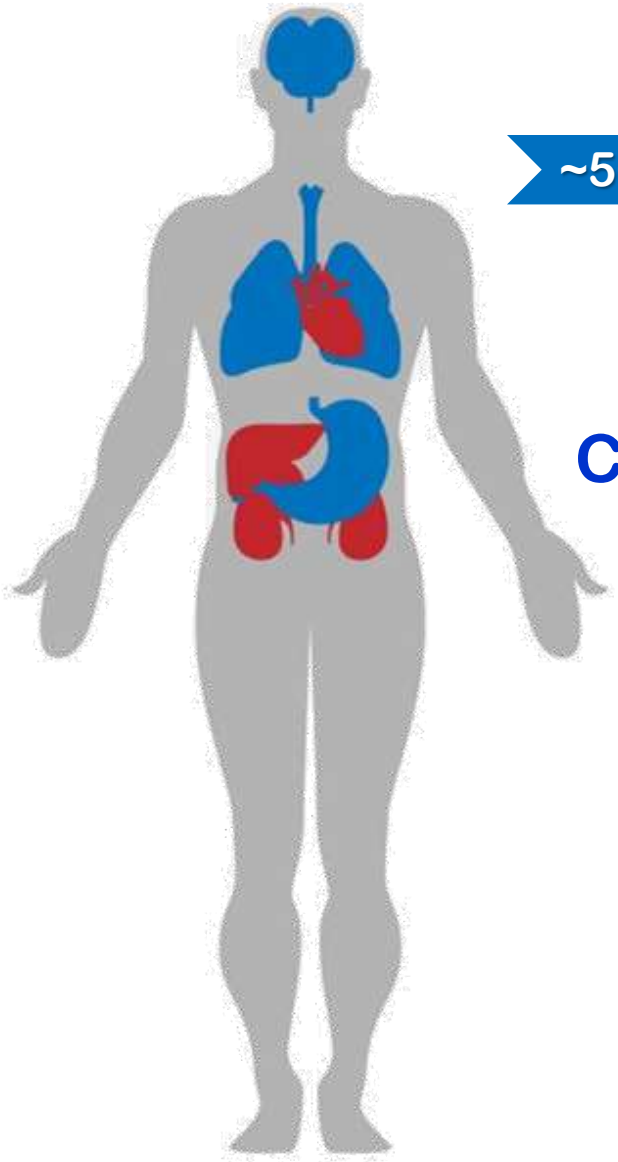
Deep learning for cell type identification based on single-cell chromatin accessibility data

Rui Jiang

Tsinghua University

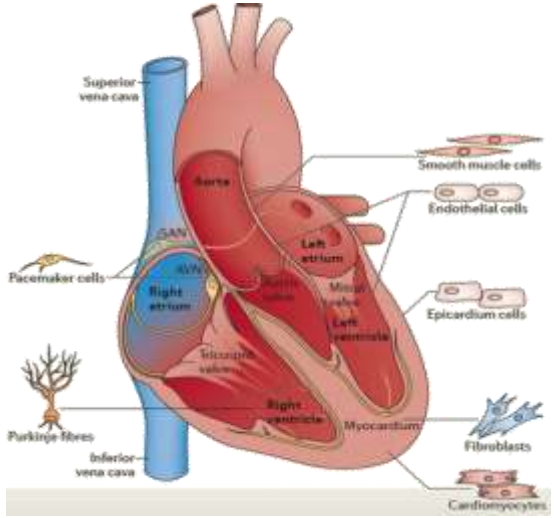
Beijing, China

Cell type

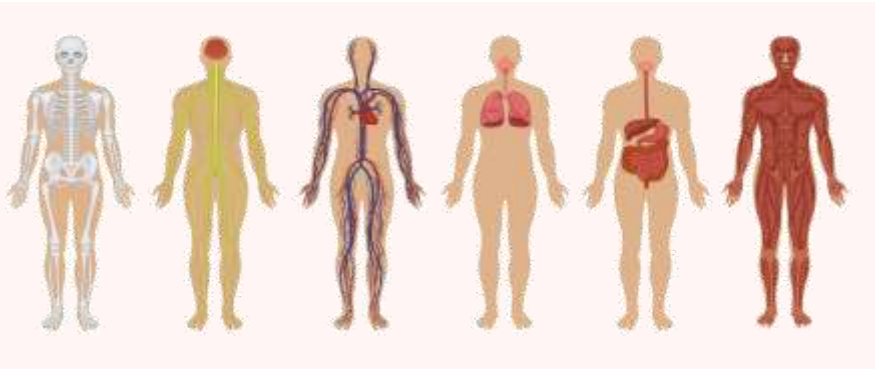


~50 trillion

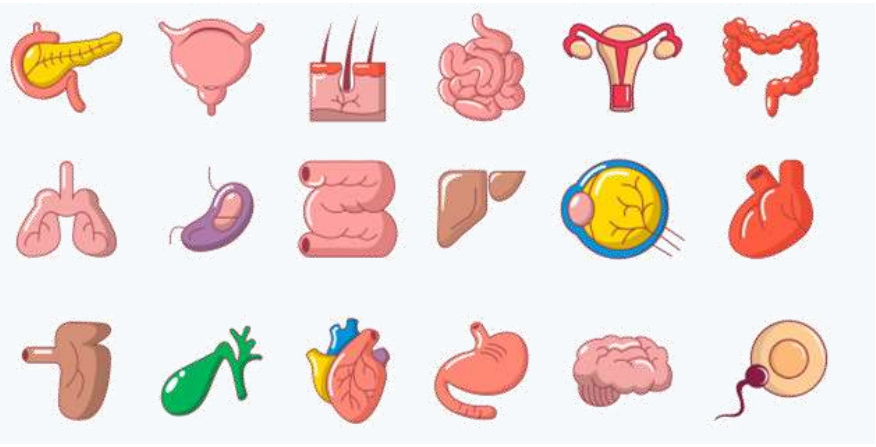
Cells sharing morphological or phenotypical features define a cell type.



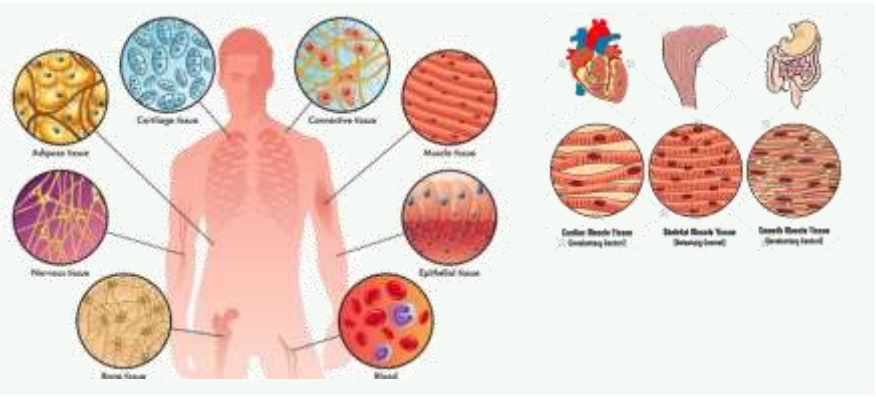
SYSTEM



ORGAN

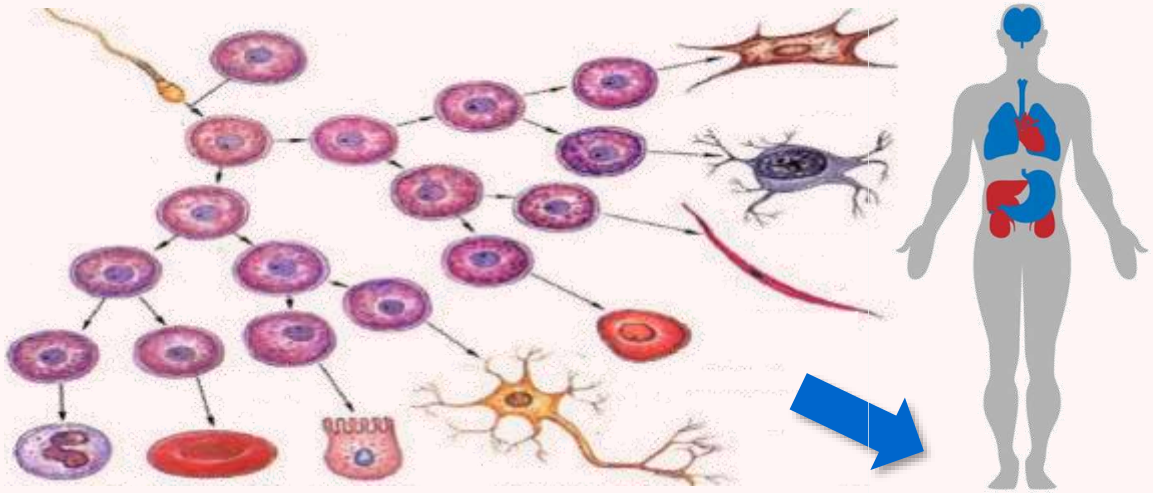


TISSUE

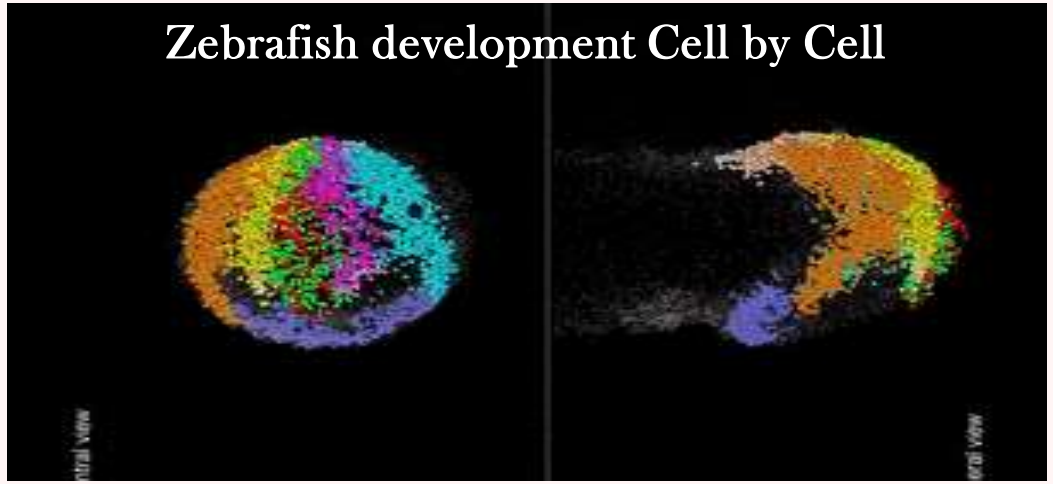


Cell type identification is crucial to biology and medicine

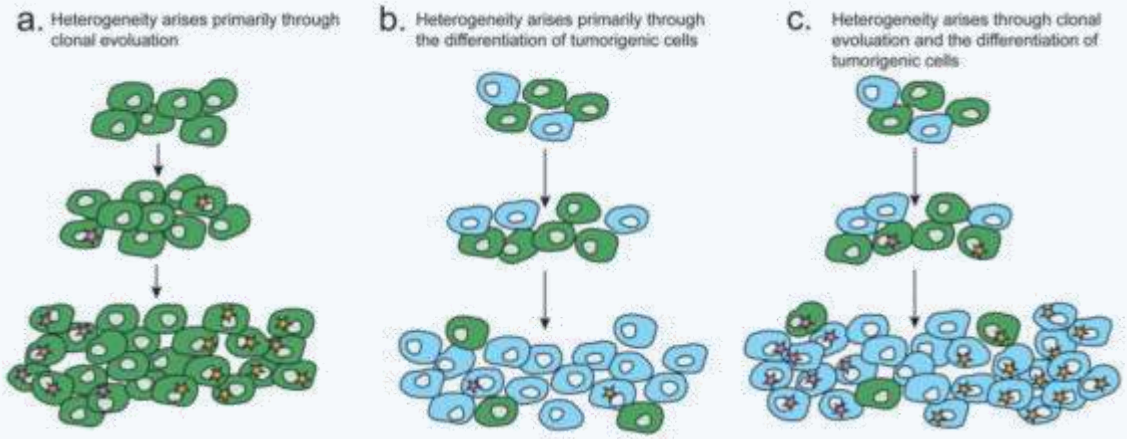
Development



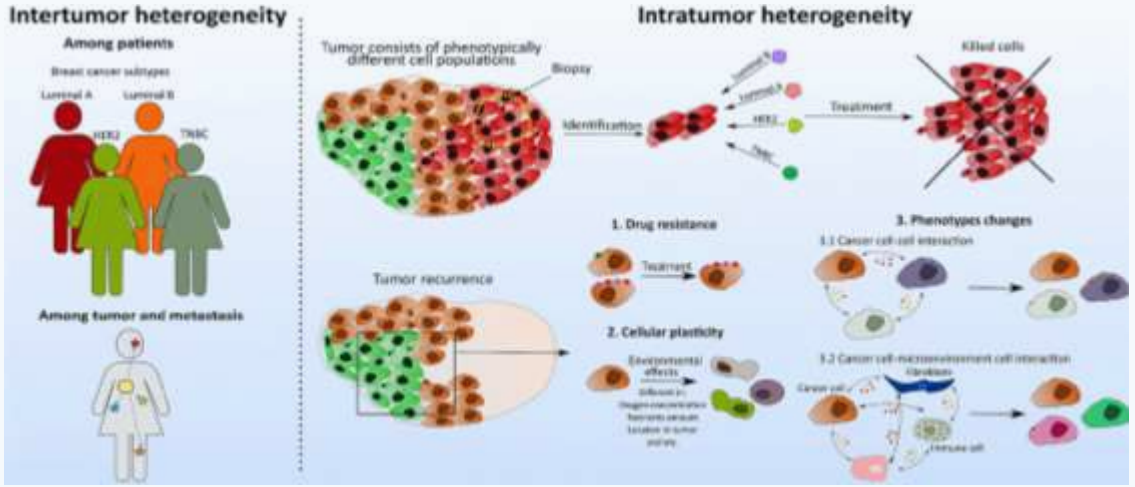
- 2018 Breakthrough of the year (Science)



