Data handling for multi-modal single cell data using muon

EMBO Practical Course
Integrative analysis of multi-omics data

EMBL
September 21, 2022


Danila Bredikhin @gtcaa


Max Frank


Ilia Kats

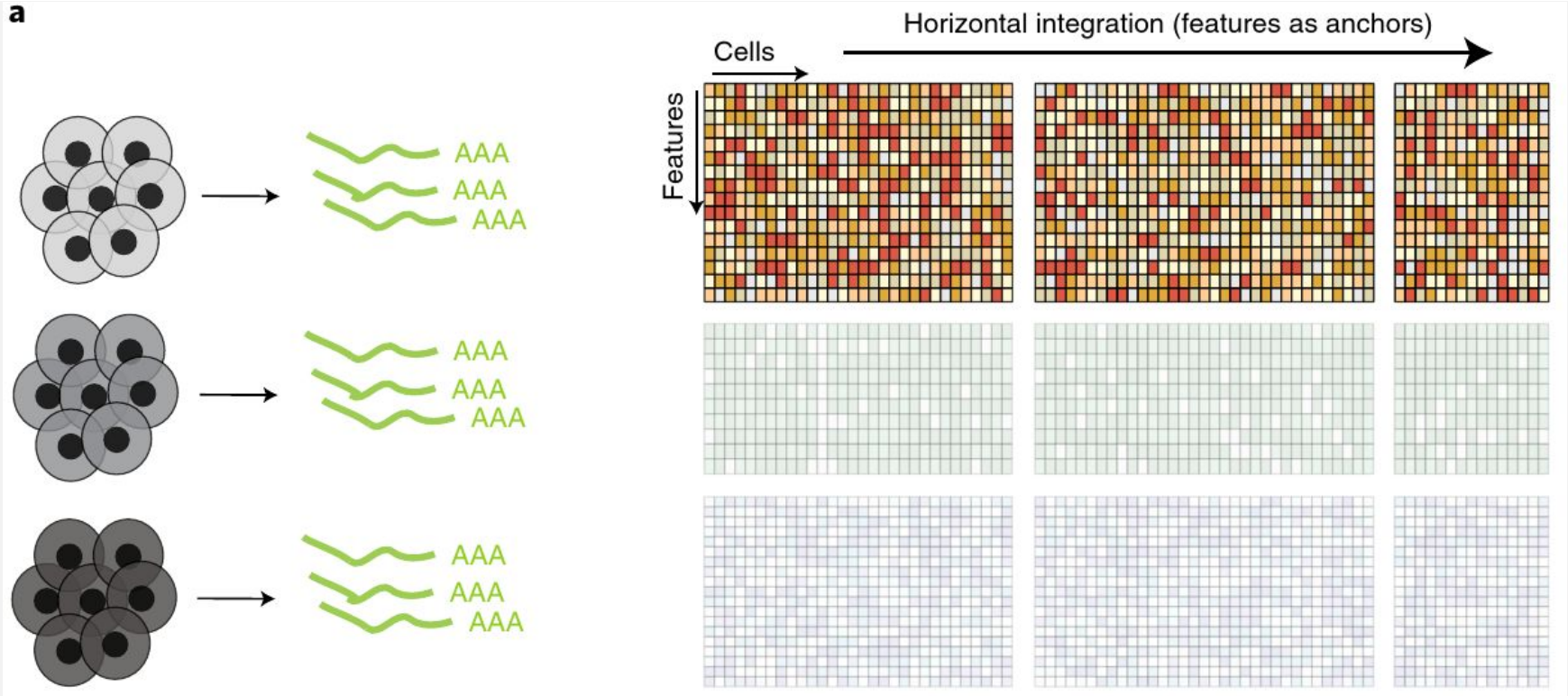
EMBL

Outline

1. Types of multimodal data integration
2. Single-cell data
3. Intro to AnnData and MuData
    a. Generic solution, primary focus is single-cell data
4. MuData for scATAC (peaks, windows), CITE-seq (prot counts), scNMT-seq data
5. Cross-language functionality (R)

