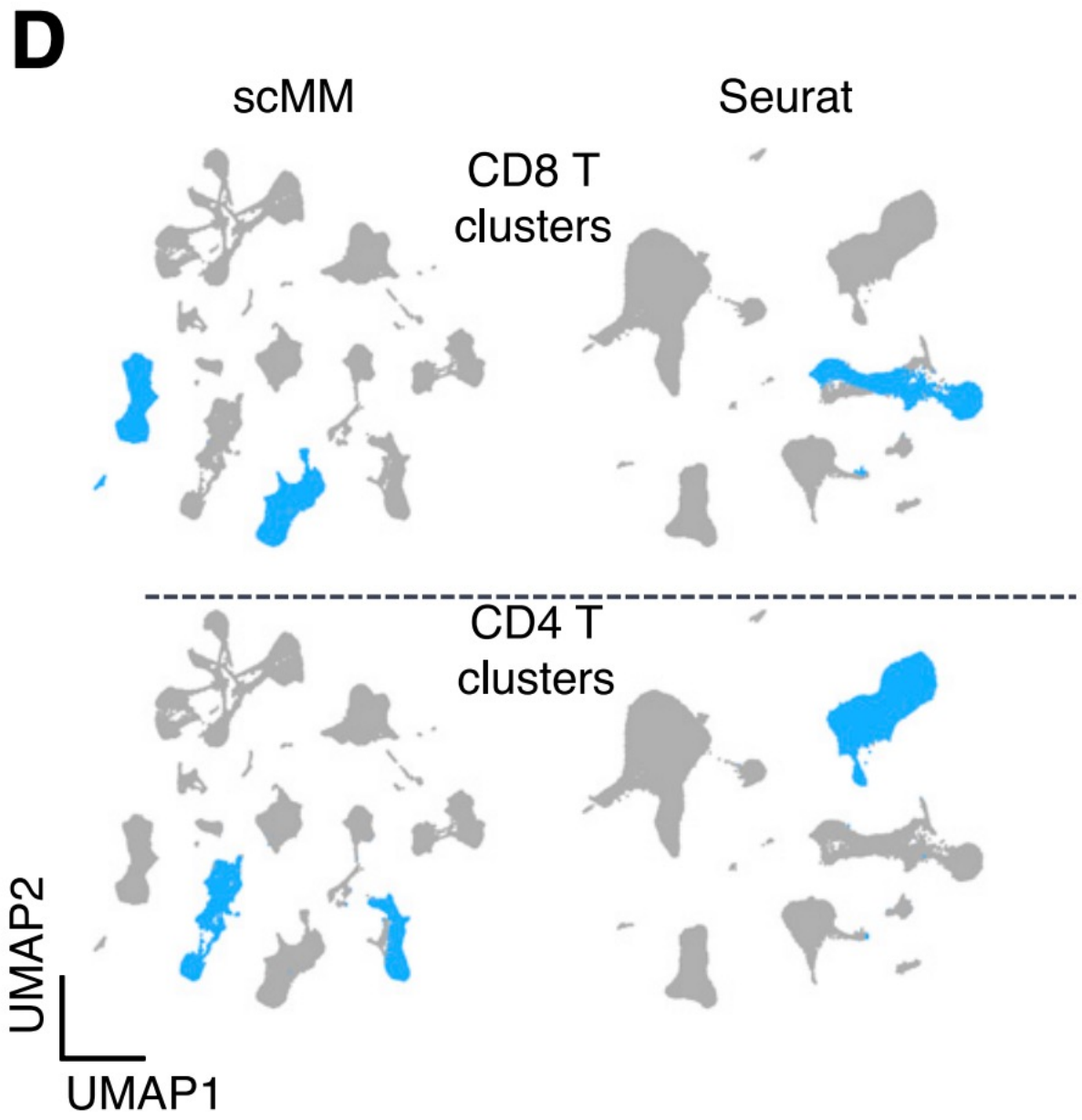
Transcriptome and Chromatin accessibility

SHARE-seq dataset (小鼠皮肤单细胞转录组和染色质可及性)