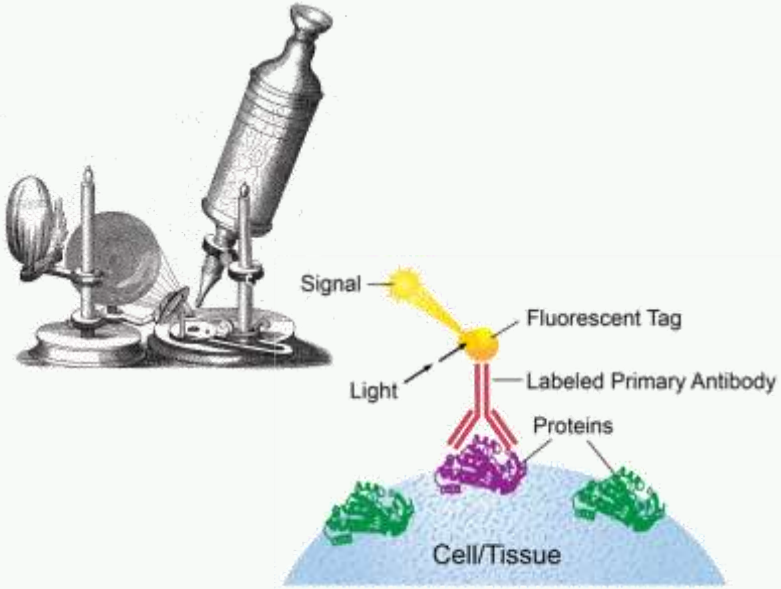
Heterogeneity

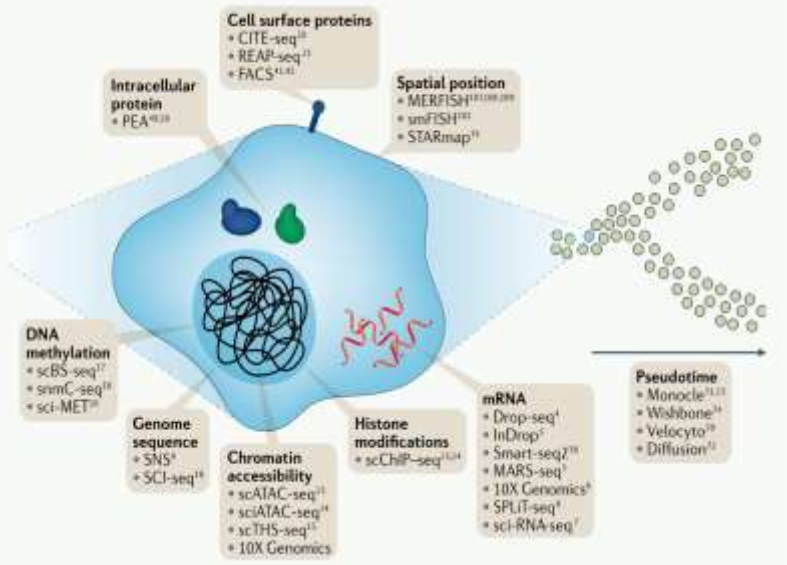


- Heterogeneity of cancer cells

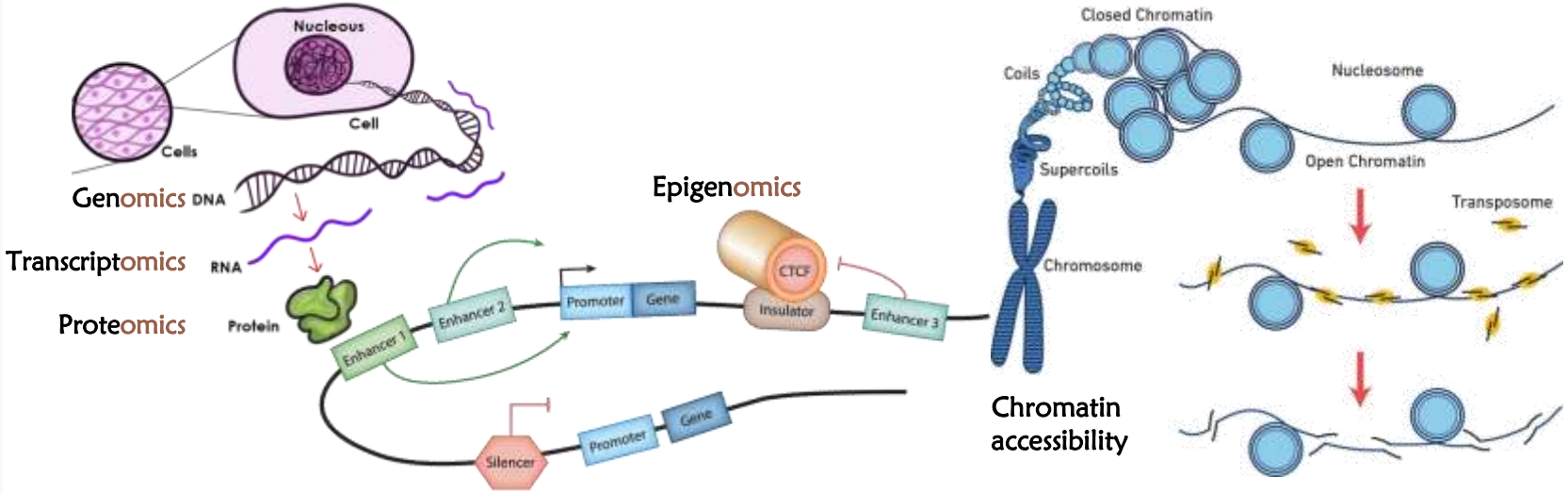


Single-cell sequencing





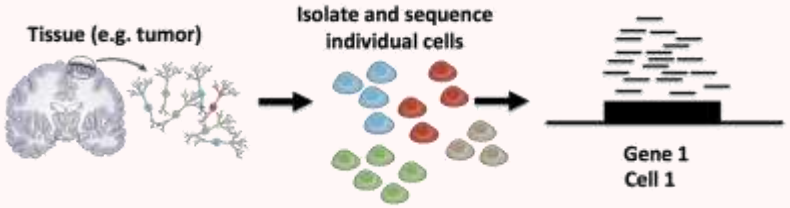
- DNA sequencing **(Methods of the year 2013)** **nature methods**
- RNA sequencing **(Methods of the year 2013)**
- Epigenome sequencing
  - **Chromatin accessibility**
  - Histone modification
  - DNA methylation
- Single-cell multi-omics **(Methods of the year 2019)**
- Spatially resolved sequencing **(Methods of the year 2020)**



Cell type identification based on single-cell data

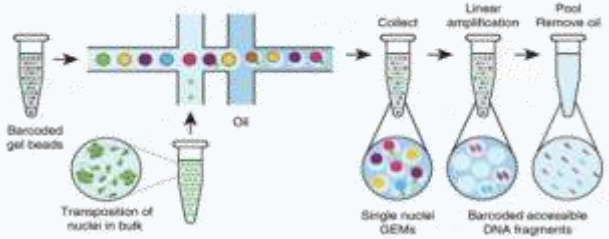
scRNA-seq

- Using gene expression
- High dimensionality (~20K)
- High sparsity (>90% zeros)



scATAC-seq

- Using chromatin accessible regions
- Ultra-high dimensionality (~1M)
- Ultra-high sparsity (>99% zeros)



Multi-Omics data

- Using both gene expression and chromatin accessible regions
- Paired for single cells
- Pared for tissues

DATA

Unsupervised

- Only use experimental data
- Discover novel cell types
- Clustering

Weakly supervised

- Add data with rough annotations
- Discover novel cell types
- Clustering

Supervised

- Add data with detailed annotations
- Annotate known cell types
- Classification

METHOD

More information used

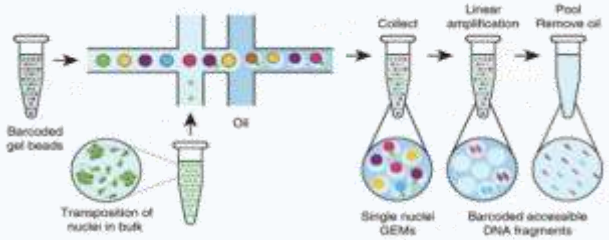
We focus on single-cell chromatin accessibility data

scRNA-seq

- Using gene expression
- High dimensionality (~20K)
- High sparsity (>90% zeros)

scATAC-seq

- Using chromatin accessible regions
- Ultra-high dimensionality (~1M)
- Ultra-high sparsity (>99% zeros)



Multi-Omics data

- Using both gene expression and chromatin accessible regions
- Paired for single cells
- Pared for tissues

DATA

METHOD

Unsupervised

- Only use experimental data
- **Simultaneous clustering and dimensionality reduction**
- **Roundtrip**
- **scDEC**

Weakly supervised

- Add data with rough annotations
- **Simultaneous clustering and dimensionality reduction**
- **RA3**
- **DC3**

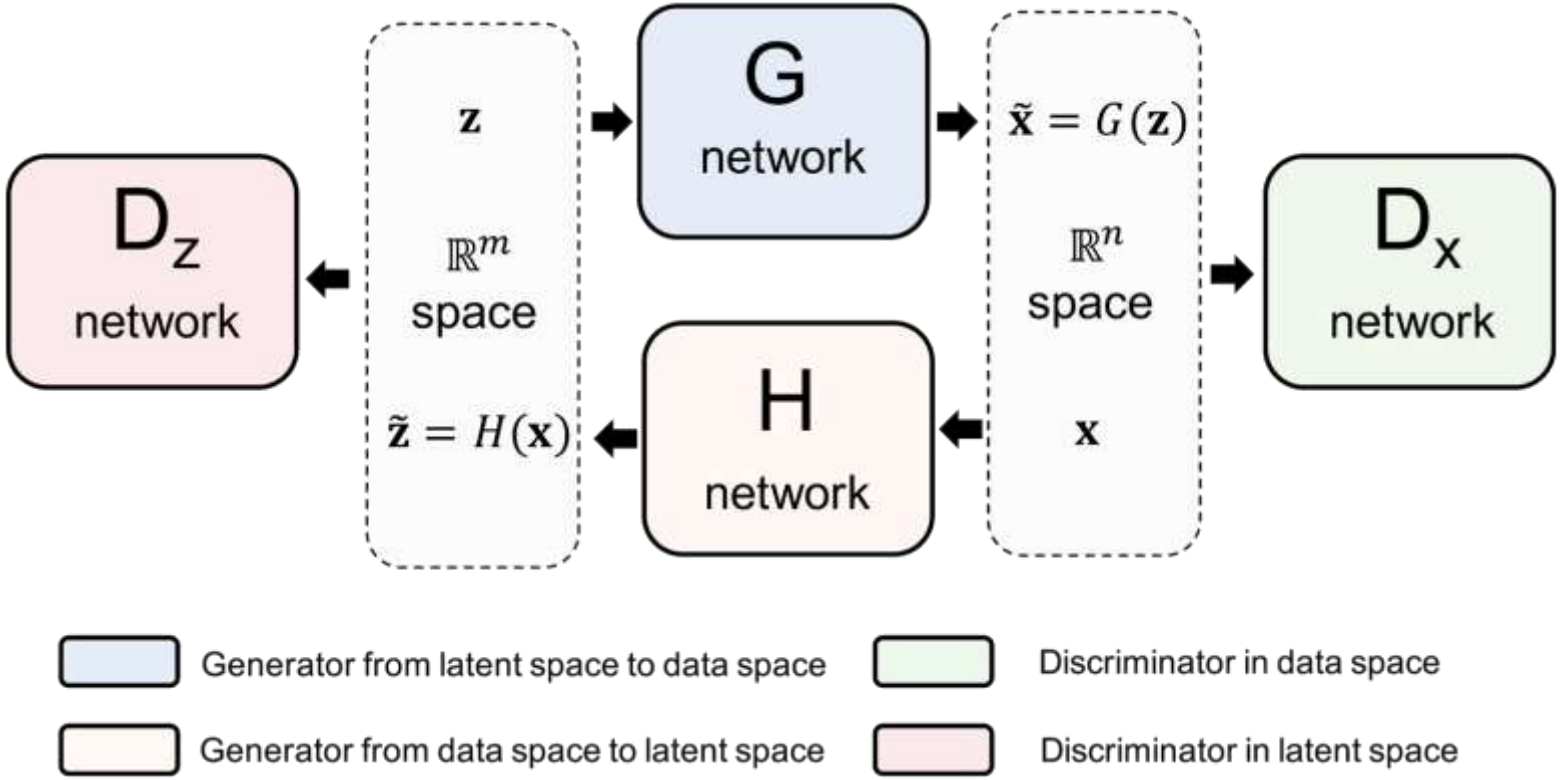
Supervised

- Add data with detailed annotations
- **Classification for cell type annotation**
- **epiAnno**
- **scGraph**

More information used

Deep generative model for density estimation **Roundtrip**

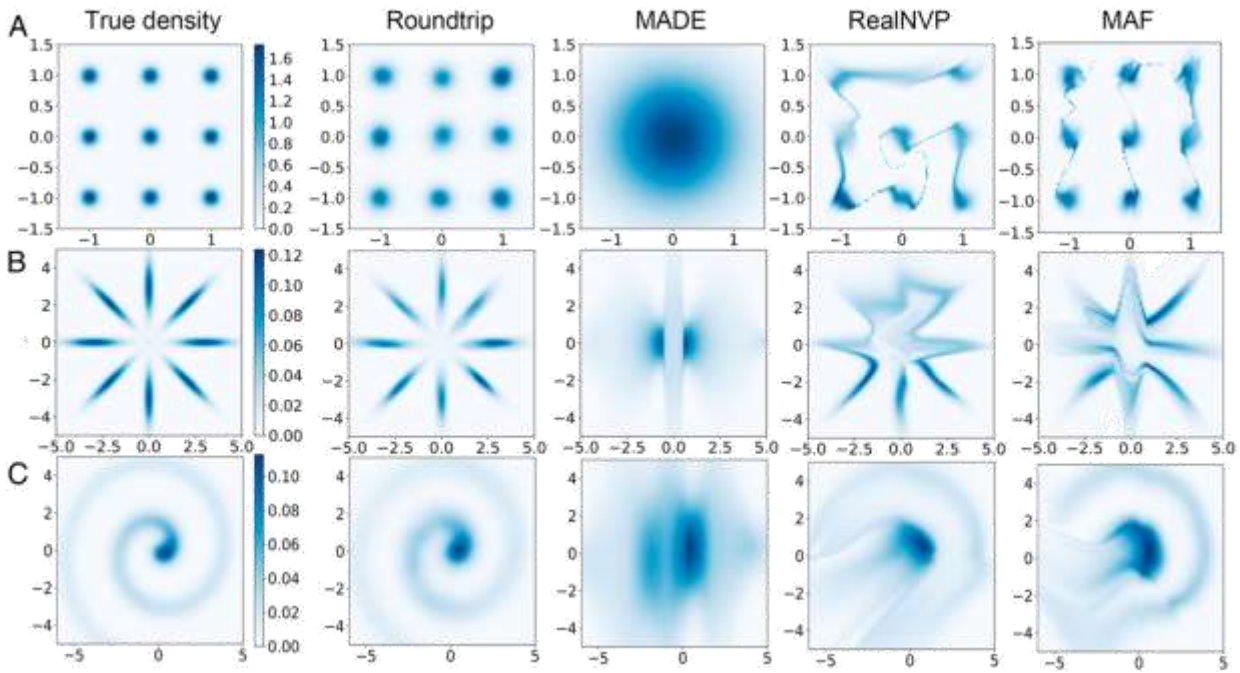
- Ultra-high dimensionality ultra-high sparsity ▸ **Liu et al, PNAS, 2021, 18(15):e2101344118**
- Demand for new theory and method for dealing with these data
- Construct **bi-directional mapping** between data space and latent space
- Utilize **importance sampling** or **Laplace approximation** to estimate density of the data