6. Notebook: AnnData / MuData. Example MuData CITE-seq dataset.

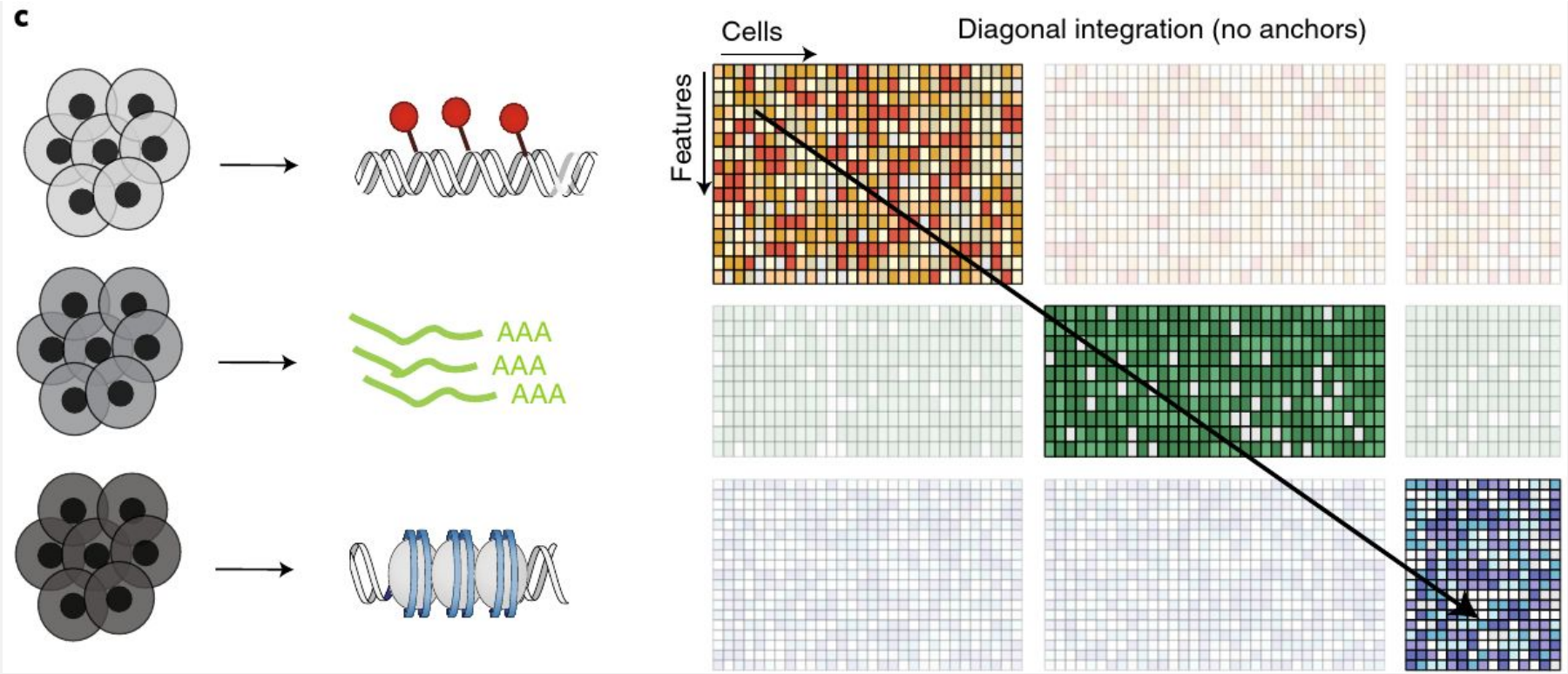
Three types of data integration



Argelaguet et al., 2021

EMBL

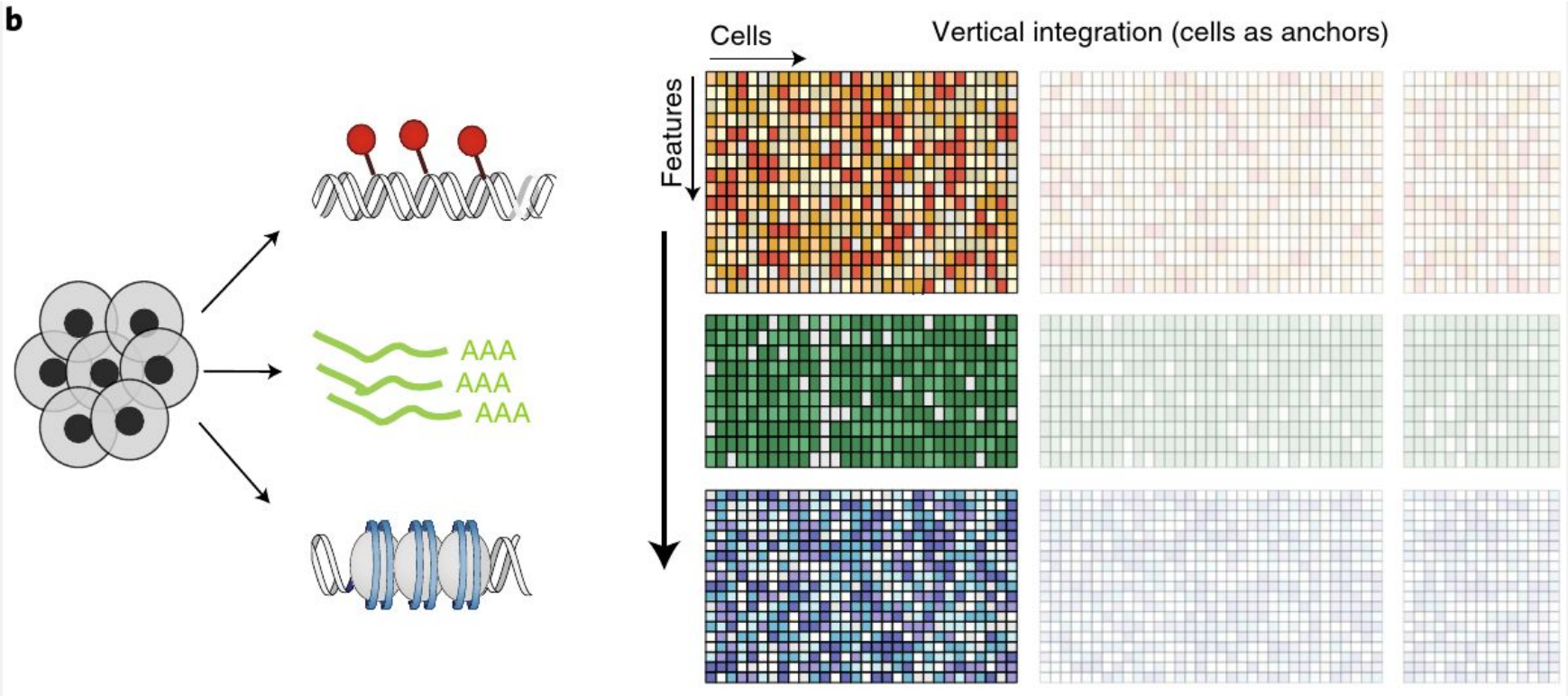
Three types of data integration



Argelaguet et al., 2021

EMBL

Three types of data integration



Argelaguet et al., 2021

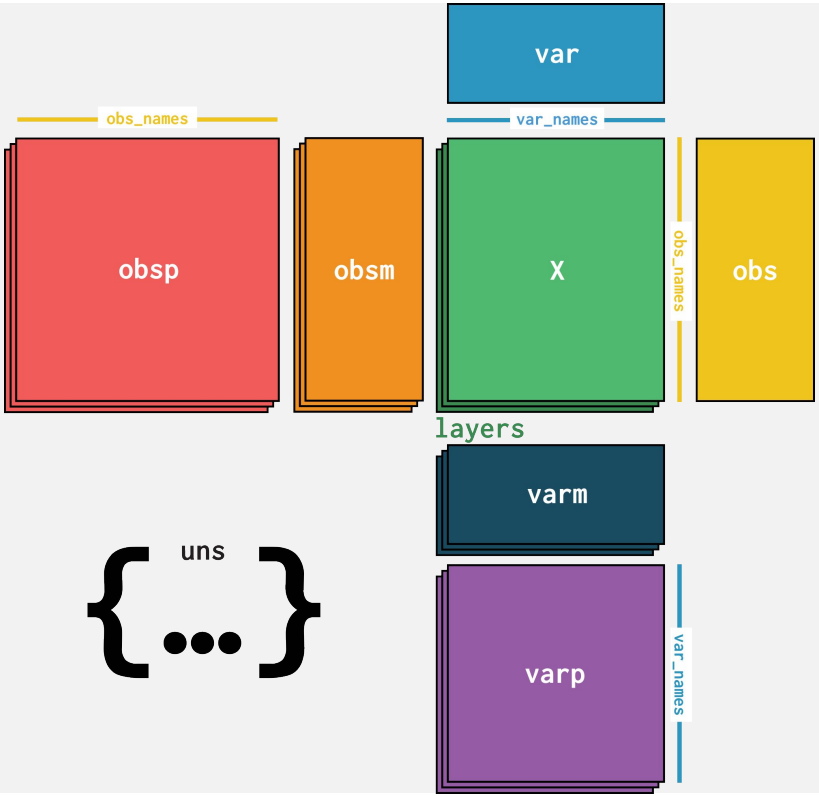
EMBL

Single-cell vs bulk

Compared to bulk (classical) sequencing data, single-cell data is

- Much sparser: many features are not detected in many cells (drop-outs)
- Noisier: Coefficient of variation for Poisson sequencing noise decreases with sequencing depth
- Contains more technical artifacts: doublets, ambient RNA

Data structures for single-cell data: AnnData



Count matrix with metadata

- `X`: count matrix
- `var`: DataFrame with gene metadata
- `var_names`: index of `var` with gene names
- `obs`: DataFrame with observation (cell) metadata
- `obs_names`: index of obs with (usually) cell barcodes
- `varm`, `obsm`: dictionaries of arrays or DataFrames, aligned to `var` and `obs`
- `varp`, `obsp`: dictionaries of arrays with pairwise annotation, e.g. pairwise distances
- `uns`: dictionary with arbitrary data

https://anndata.readthedocs.io

EMBL

Data structures for single-cell data: AnnData

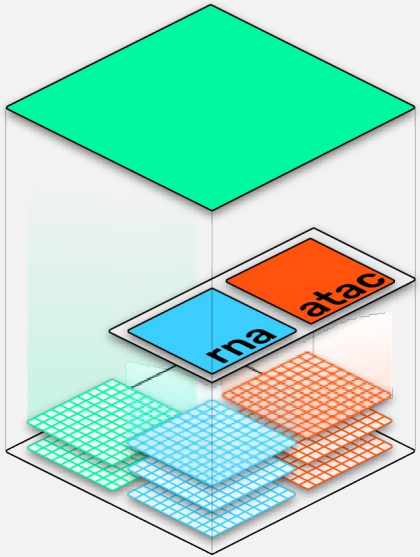
```
Python 3.10.6 (main, Aug  3 2022, 17:39:45) [GCC 12.1.1 20220730]
Type 'copyright', 'credits' or 'license' for more information
IPython 8.5.0 -- An enhanced Interactive Python. Type '?' for help.

In [1]: import anndata
   ...: import numpy as np
   ...: from scipy.sparse import csr_matrix
   ...: counts = csr_matrix(np.random.poisson(1, size=(100, 2000)))
   ...: adata = anndata.AnnData(counts, dtype=np.float32)
   ...: adata
   ...:
Out[1]: AnnData object with n_obs × n_vars = 100 × 2000
```

```
In [6]: adata
Out[6]:
AnnData object with n_obs × n_vars = 100 × 2000
    obs: 'cell_type'
```

EMBL

Data structures for single-cell multimodal data: MuData



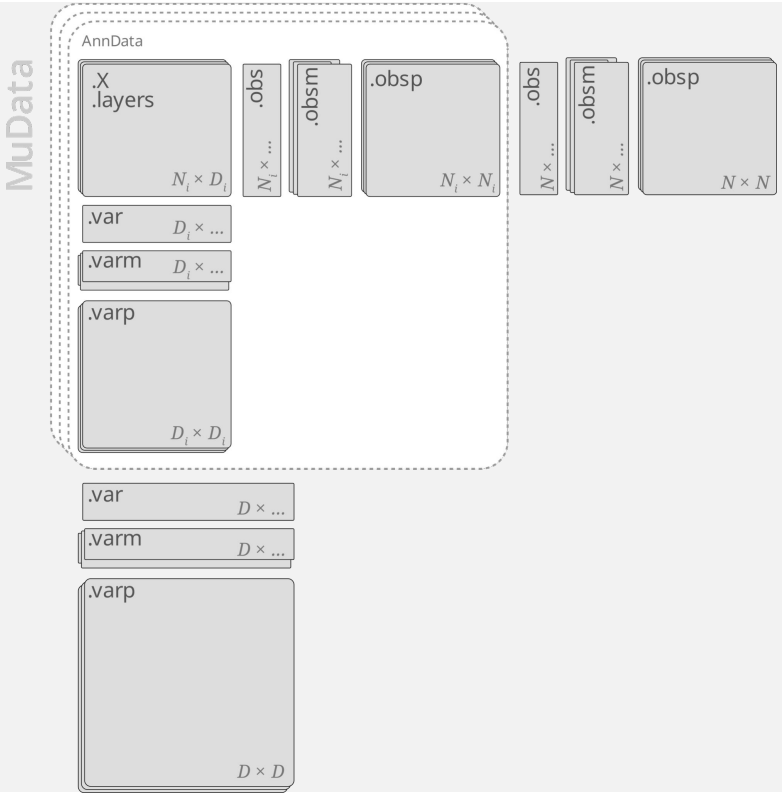
MuData –
a multimodal
data container

Individual
modalities

Count matrices
Metadata
Derived data

Dictionary of AnnData objects

Data structures for single-cell multimodal data: MuData



- AnnData objects are aligned to observations
- Assumption: each AnnData has different variables, but observations can overlap
- Joint subsetting of AnnData objects
- Additional metadata on the global level
- Easy access to individual modalities