UMAP Visualization

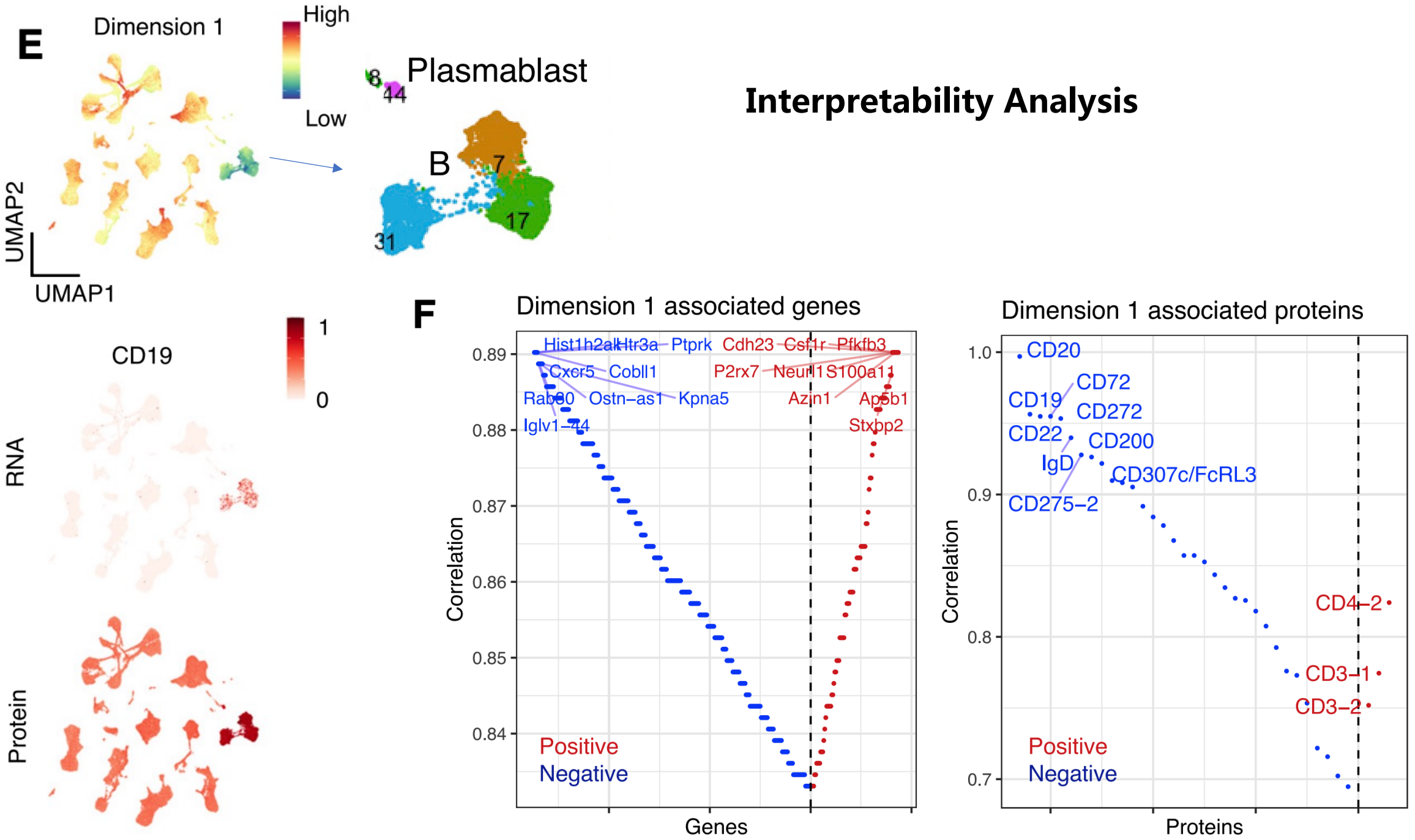


UMAP Visualization



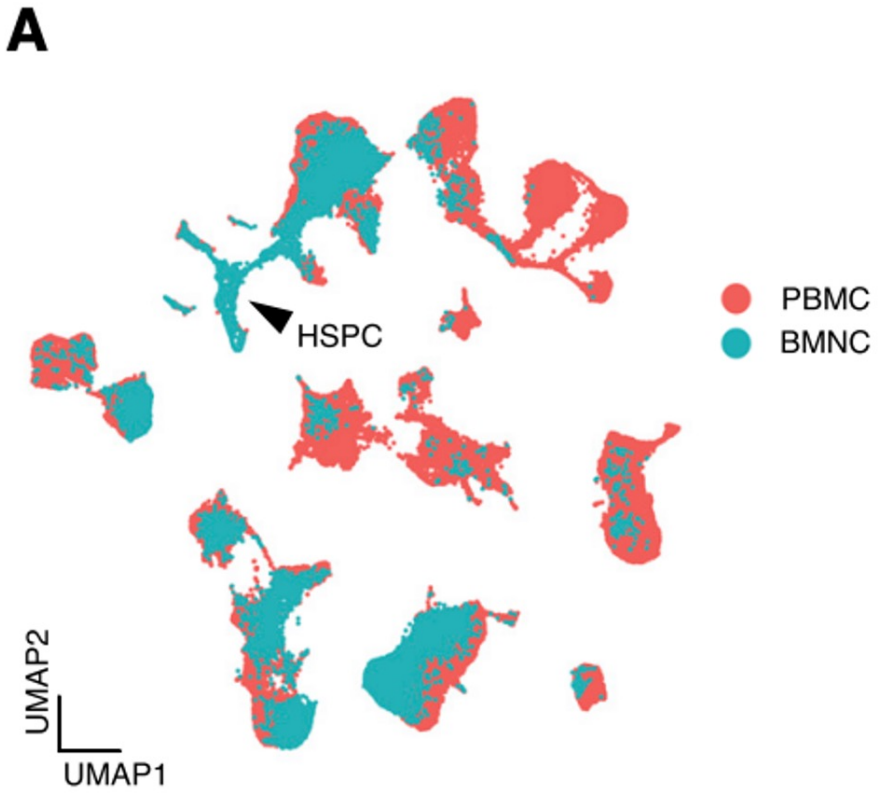
MultiModal Heterogeneity Discover

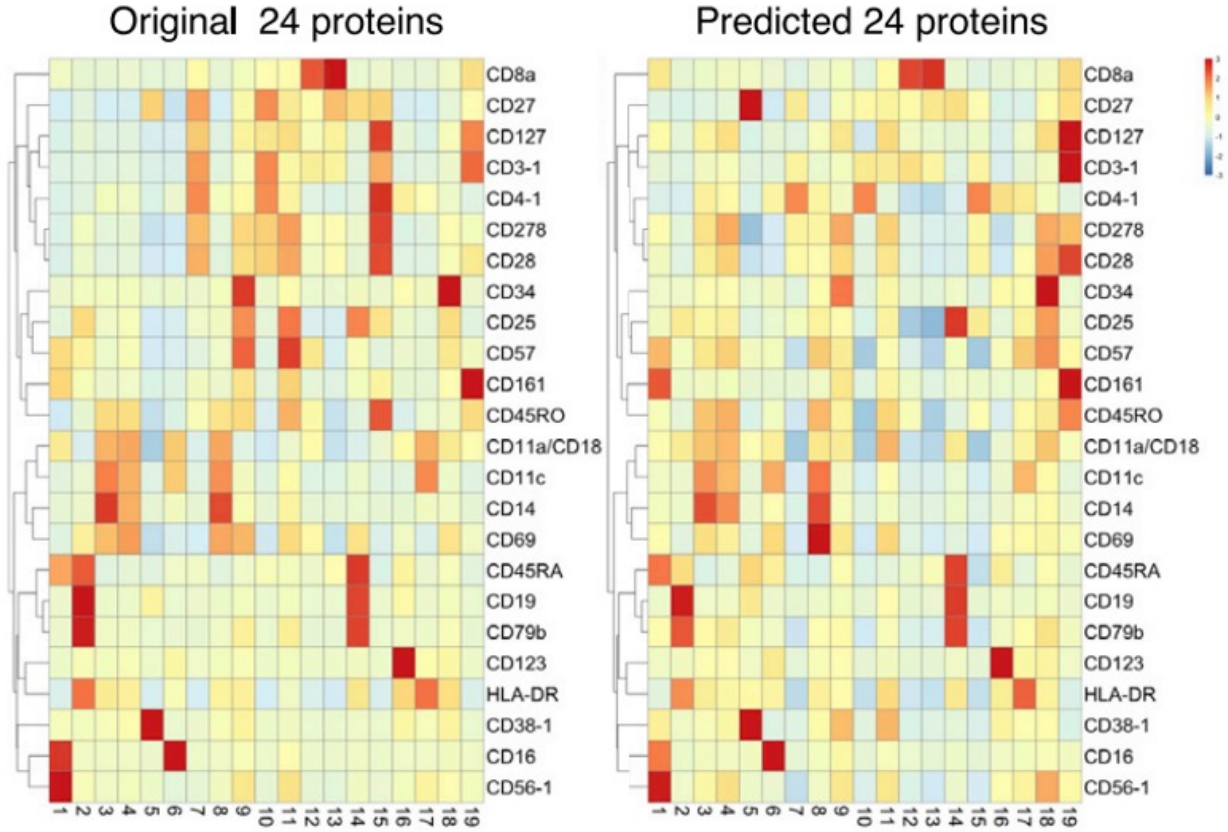
Interpretability Analysis



Crossmodal Prediction

transcriptome-to-protein crossmodal estimation





training with PBMC and test with BMNC

$$p(x_{n,2} \mid z_{n,1})$$

SHARE-seq dataset