Deep generative model for density estimation **Roundtrip**

- Performance is superior to existing methods

Simulated data



Real data

	AReM	CASP	HEPMASS	BANK	YPMSD
KDE	6.26 ± 0.07	20.47 ± 0.10	−25.46 ± 0.03	15.84 ± 0.12	247.03 ± 0.61
MADE	6.00 ± 0.11	21.82 ± 0.23	−15.15 ± 0.02	14.97 ± 0.53	273.20 ± 0.35
RealNVP	9.52 ± 0.18	26.81 ± 0.15	−18.71 ± 0.02	26.33 ± 0.22	287.74 ± 0.34
MAF	9.49 ± 0.17	27.61 ± 0.13	−17.39 ± 0.02	20.09 ± 0.20	290.76 ± 0.33
Roundtrip	**11.74** ± 0.04	**28.38** ± 0.08	**−4.18** ± 0.02	**35.16** ± 0.14	**297.98** ± 0.52

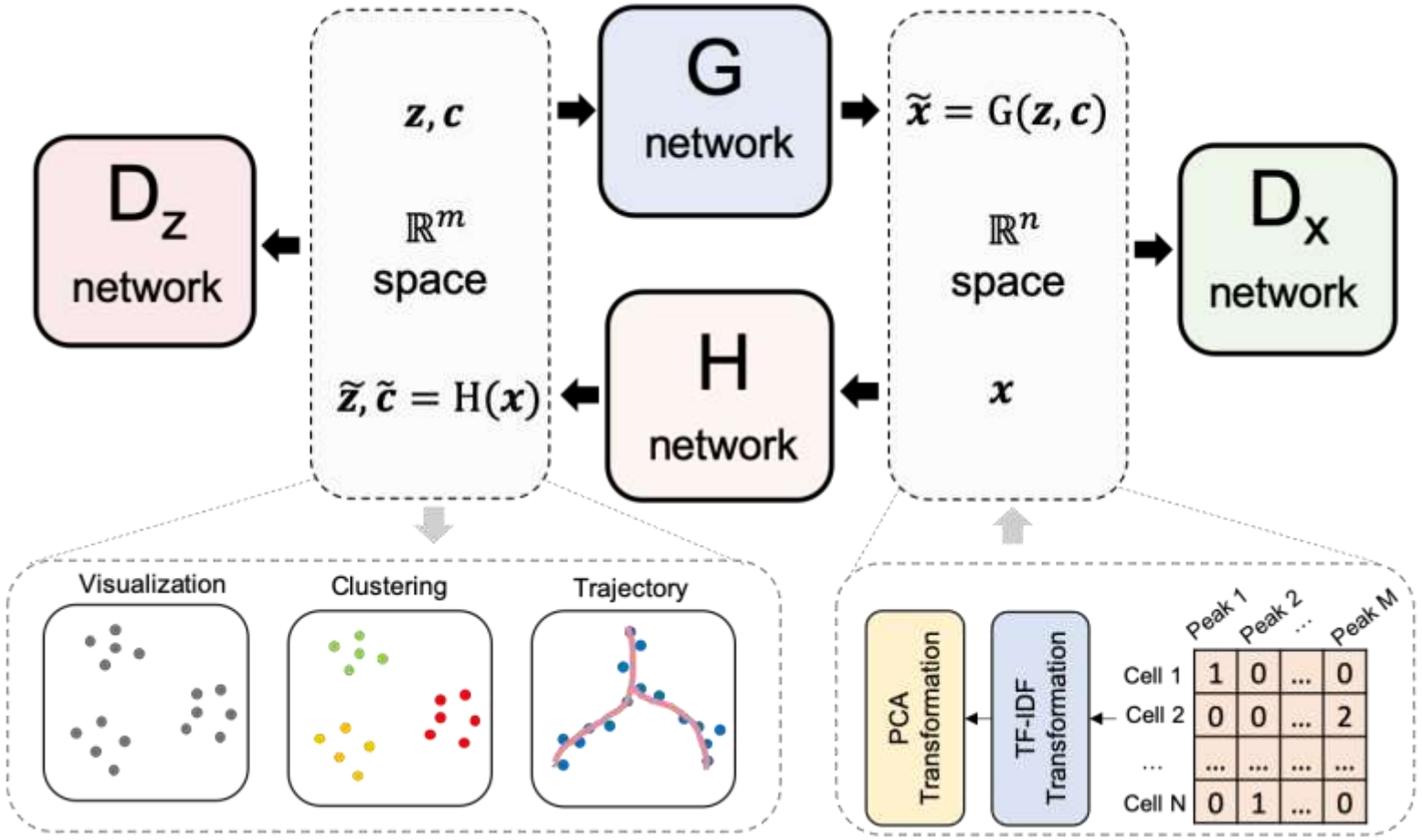
- Solve machine learning tasks
  - Supervised classification
  - Dimensionality reduction
  - Data generation
  - Outlier detection
  - Unsupervised clustering

- AReM: 6 D, activity recognition
- CASP: 9 D, protein tertiary structure
- HEPMASS: 21D, particle collision
- BANK: 17D, marketing campaign
- YPMSD: 90D, audio features of songs

Deep generative model for cell type identification **scDEC**

- Unsupervised cell clustering
- Incorporating cell type label in the latent space
- Simultaneous dimensionality reduction and cell clustering

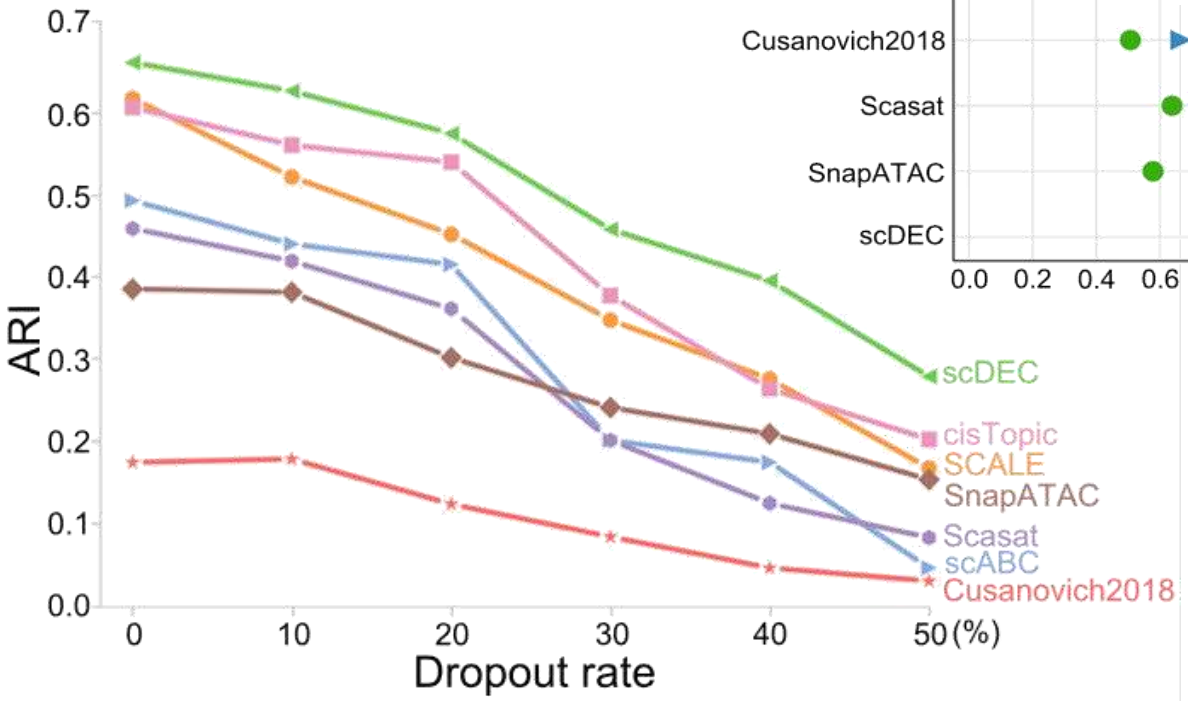
- **Liu et al, Nature Machine Intelligence, 2021, 3:536-544**

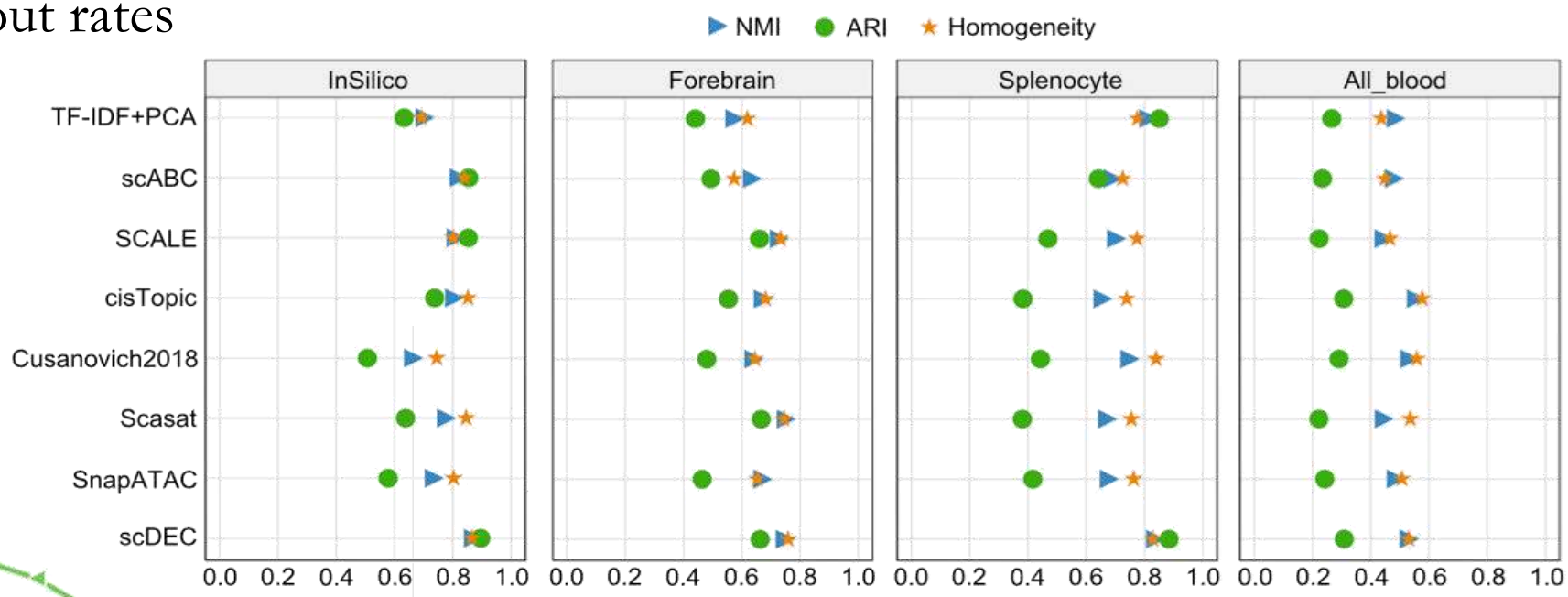


High performance

scDEC

- Superior to 7 existing methods in 4 data sets according to 3 evaluation criteria
- Robust to different dropout rates

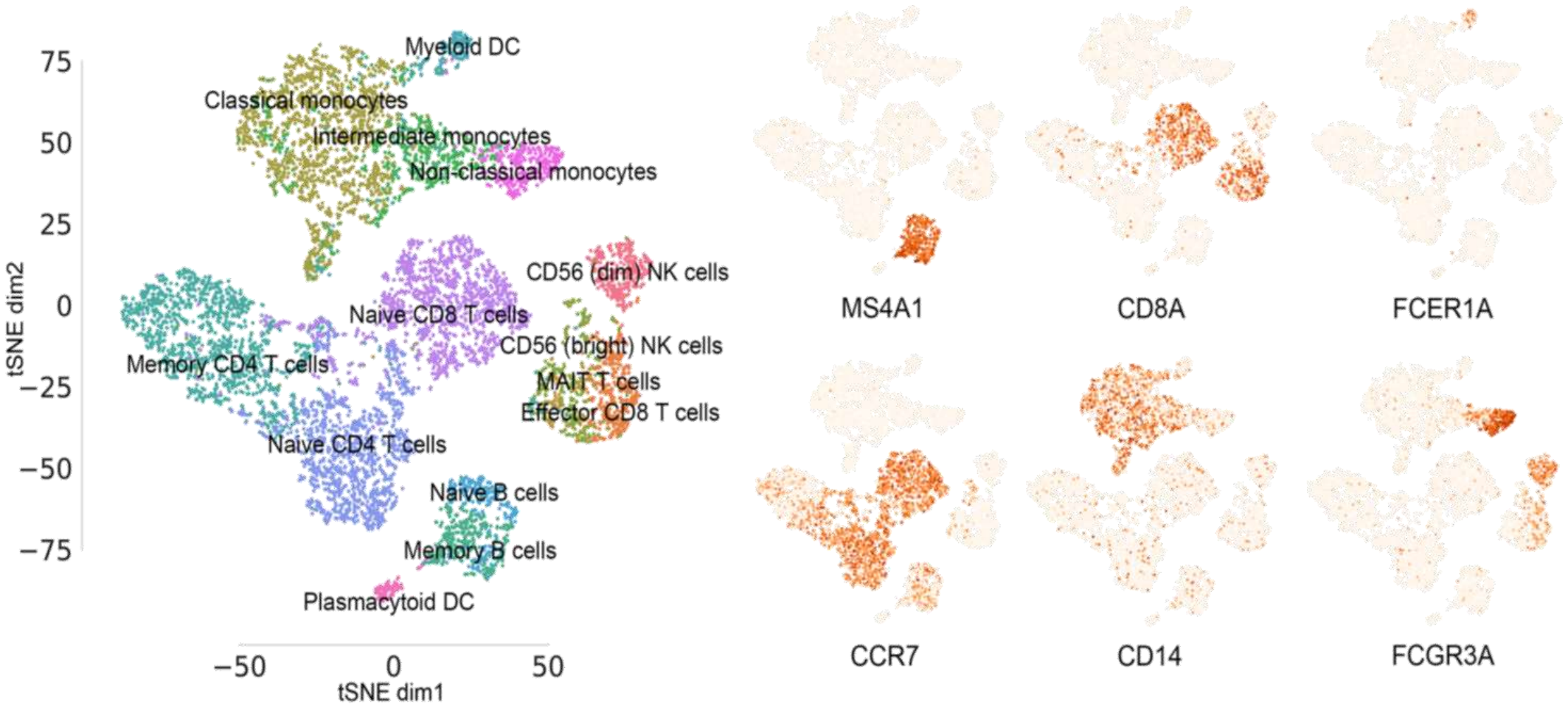




Visualization

scDEC

- Cell clustering with the integration of paired scRNA-seq and scATAC-seq data



10x Genomics PBMC10k data set (scRNA-seq and scATAC-seq paired for single cells)