EMBL

Data structures for single-cell multimodal data: MuData

```
In [20]: import mudata as md

In [21]: bdata.var_names = [f"Protein_{i}" for i in range(bdata.n_vars)]

In [22]: mudata = md.MuData({"rna": adata, "prot": bdata})

In [23]: mudata
Out[23]:
MuData object with n_obs × n_vars = 100 × 4000
  2 modalities
    rna:        100 x 2000
      obs:      'cell_type'
      layers:   'log_transformed'
    prot:       35 x 2000
      obs:      'cell_type'
      layers:   'log_transformed'
```

EMBL

1. Gene expression
2. ...
3. ...
4. ...
5. ...
6. ...

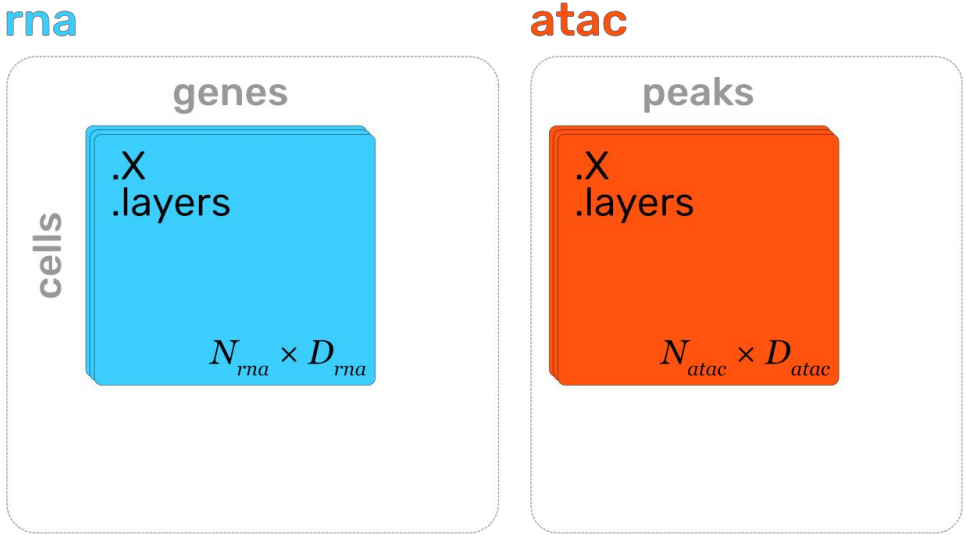
Multimodal single-cell datasets

1. Gene expression
2. Chromatin accessibility
3. DNA methylation
4. Histone modifications
5. Chromatin conformation
6. Protein abundance (intranuclear, surface)

Multimodal single-cell datasets

MuData for ...

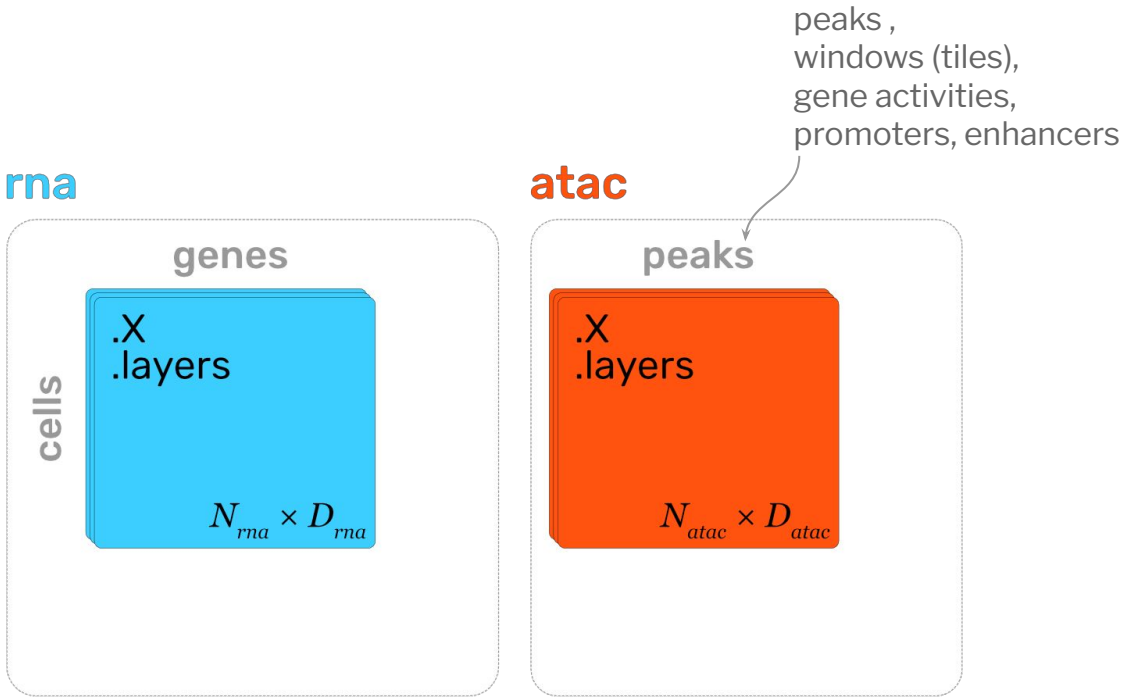
1. scRNA+scATAC-seq
   a. SNARE-seq
   b. SHARE-seq
   c. 10x Multiome
2. CITE-seq
3. scNMT-seq



EMBL

MuData for ...

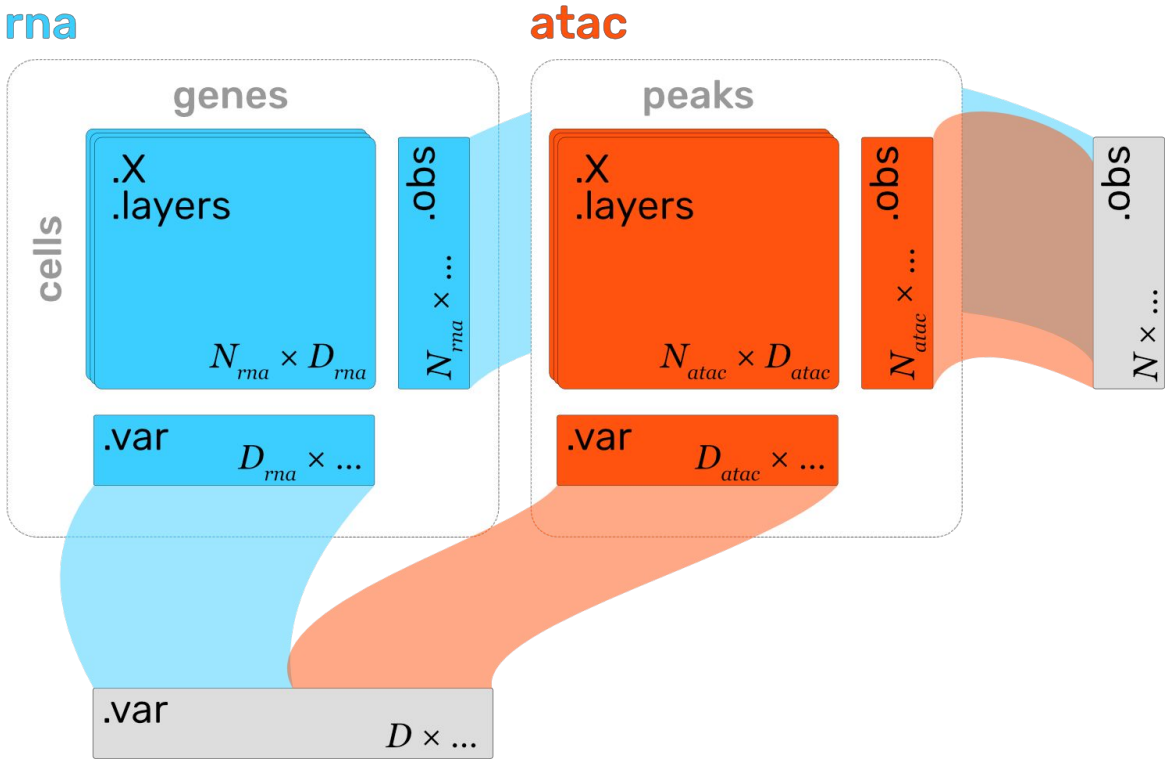
1. scRNA+scATAC-seq
   a. SNARE-seq
   b. SHARE-seq
   c. 10x Multiome
2. CITE-seq
3. scNMT-seq



EMBL

MuData for ...