Weakly supervised learning

Unsupervised

- Only use experimental data
- Low requirement

- Still suffer from such data features as high noise and batch effects in single-cell data
- Hard to identify rare cell types

Weakly supervised

- Abundant bulk sequencing data also contain cell type information
  - Summary information: Mean, variance
  - Low noise, low sparsity: helpful to overcome high noise, high sparsity problems
- Existing approaches make use of reference data
  - Calculate correlation between single cells and reference data (scRNA-seq)
    - Li et al, Nature Genetics, 2017, 49(5):708-18
  - Apply PCA to reference data, then map single cell data (scATAC-seq)
    - Buenrostro et al, Cell, 2018, 173(6):1535-48
    - Lareau et al, Nature Biotechnology, 2019, 37(8):916-24
- Limitation
  - Assume biological variation is identical in single-cell data and reference data

Weakly supervised generative model

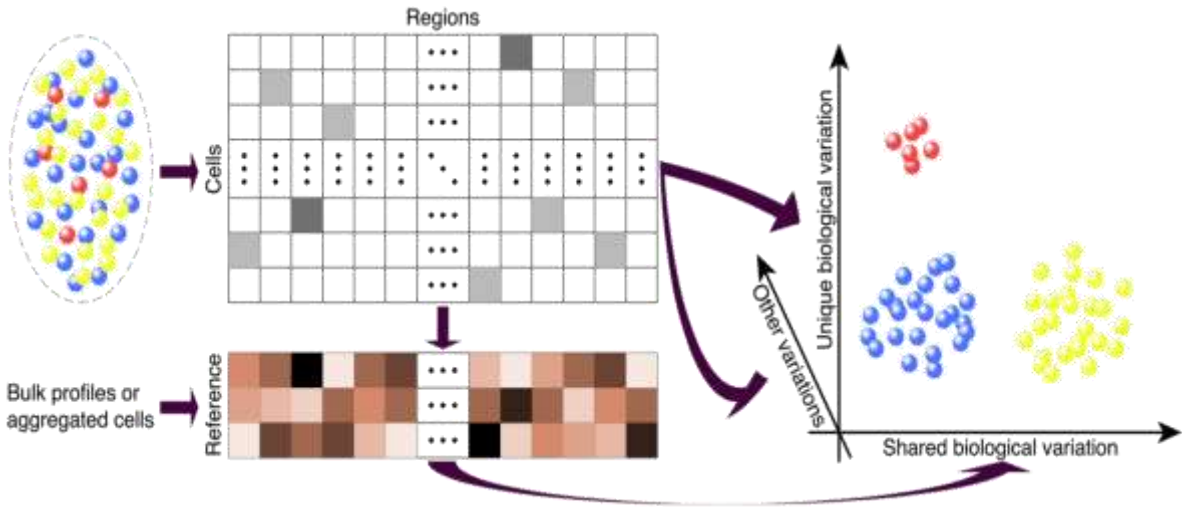
RA3

- Decompose variation to:
- **Chen et al, Nature Communications, 2021, 12:2177**

① Shared biological variation between single-cell and bulk data

② Unique biological variation in single-cell data

③ Other technical variation



$$\mathbf{y}_j|\boldsymbol{\lambda}_j \sim \mathcal{N}_p(\boldsymbol{\lambda}_j, \sigma^2\mathbf{I}_p),$$

$$\boldsymbol{\lambda}_j = \boldsymbol{\beta}\mathbf{x}_j + \mathbf{W}\mathbf{h}_j, \boldsymbol{\lambda}_j \in \mathbb{R}^{p\times 1},$$

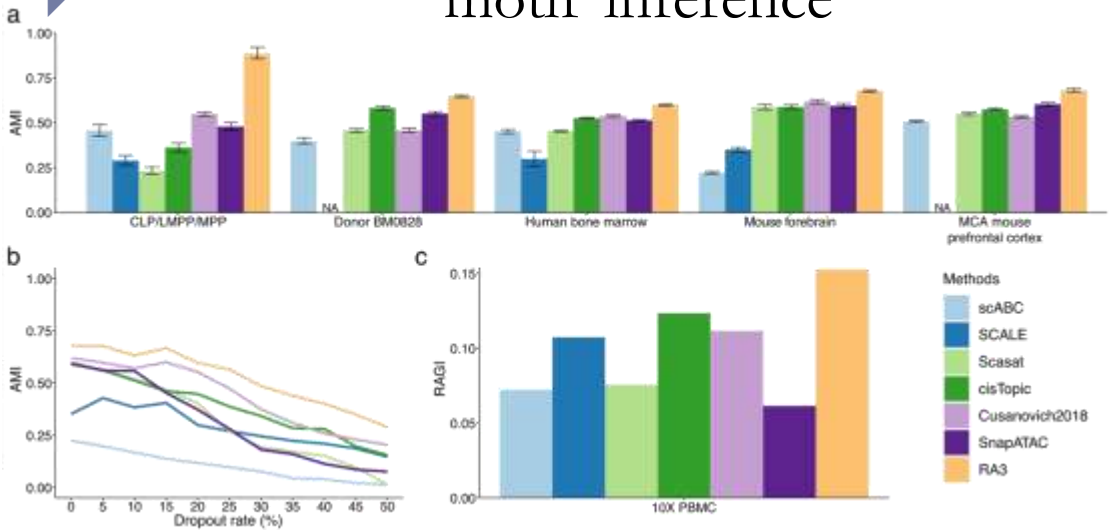
$$\mathbf{W}\mathbf{h}_j = \mathbf{W}_1\mathbf{h}_{j_1} + \mathbf{W}_2\mathbf{h}_{j_2} + \mathbf{W}_3\mathbf{h}_{j_3}$$

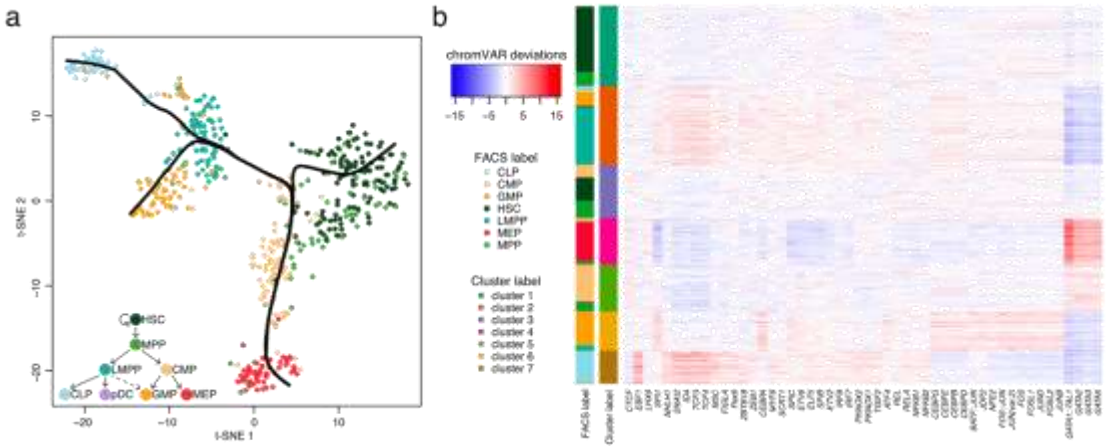
① ② ③

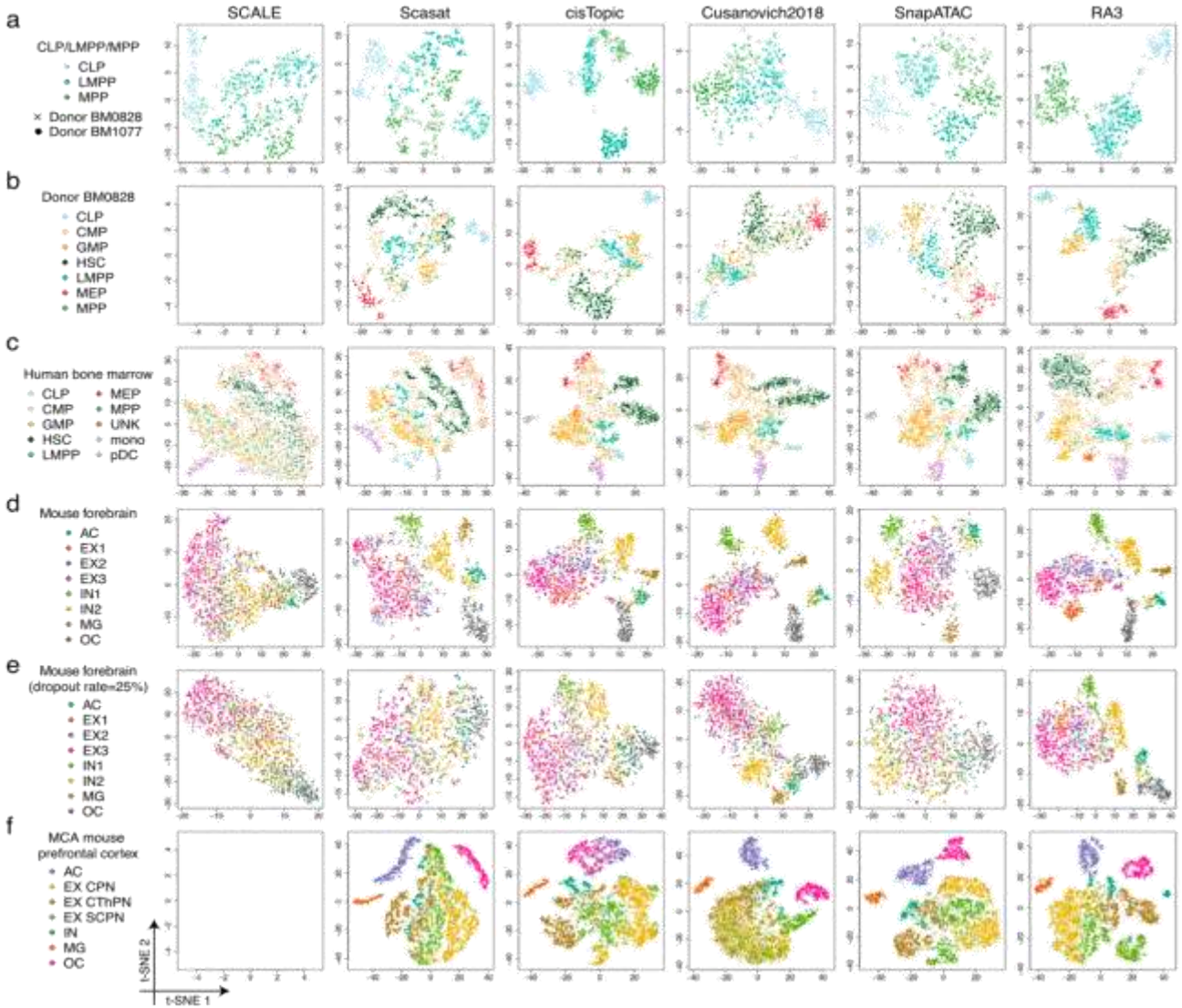
Superior performance

RA3

- Visualization and cell clustering
- Applications: trajectory inference
- motif inference