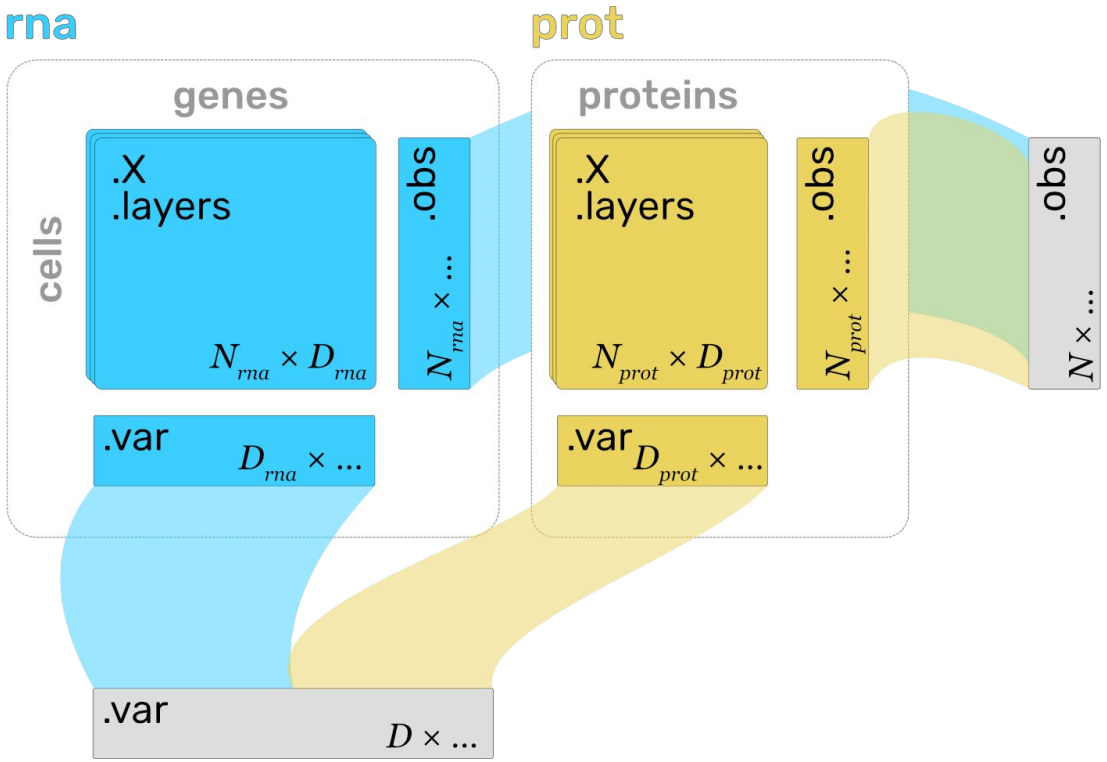
1. scRNA+scATAC-seq
   a. SNARE-seq
   b. SHARE-seq
   c. 10x Multiome
2. CITE-seq
3. scNMT-seq



EMBL

MuData for ...

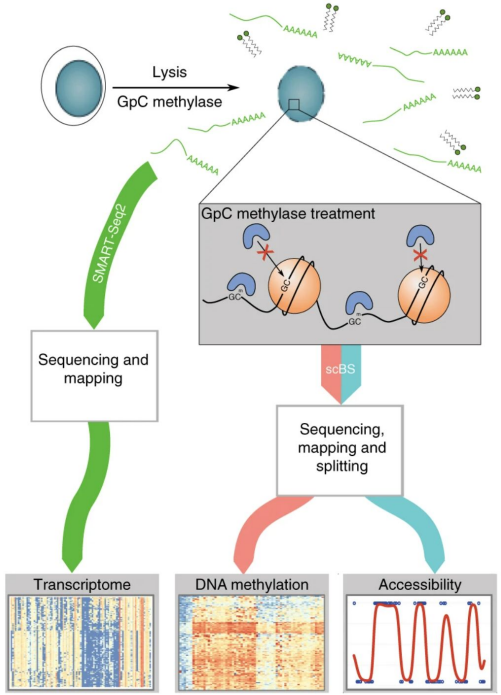
1. scRNA+scATAC-seq
   a. SNARE-seq
   b. SHARE-seq
   c. 10x Multiome
2. CITE-seq
3. scNMT-seq

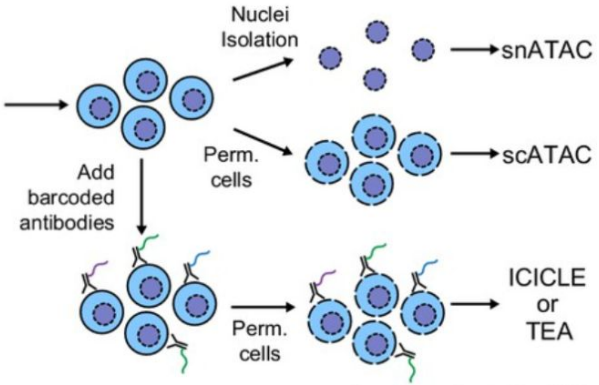


EMBL

MuData for ...

1. scRNA+scATAC-seq
   a. SNARE-seq
   b. SHARE-seq
   c. 10x Multiome
2. CITE-seq
3. scNMT-seq, TEA-seq, ...





Clark, Argelaguet et al., 2018; Swanson et al., 2021

EMBL

Language-agnostic data serialisation

1. **.h5ad/.h5mu**
2. **MuData(MAE)**
3. MuDataSeurat
4. Muon.jl

```
> library(MuData)
> (mdata ← readH5MU("pbmc5k_citeseq.h5mu", backed = TRUE))
# A MultiAssayExperiment object of 2 listed
#  experiments with user-defined names and respective classes.
#  Containing an ExperimentList class object of length 2:
#  [1] prot: SingleCellExperiment with 32 rows and 3891 columns
#  [2] rna: SingleCellExperiment with 17806 rows and 3891 columns
> class(assays(mdata[["prot"]])[['counts']])
# [1] "DelayedMatrix"
# attr(,"package")
# [1] "DelayedArray"
```

EMBL

Language-agnostic data serialisation

1. **.h5ad/.h5mu**
2. MuData(MAE)
3. **MuDataSeurat**
4. Muon.jl

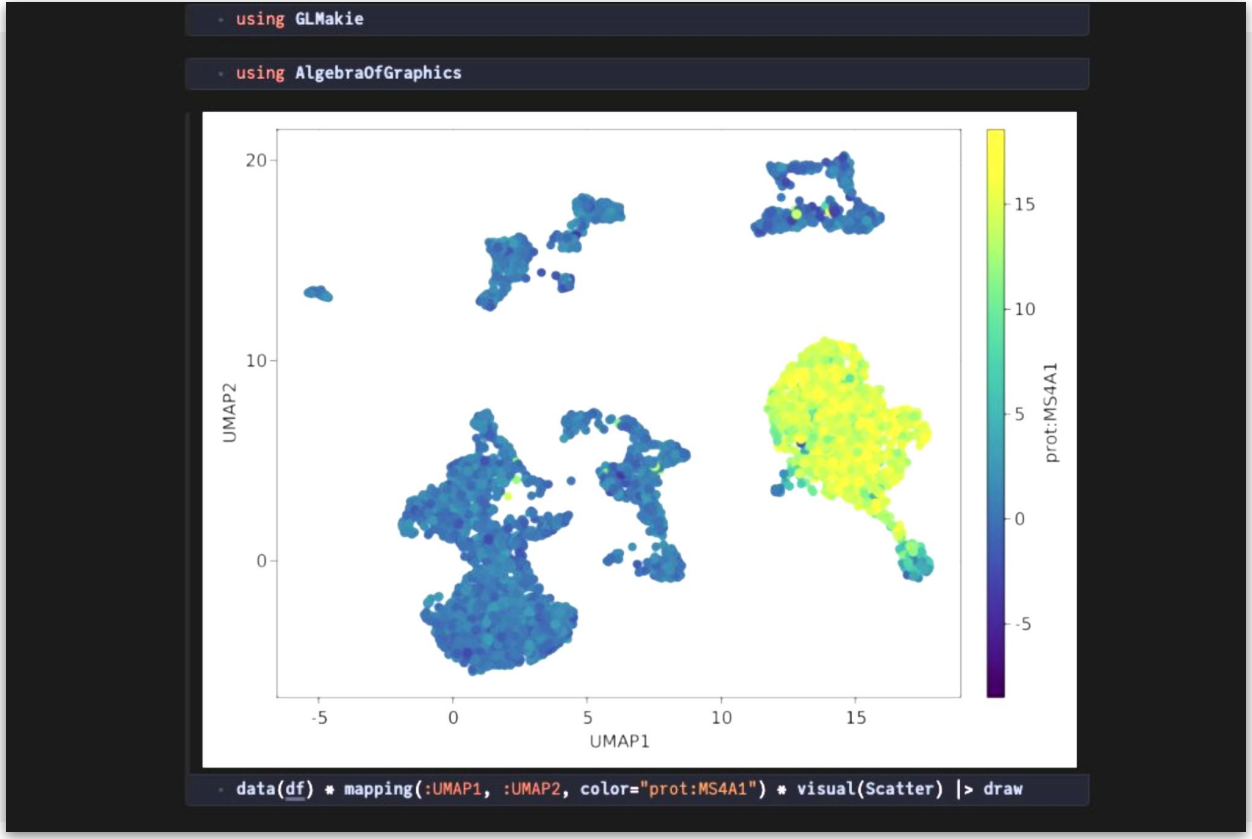
```
library(MuDataSeurat)
(mdata <- ReadH5MU("pbmc5k_citeseq.h5mu"))
# An object of class Seurat
# 17838 features across 3891 samples within 2 assays
# Active assay: prot (32 features, 32 variable features)
#  1 other assay present: rna
#  6 dimensional reductions calculated: MOFA, UMAP, protPCA, #
protUMAP, rnaPCA, rnaUMAP
```