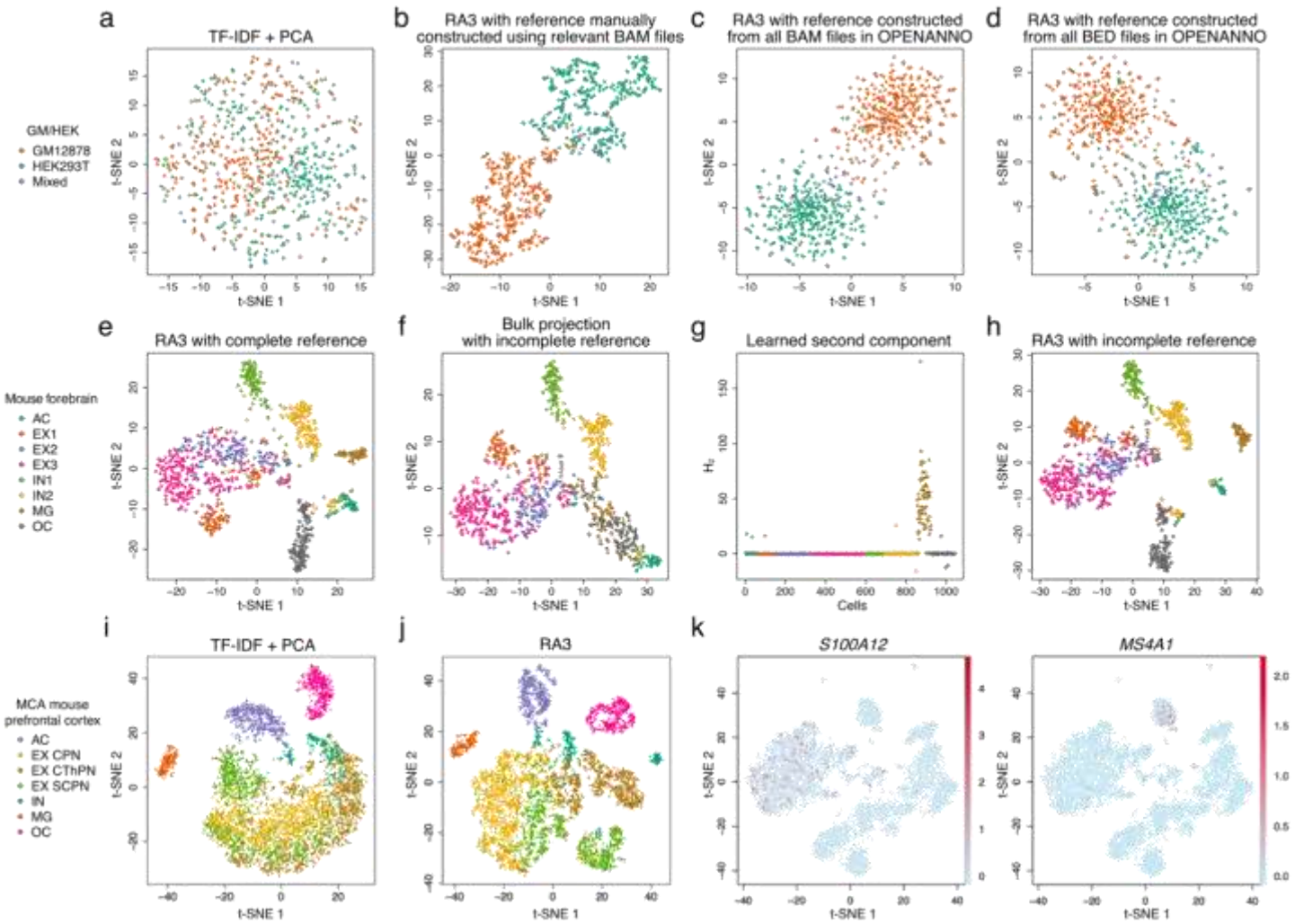




Different means for obtaining reference data

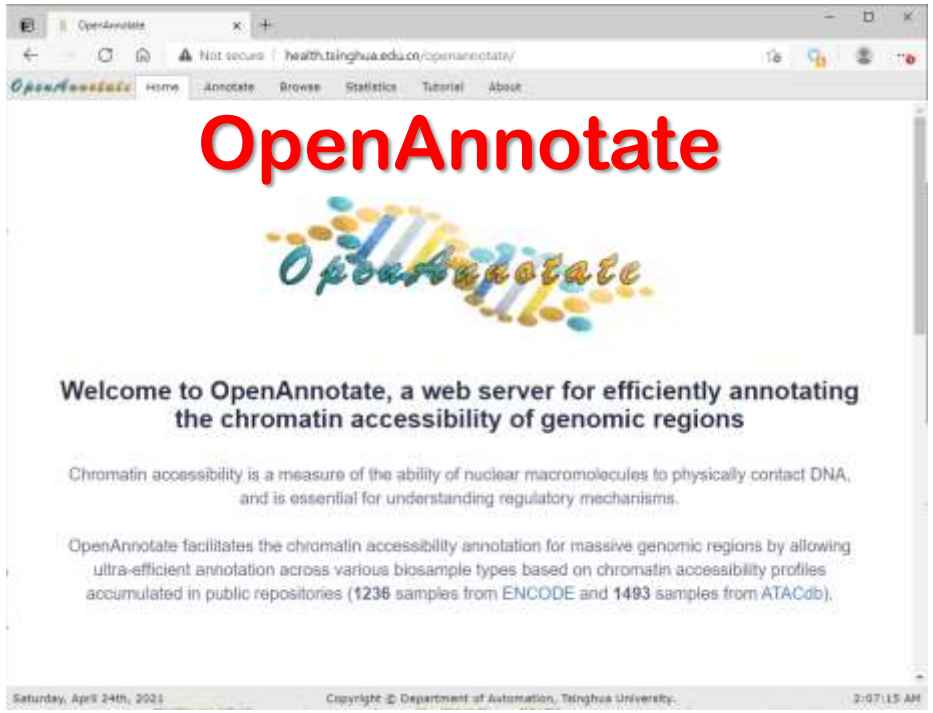
RA3

- All these means of using reference data are effective



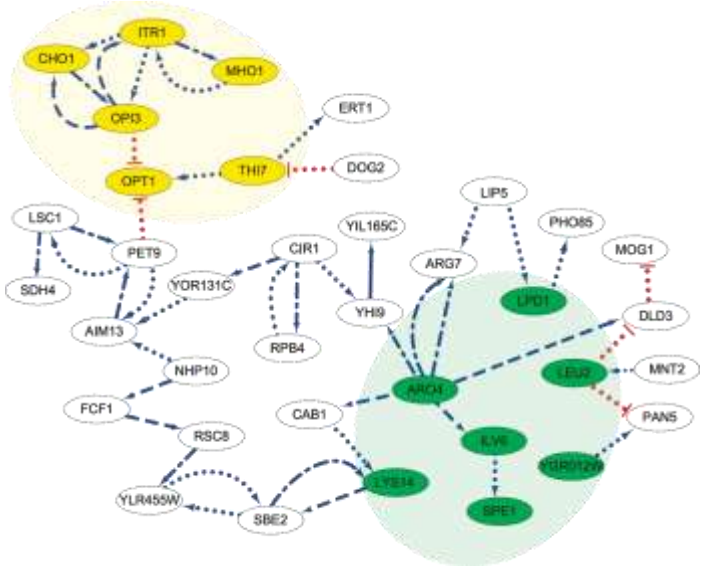
- Bulk ATAC-seq
- Bulk DNase-seq
- Aggregated scATAC-seq data

Chen et al, NAR, 2021, 49(W1):W383-W490

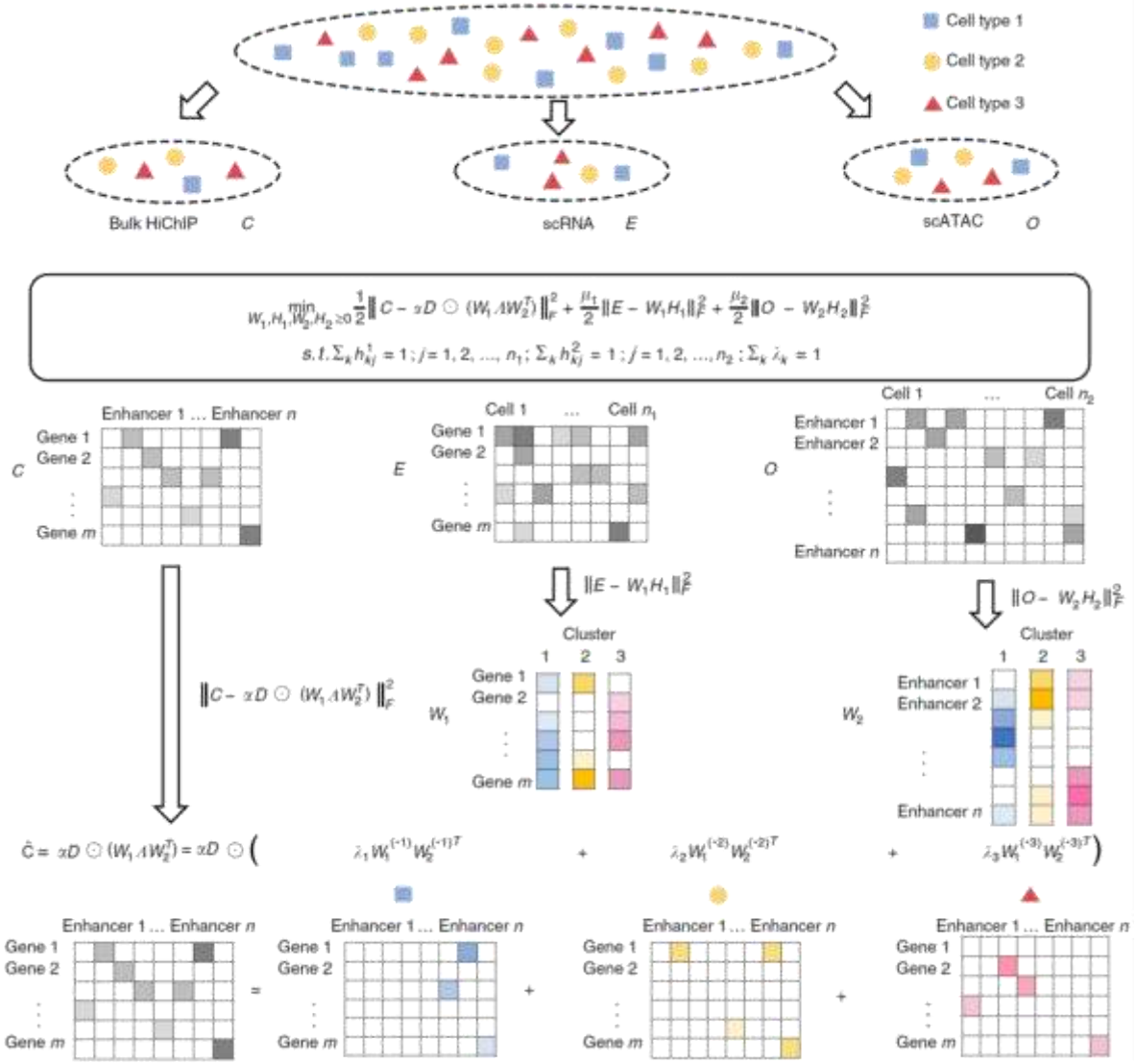


Regulatory network construction via matrix factorization **DC3**

- Multi-omics data integration
  - Single-cell RNA-seq data
  - Single-cell ATAC-seq data
  - Bulk HiChIP data
- Deconvolution of the bulk data
  - Obtain cell clusters (cell types)
  - Obtain cell type specific regulatory relationships
  - Obtain cell type specific regulatory networks



- **Zeng et al, Nature Communications, 2019, 10:4613**

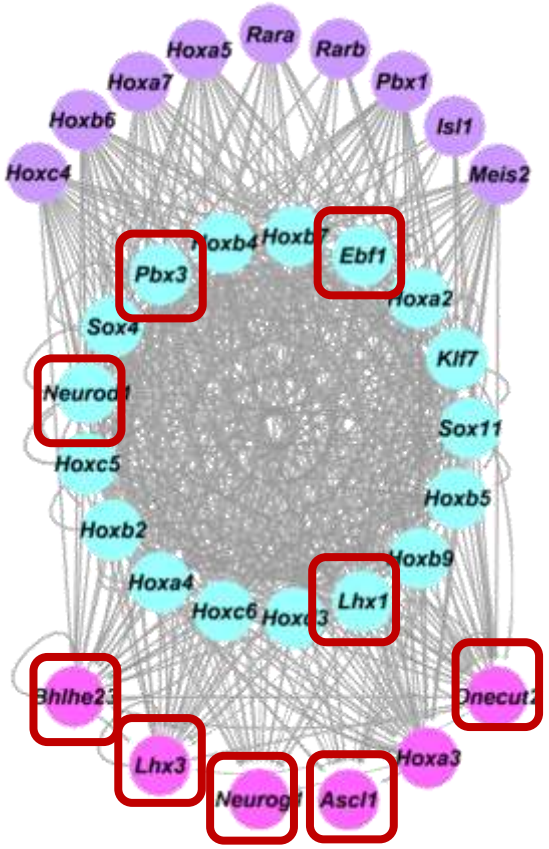


Cell type specific regulatory networks

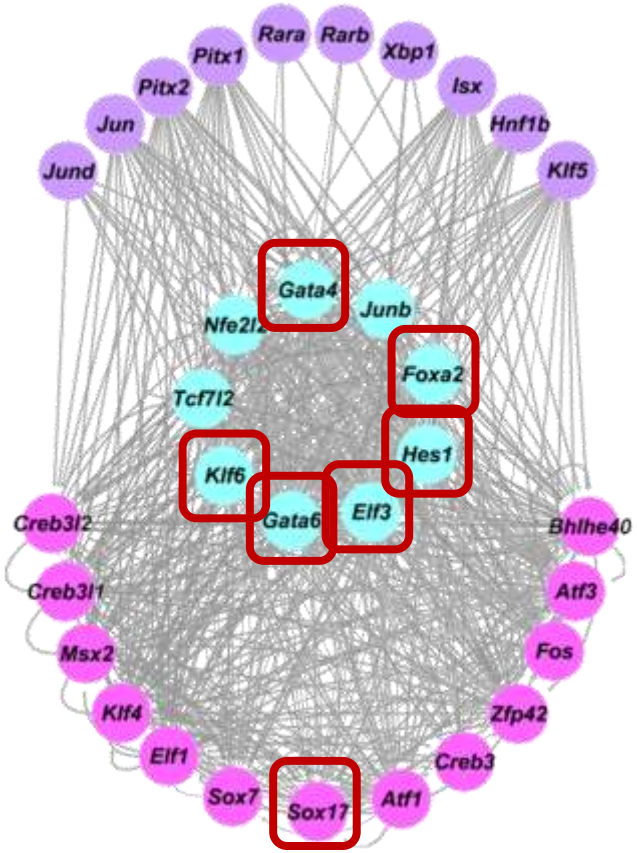
DC3

- Regulatory networks in the differentiation of mouse embryonic stem cells (mESC)