Language-agnostic data serialisation

1. **.h5ad/.h5mu**
2. MuData(MAE)
3. MuDataSeurat
4. **Muon.jl**

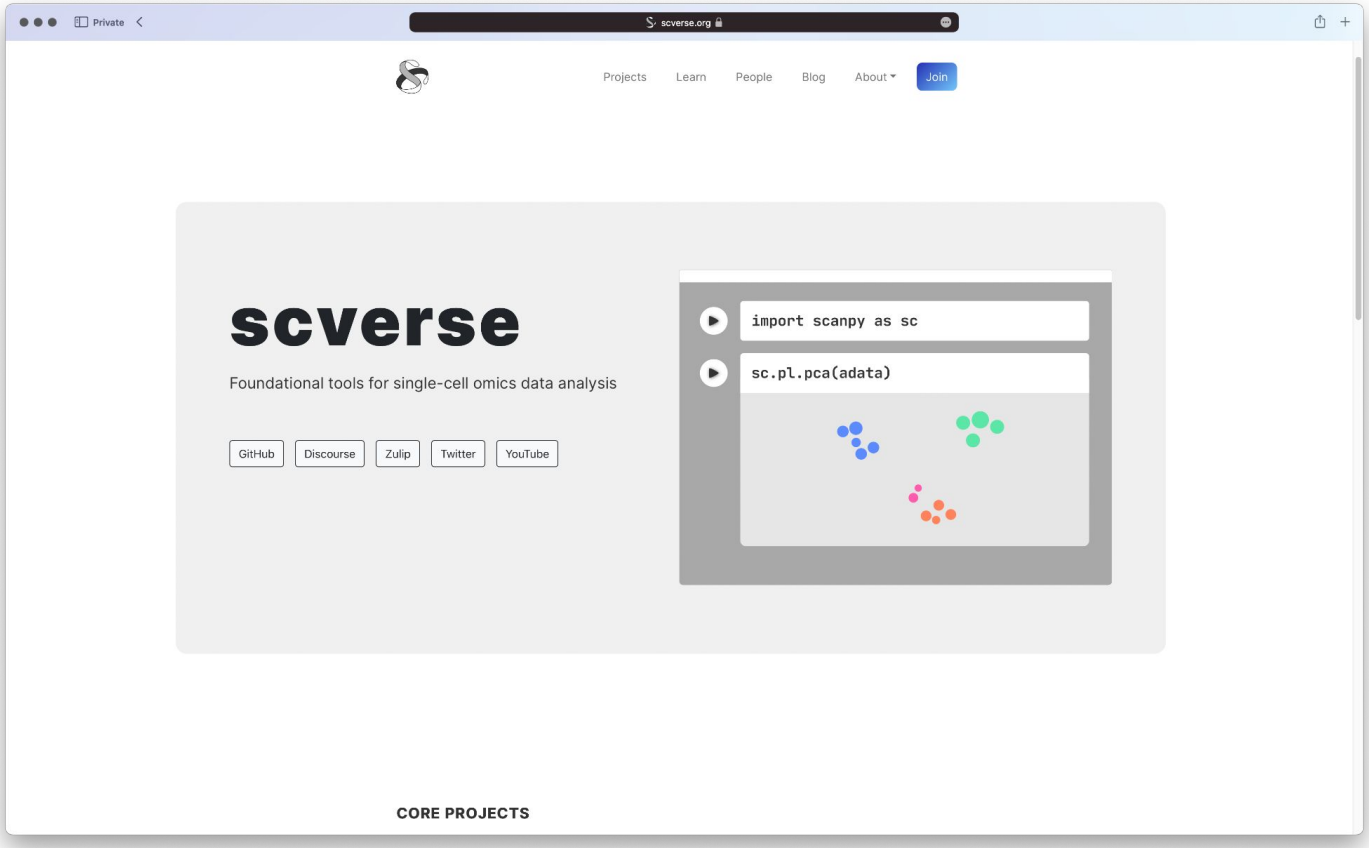
Language-agnostic data serialisation

1. **.h5ad/.h5mu**
2. MuData(MAE)
3. MuDataSeurat
4. **Muon.jl**



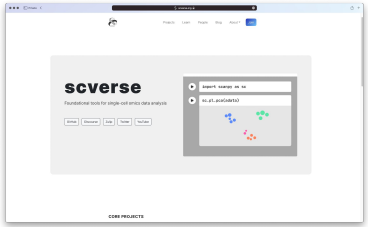
EMBL

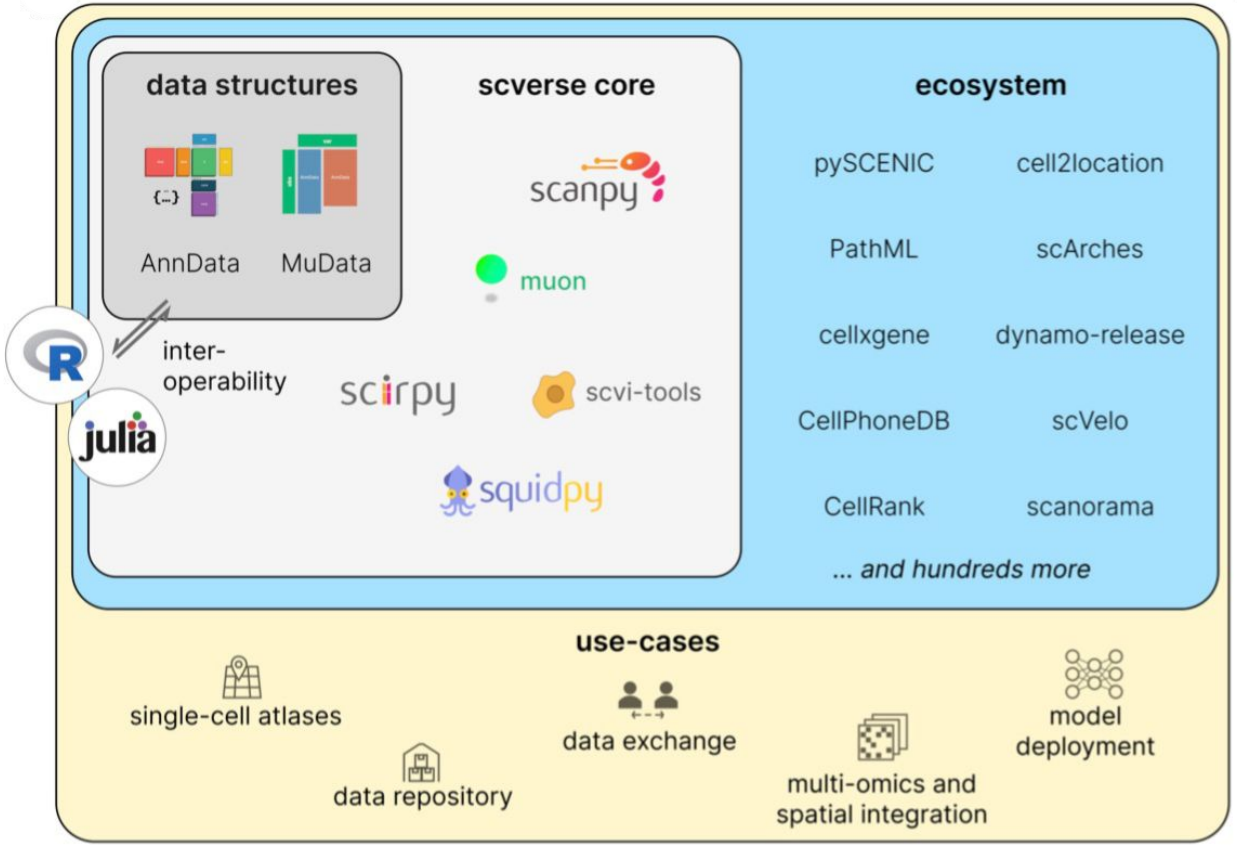
scverse



EMBL

scverse

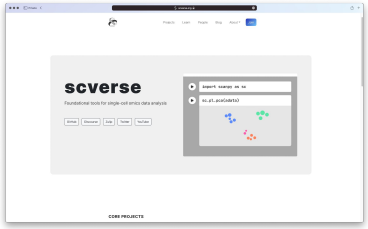


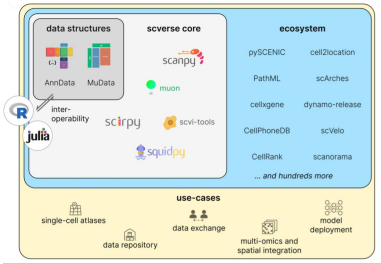


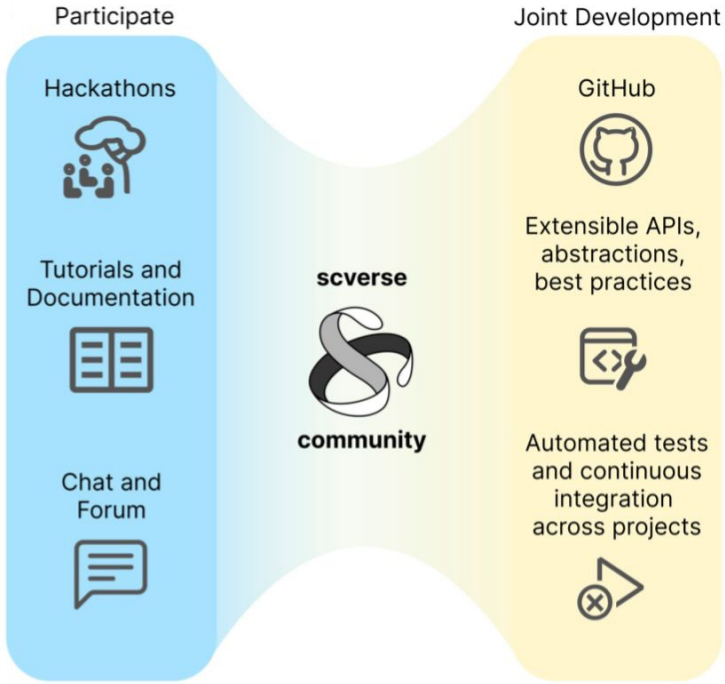
EMBL

scverse

scverse.org github.com/scverse scverse.zulipchat.com