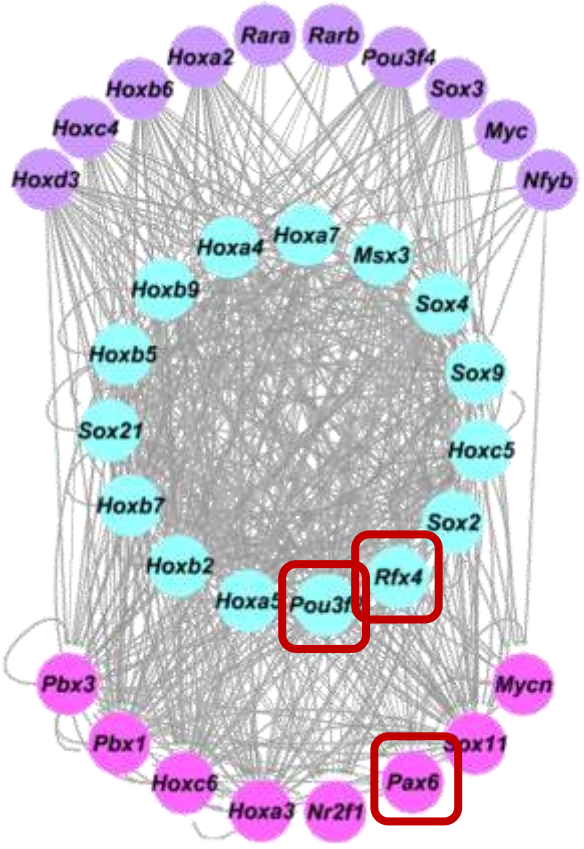
Neural commitment



Mesendoderm development



Brain development



Supervised learning

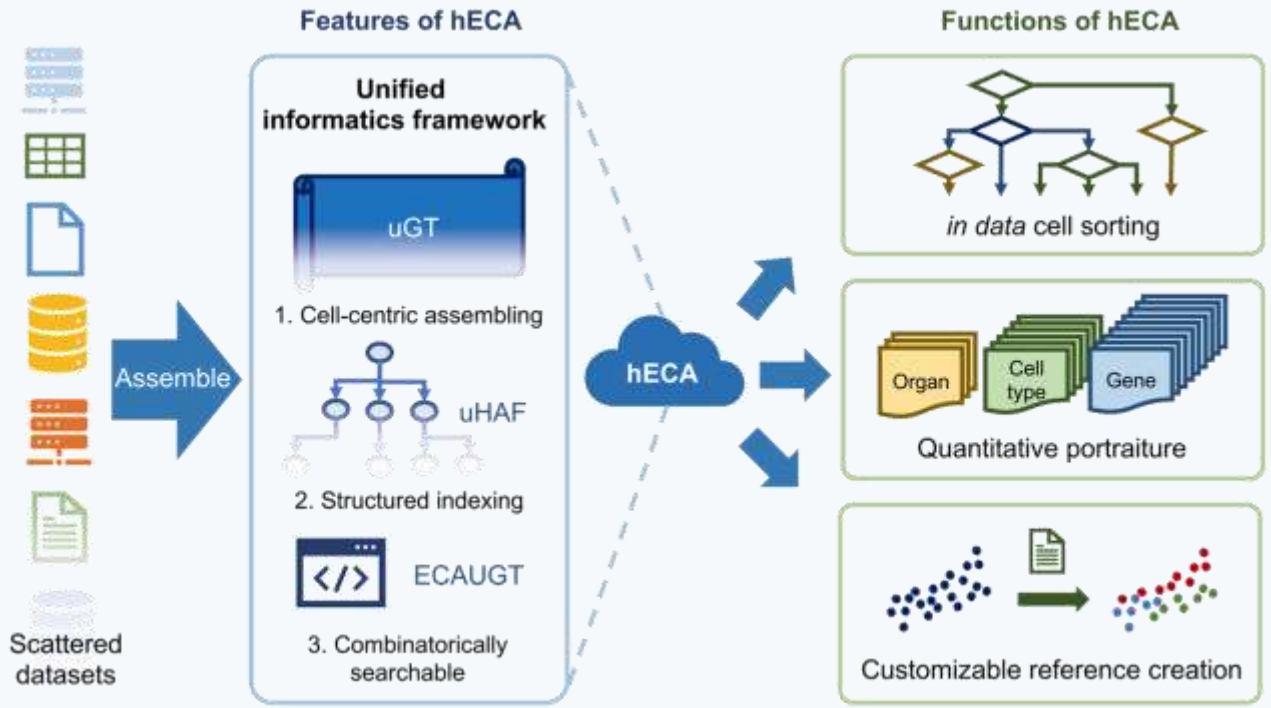
Weakly supervised

- Data of rough annotations
- Broad scope of application
- Due to imprecise annotation, only limited information is incorporated

Supervised

- Cell atlases often provide curated cell type annotations
- How to make use of such information to annotate known cell types?

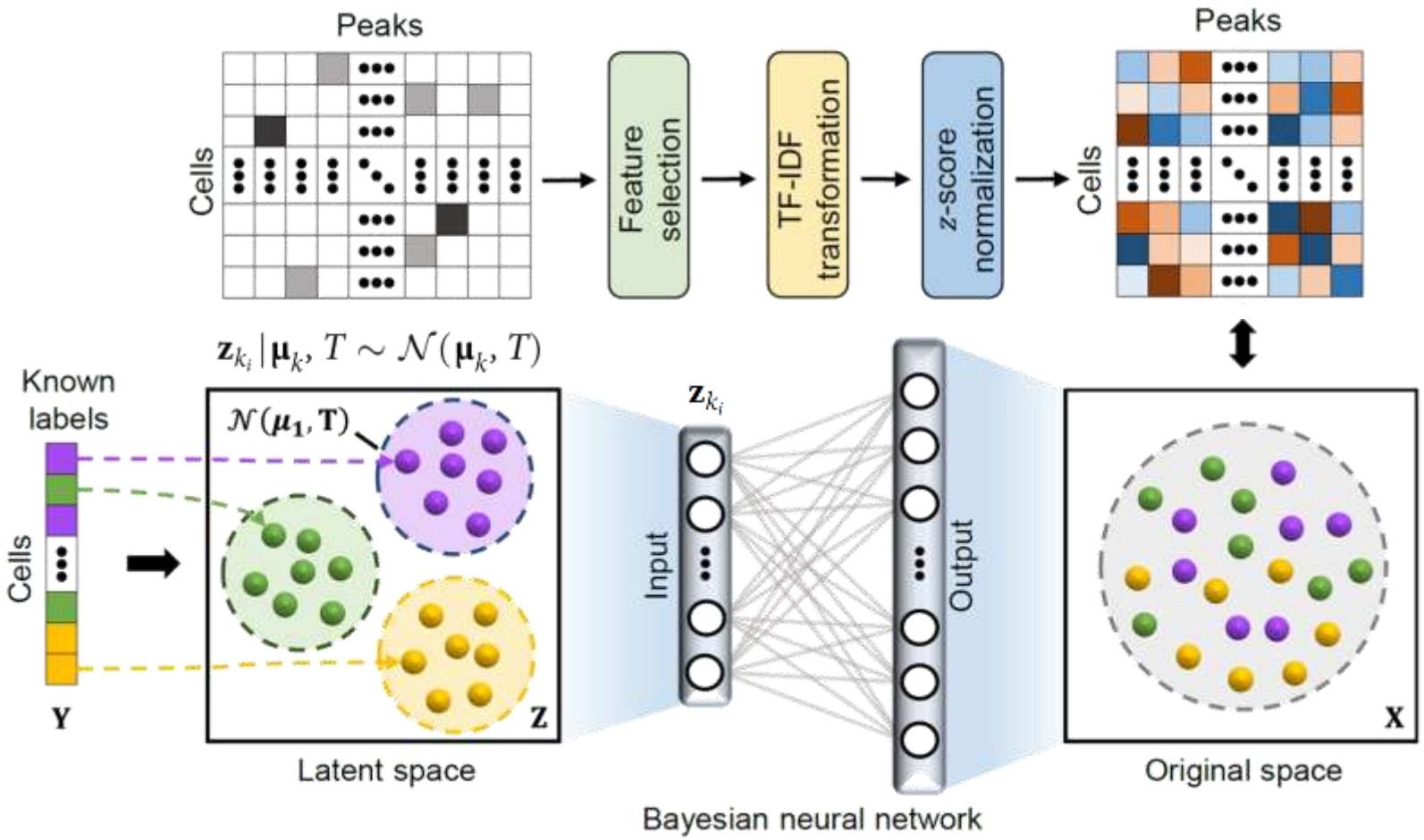
hECA: The cell-centric assembly of a cell atlas

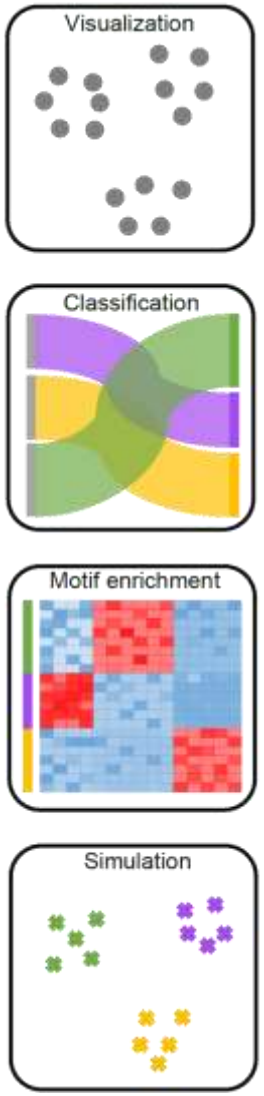


Bayesian neural network for cell type annotation **epiAnno**

- Generative model with a supervised training procedure
- Designed for scATAC-seq data

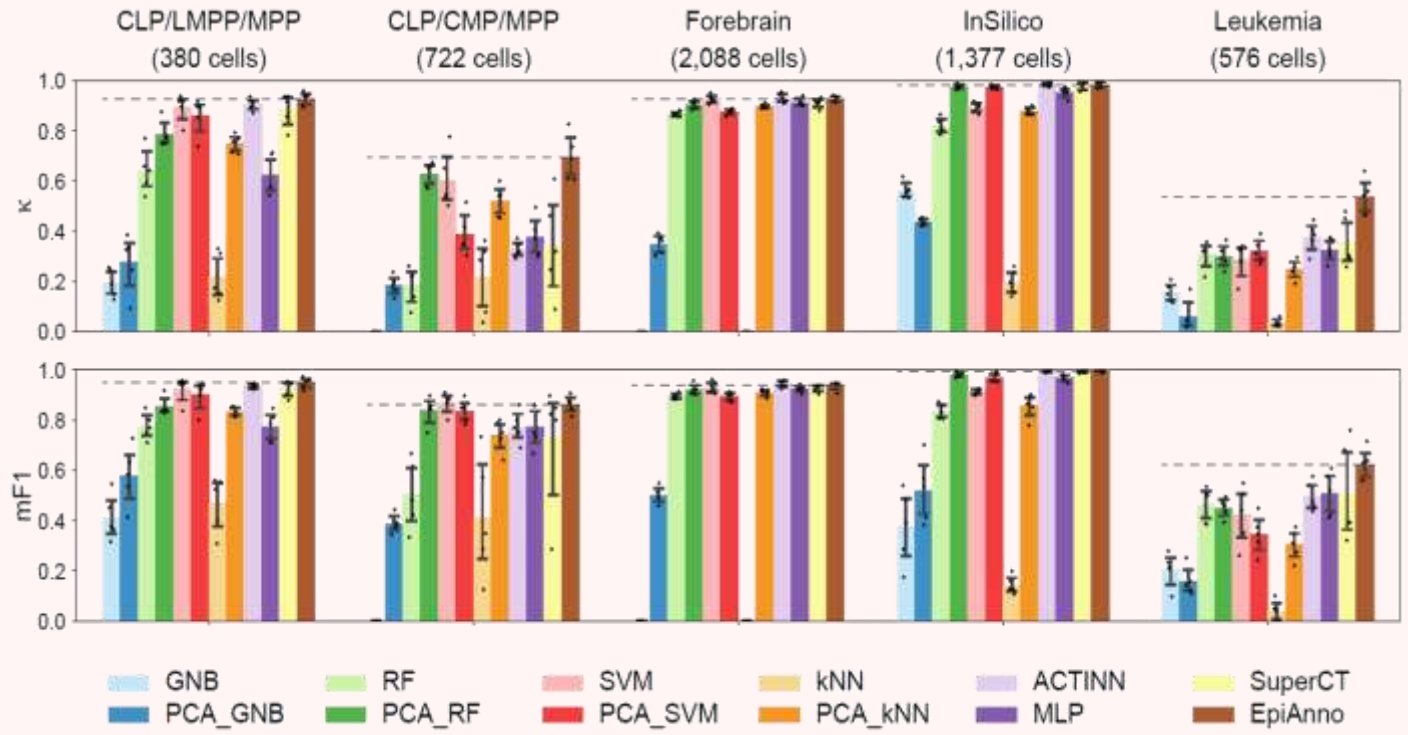
- **Chen et al, Nature Machine Intelligence, 2022, 4:116–126**



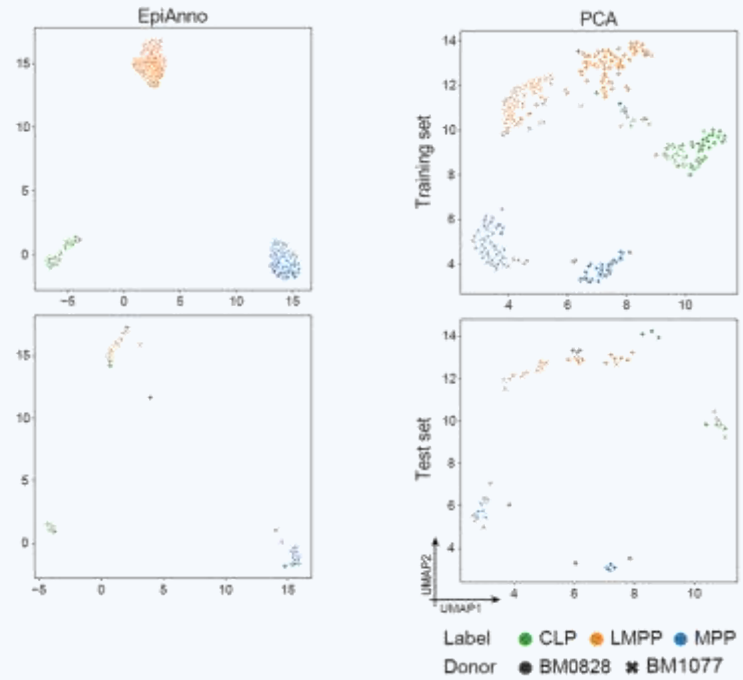


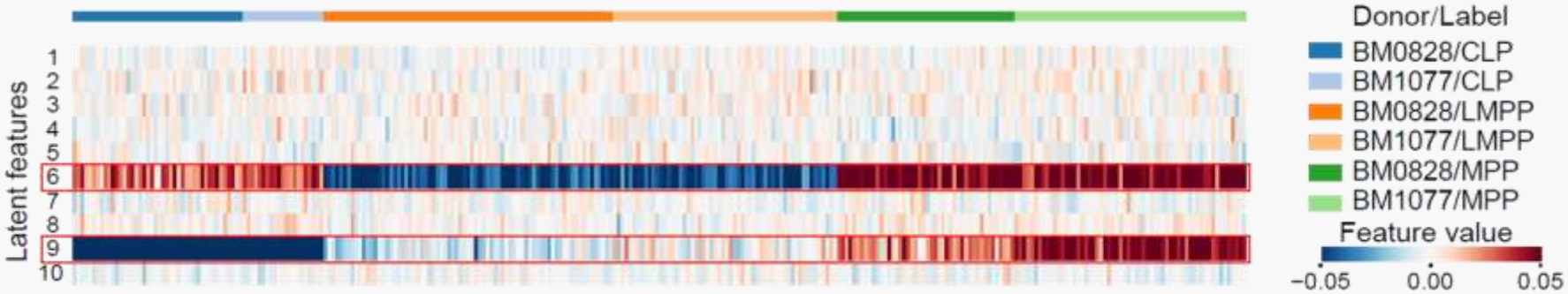
Accurate and interpretable cell type annotations **epiAnno**