twitter.com/scverse_team discourse.scverse.org/

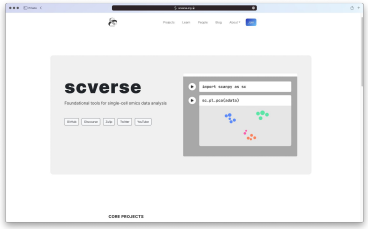


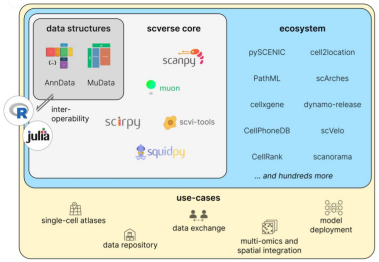


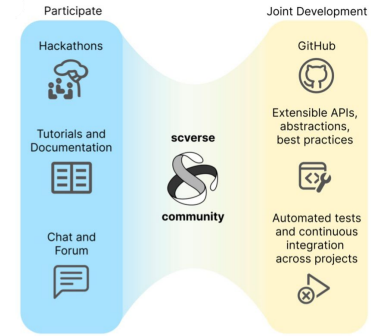


EMBL

scverse







Core people...

Danila Bredikhin | Adam Gayoso | Lukas Heumos | Ilia Kats | Giovanni Palla | Gregor Sturm | Isaac Virshup

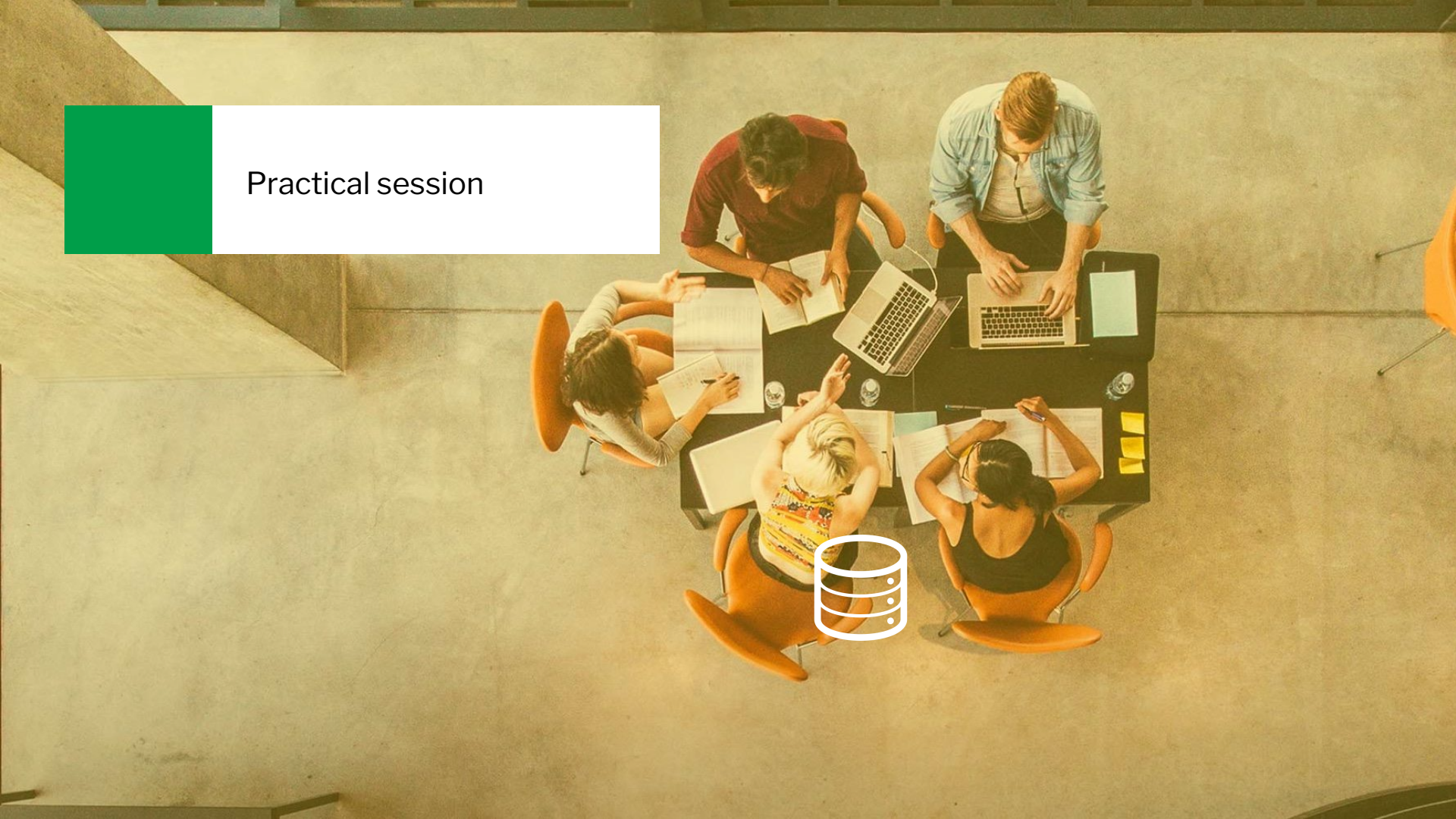
Francesca Finotello | Oliver Stegle | Fabian Theis | Alex Wolf | Nir Yosef

Bonnie Berger | Dana Pe'er | Aviv Regev | Sarah Teichmann

... and hundreds contributors!