- Performance is superior to 11 baseline methods



- Effectively remove batch effects

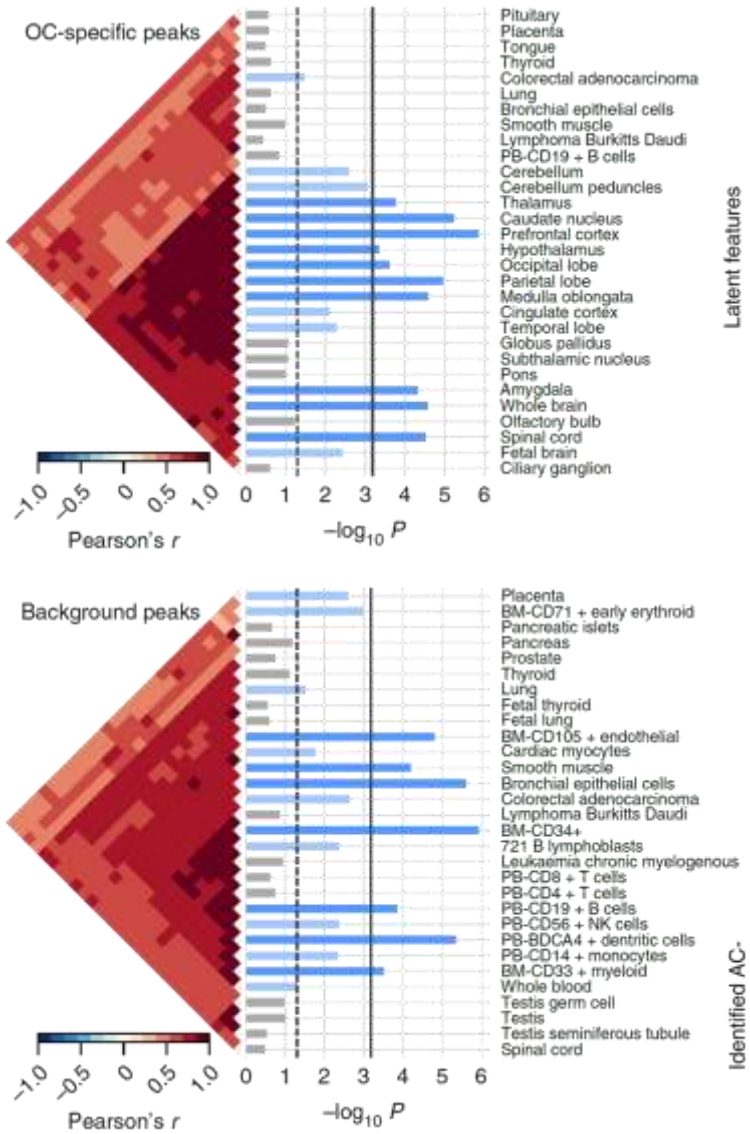




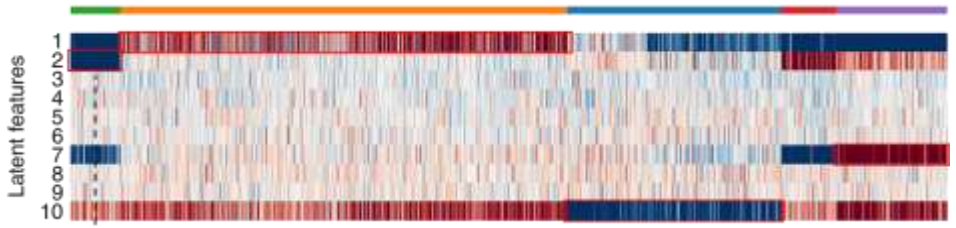
- Latent features have excellent interpretability

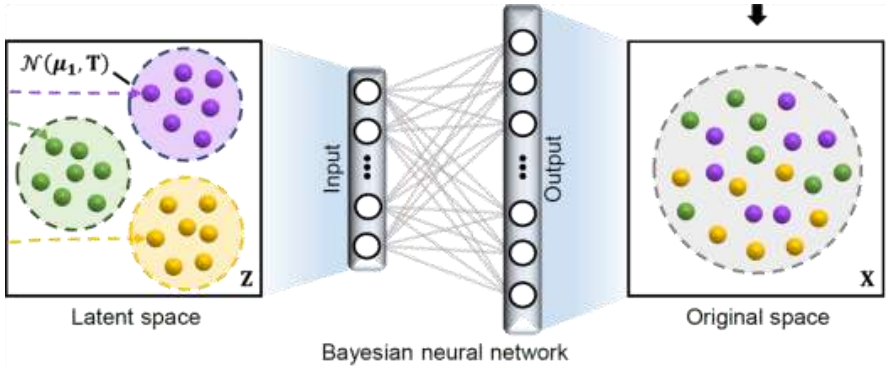
Tissue specificity of cell type specific peaks

epiAnno

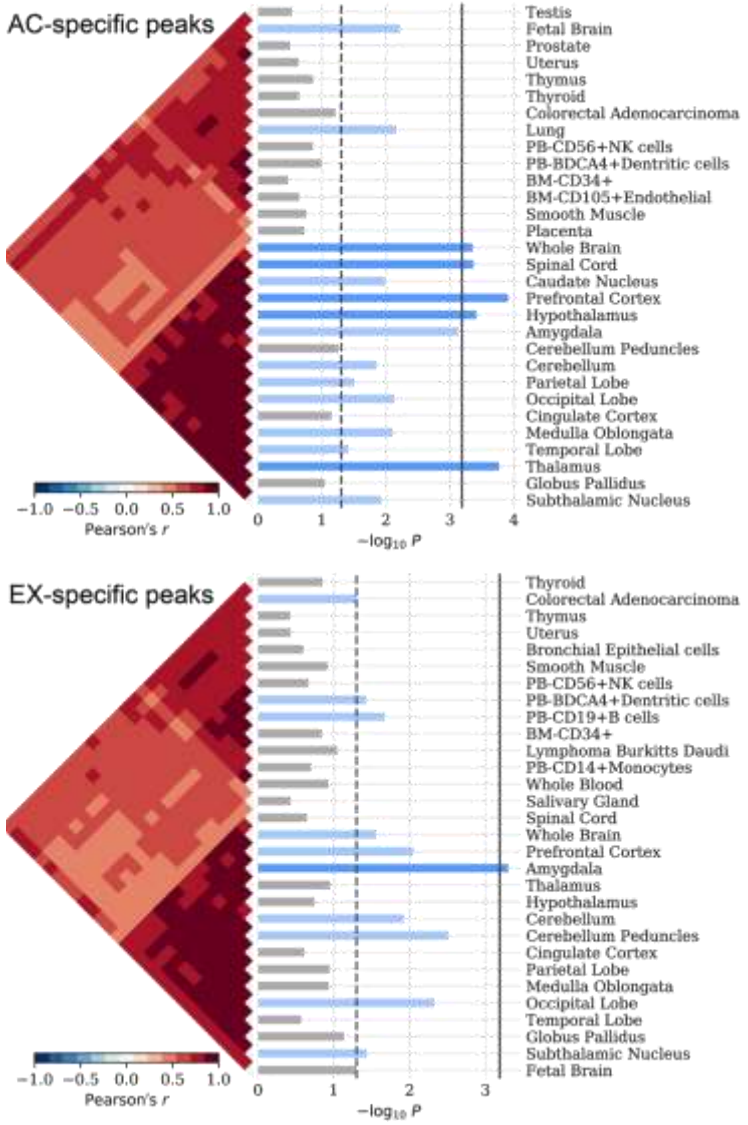


Latent feature 2 alone can well characterize astrocyte cells





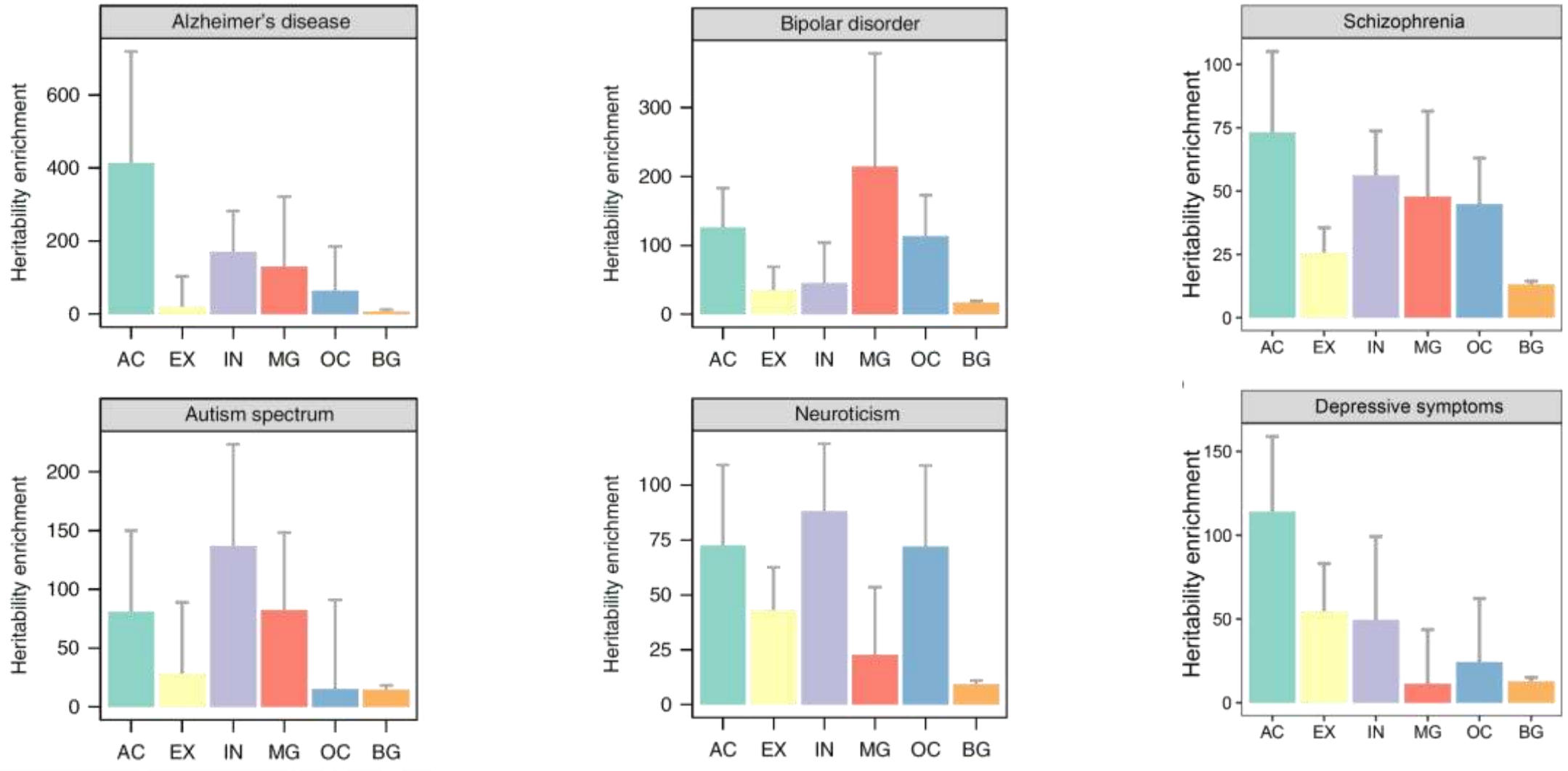
Chromatin accessible regions of astrocyte cells show specificity against those of other cell types



Disease association of cell type specific peaks

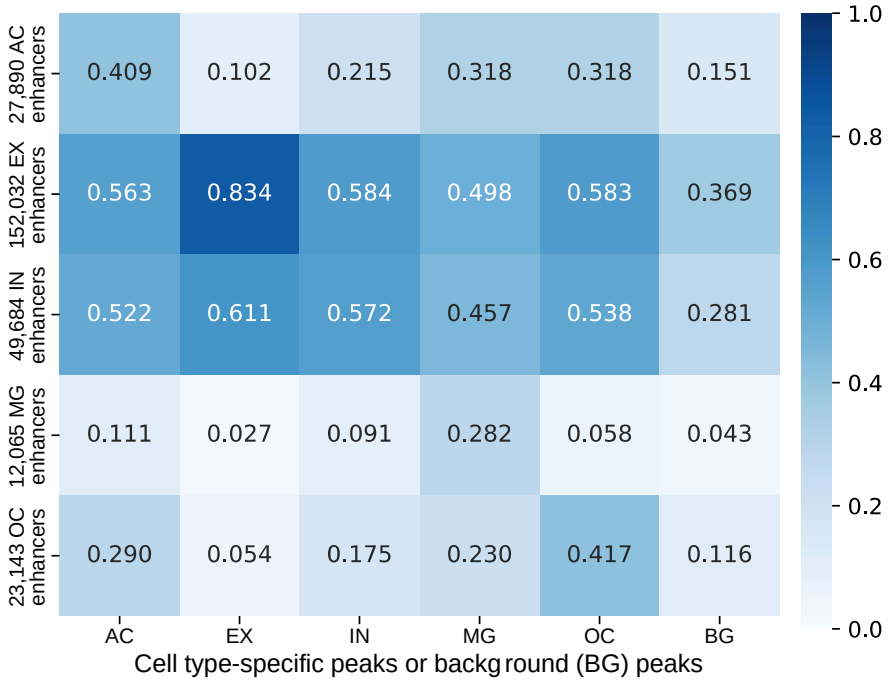
epiAnno

- SNPs associated with neurological diseases are enriched in peaks specific to brain-related cell types



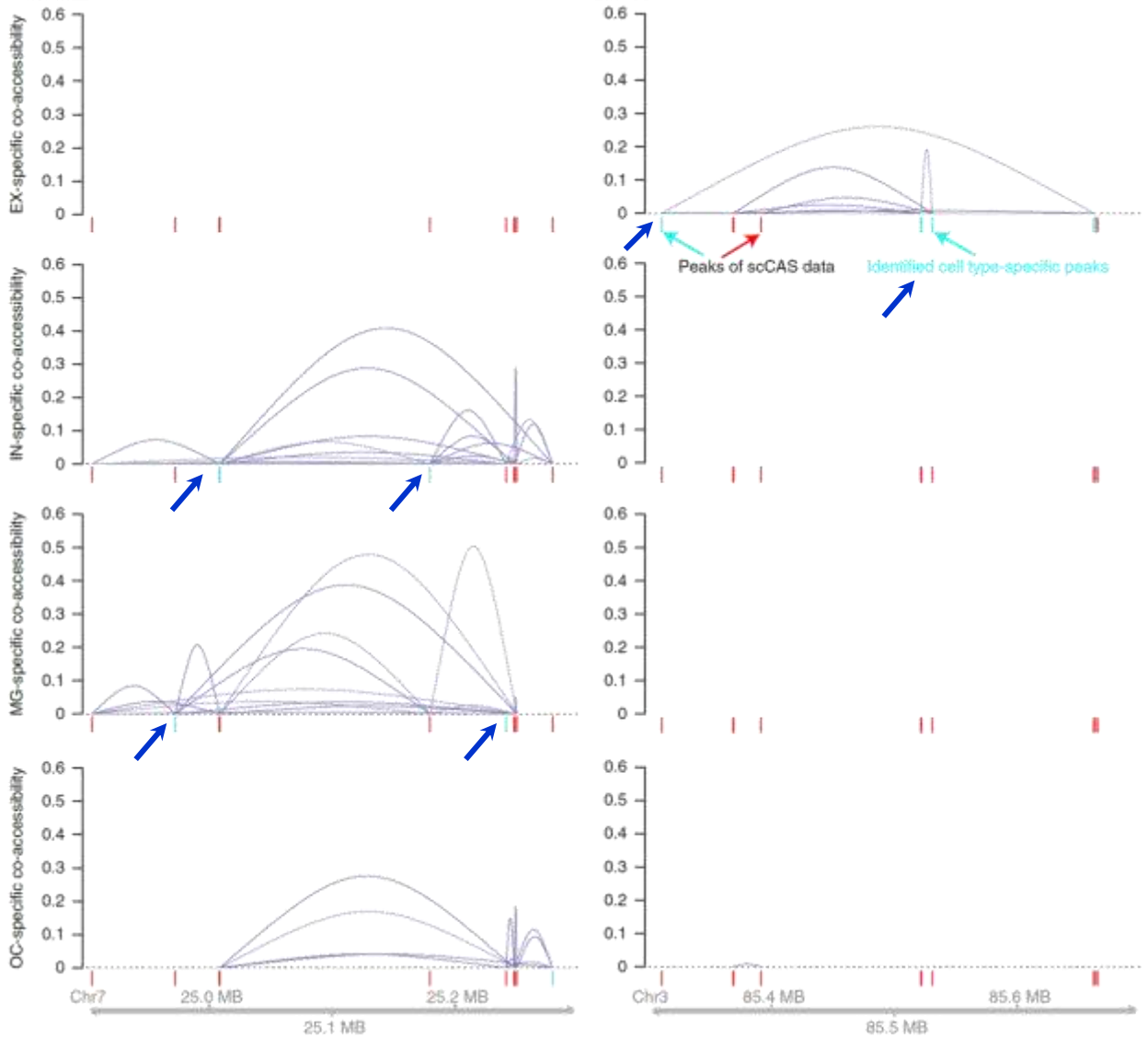
Regulatory elements and cell type specific peaks **epiAnno**

- Cell type specific overlapping with enhancers



Cell type specific peaks show higher overlap with enhancers specific to the same cell type

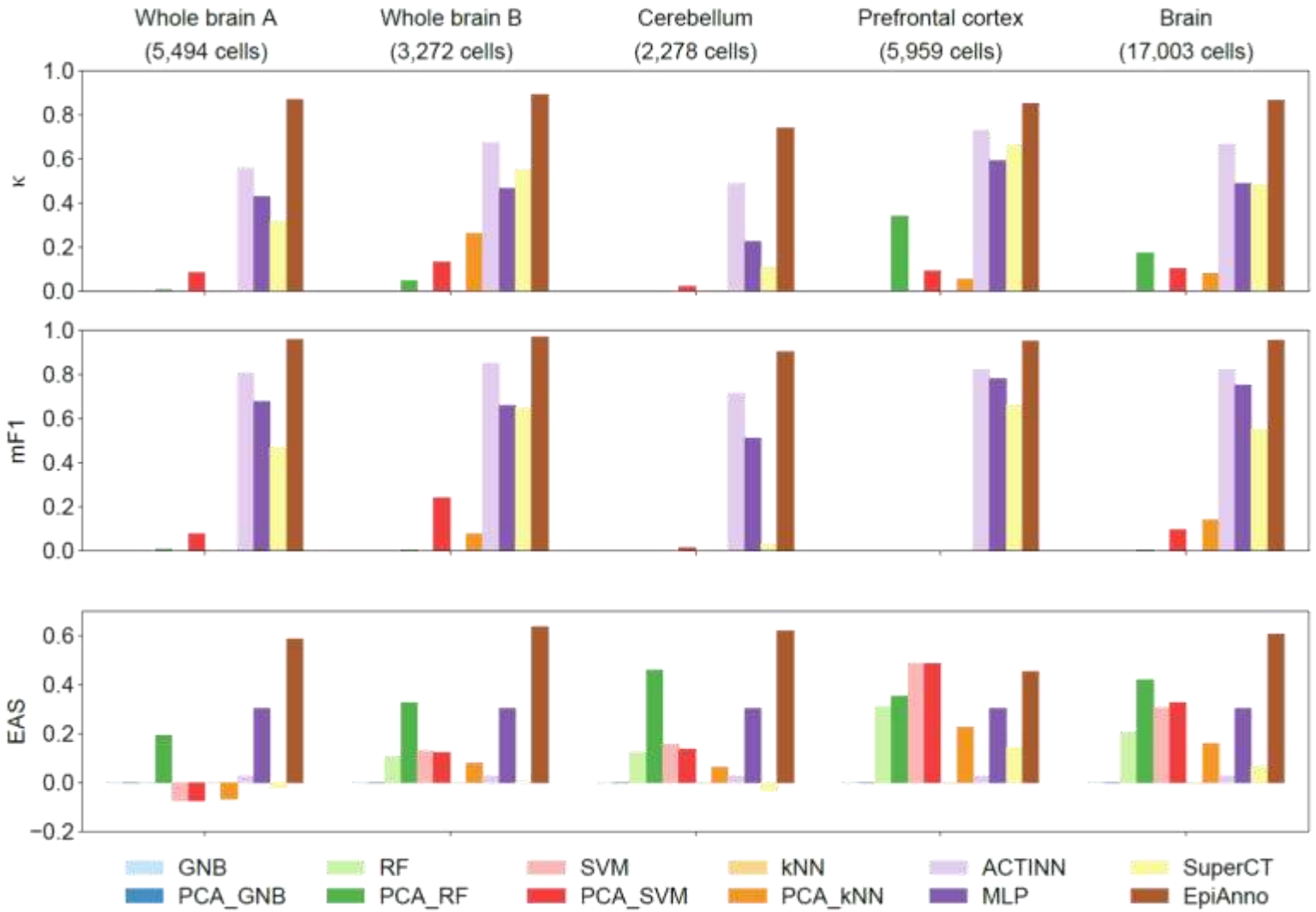
- Cell type specific interaction between peaks

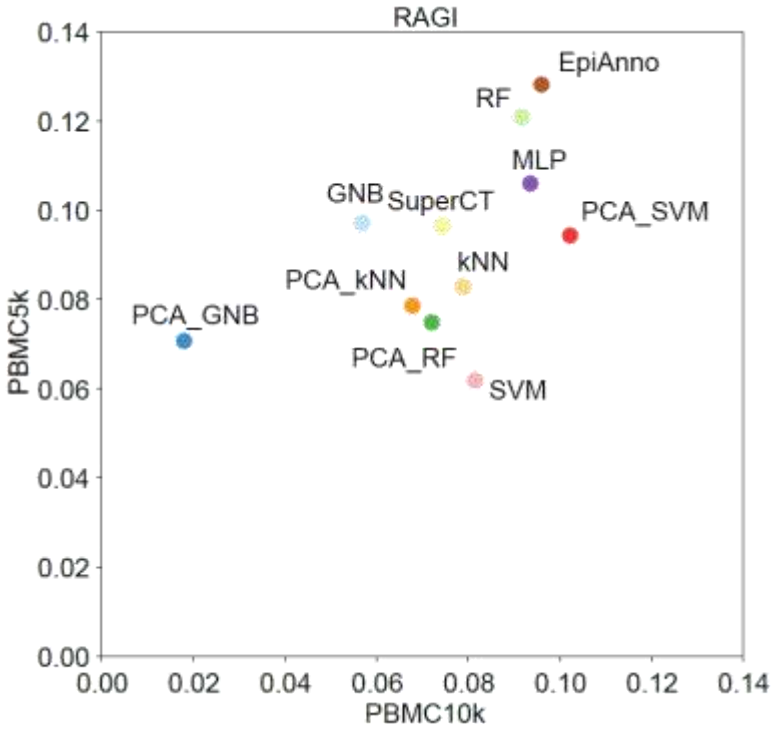


Accurate annotation of newly sequenced data

epiAnno

- Performance is superior to baseline methods

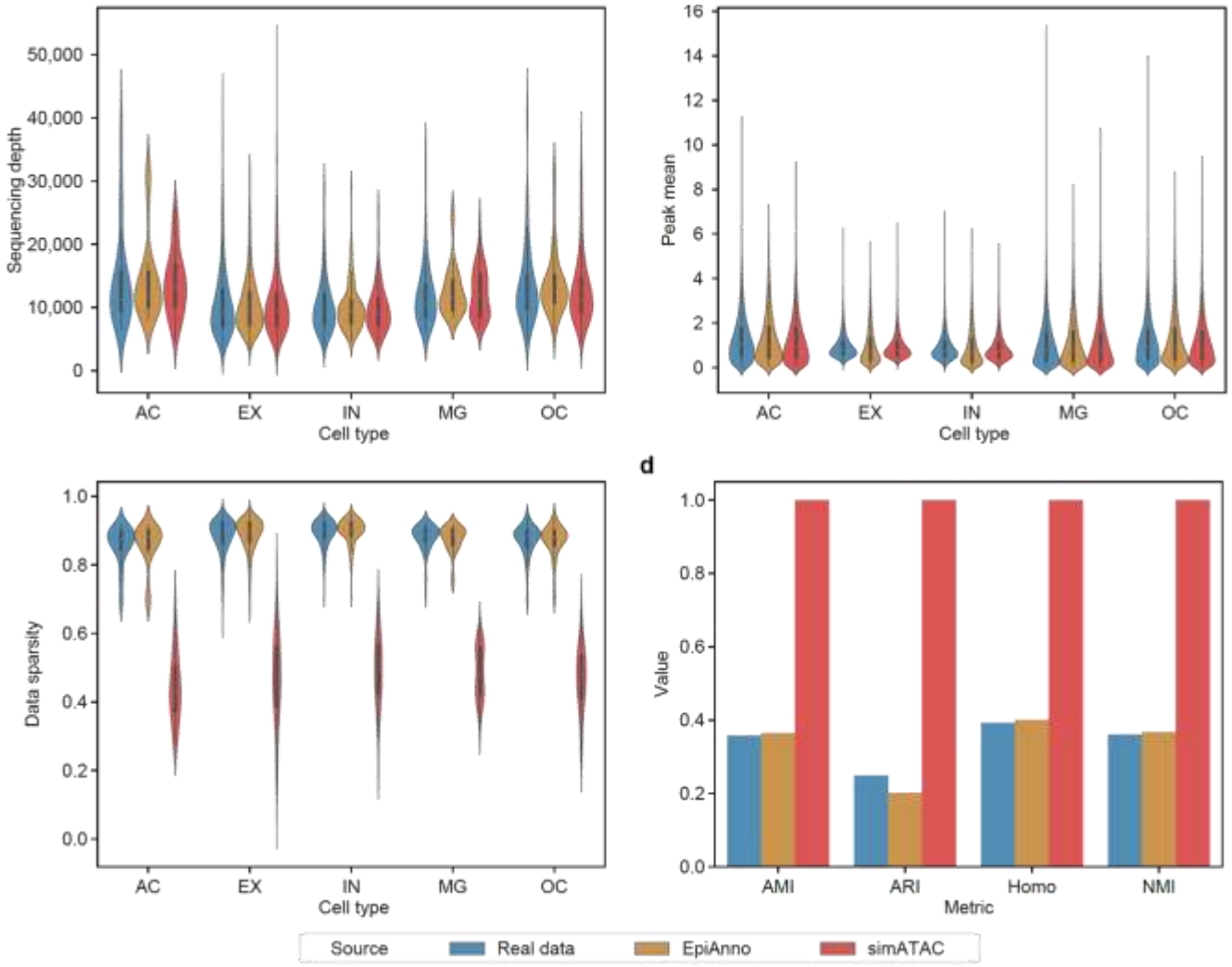


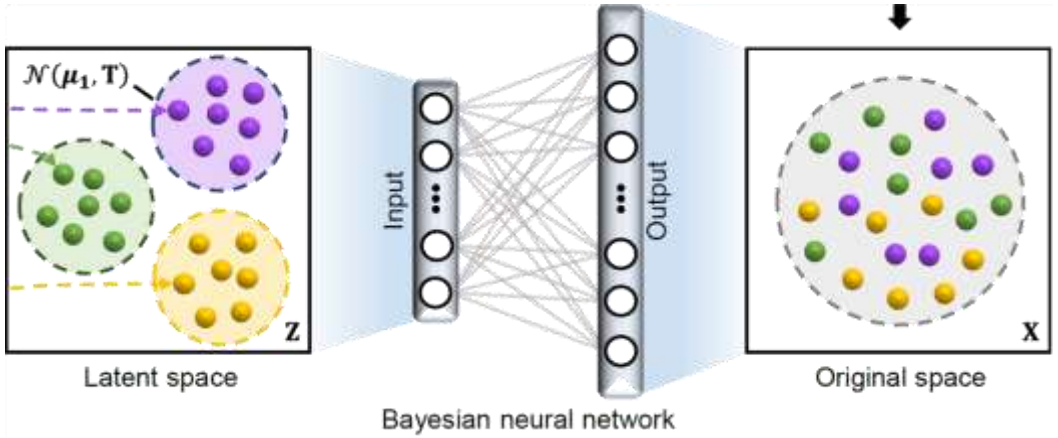


Simulation of scATAC-seq data

epiAnno

- Generated data show high similarity with real data





Deep learning methods for cell type identification based on single-cell chromatin accessibility data

Unsupervised	Weakly supervised	Supervised
Roundtrip scDEC VPAC	RA3 DC3 stPlus (Bioinformatics, 2021)	epiAnno scGraph (Bioinformatics, 2022)
Discover new cell types Simultaneous clustering and dimension reduction without prior knowledge	Discover new cell types Simultaneous clustering and dimension reduction with reference data	Annotate known cell types Simultaneous classification and dimension reduction with well annotated data
Higher requirement for data annotation		
Need manual curation of cell types	Need manual curation of cell types	Do not need manual curation of cell types
Higher requirement for manual curation		

Acknowledgements

Wing Hung Wong

Zhixiang Lin

Qiao Liu

Roundtrip
scDEC

Wanwen Zeng

DC3

Shengquan Chen

scDEC
RA3
epiAnno

Xiaoyang Chen

epiAnno

MOST

NHC

NSFC

Tsinghua University

Rui Jiang

Thank you very much!