EMBL

Practical session



Single-cell multi omics integration

EMBO Practical Course
Integrative analysis of multi-omics data

EMBL
September 21, 2022


Danila Bredikhin @gtcaa


Max Frank


Ilia Kats

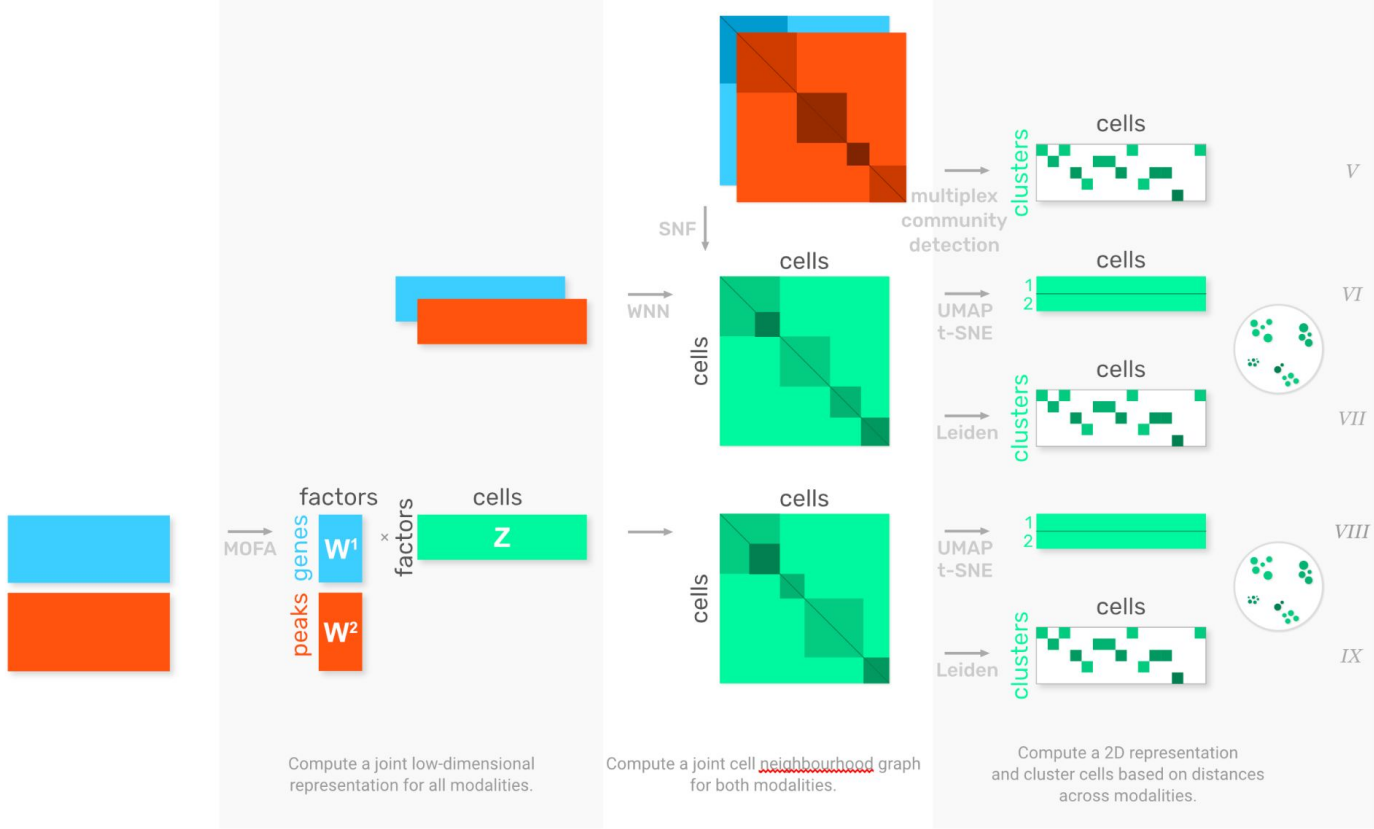
EMBL

Outline

1. Integration strategies (MOFA, WNN, DL)
2. Integration examples
    a. scRNA + scATAC integration
    b. CITE-seq integration
    c. TEA-seq

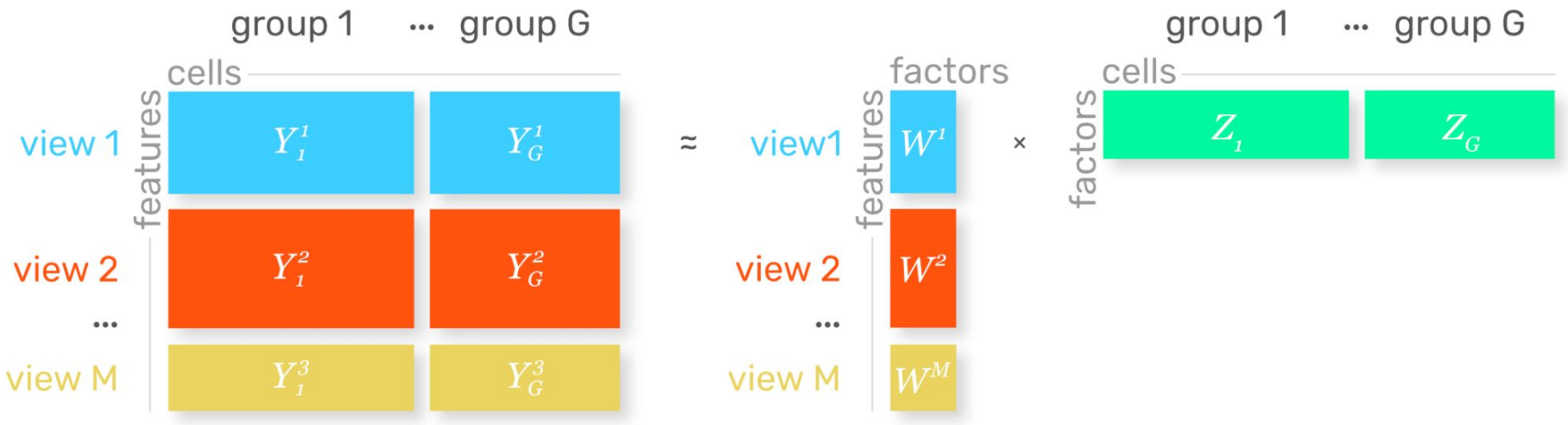
3. Notebook: CITE-seq data integration with MOFA

Different integration strategies with a uniform API



EMBL

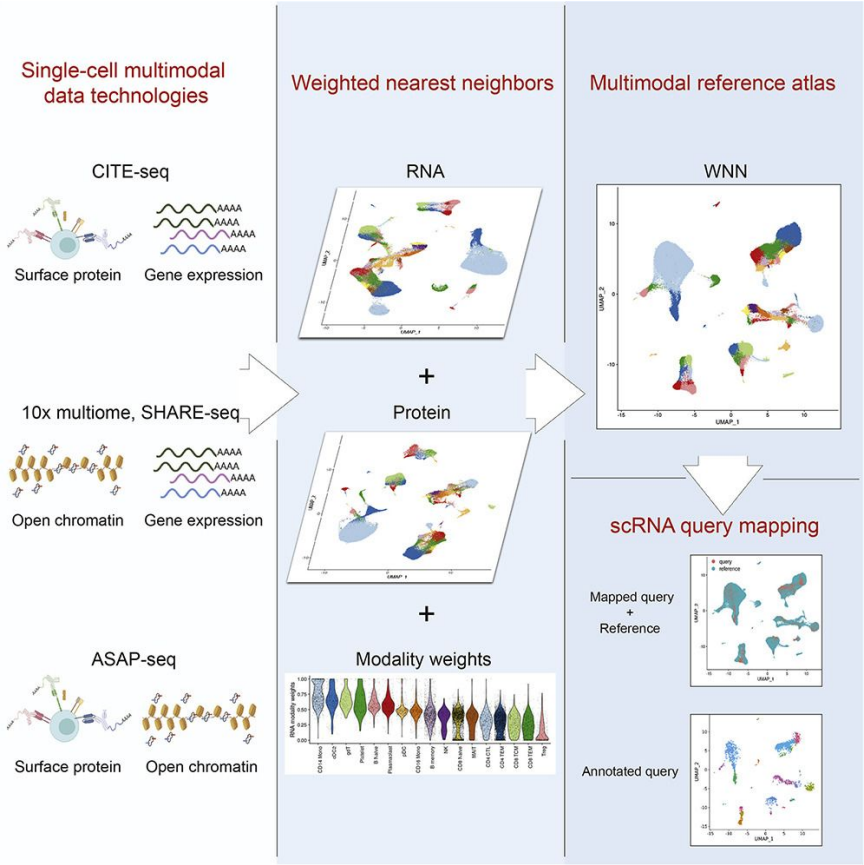
Multi-omics factor analysis (v2)



Argelaguet, Arnol, Bredikhin et al., 2020

EMBL

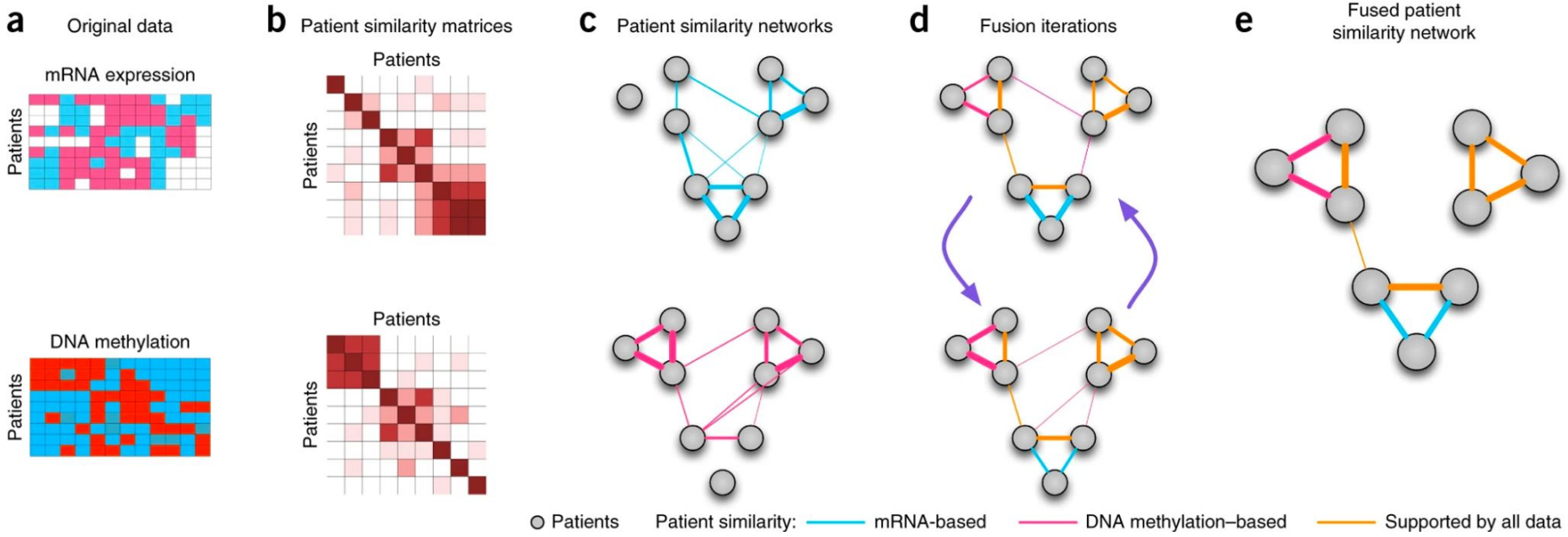
Weighted nearest neighbours (2 modalities)



Hao, Hao et al., 2021

EMBL

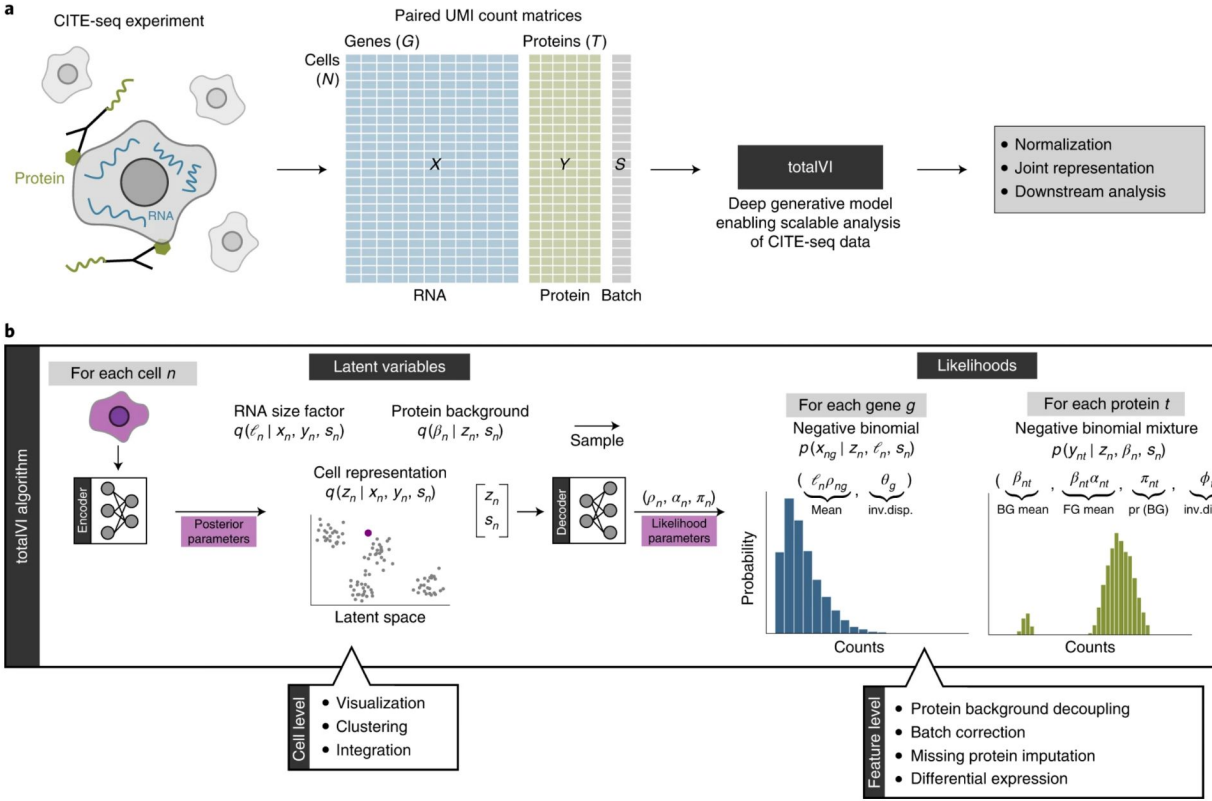
Similarity network fusion



Eang et al., 2014

EMBL

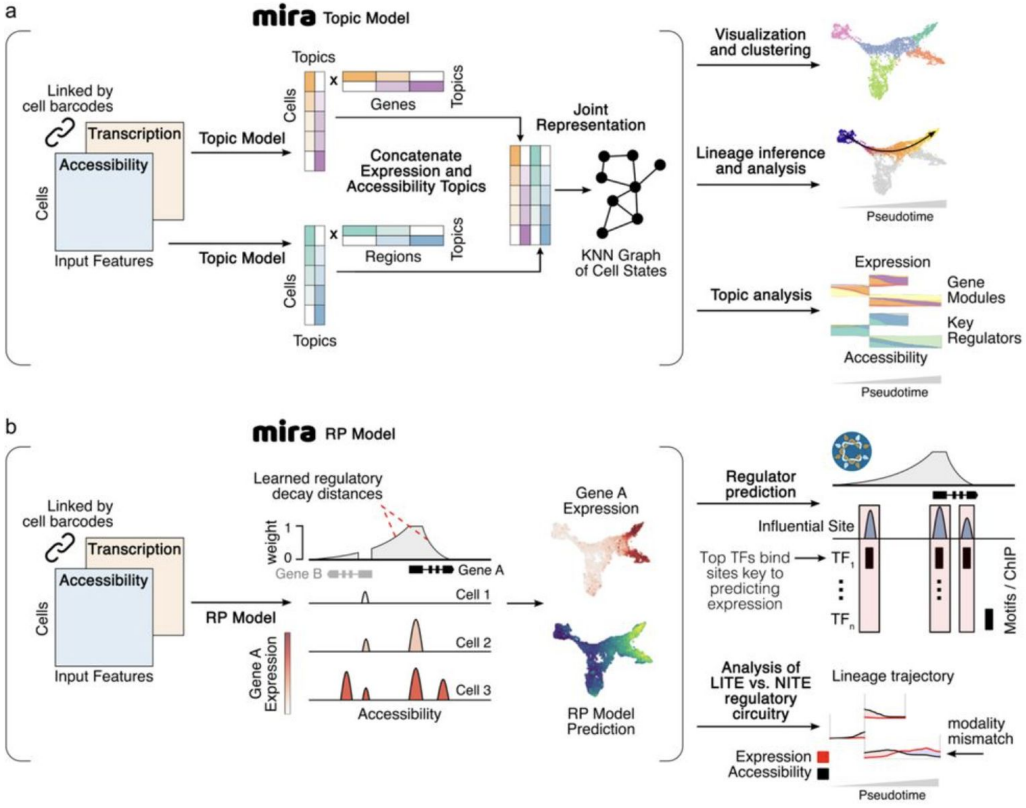
Deep learning-based (example: TotalVI)



Gayoso et al., 2021

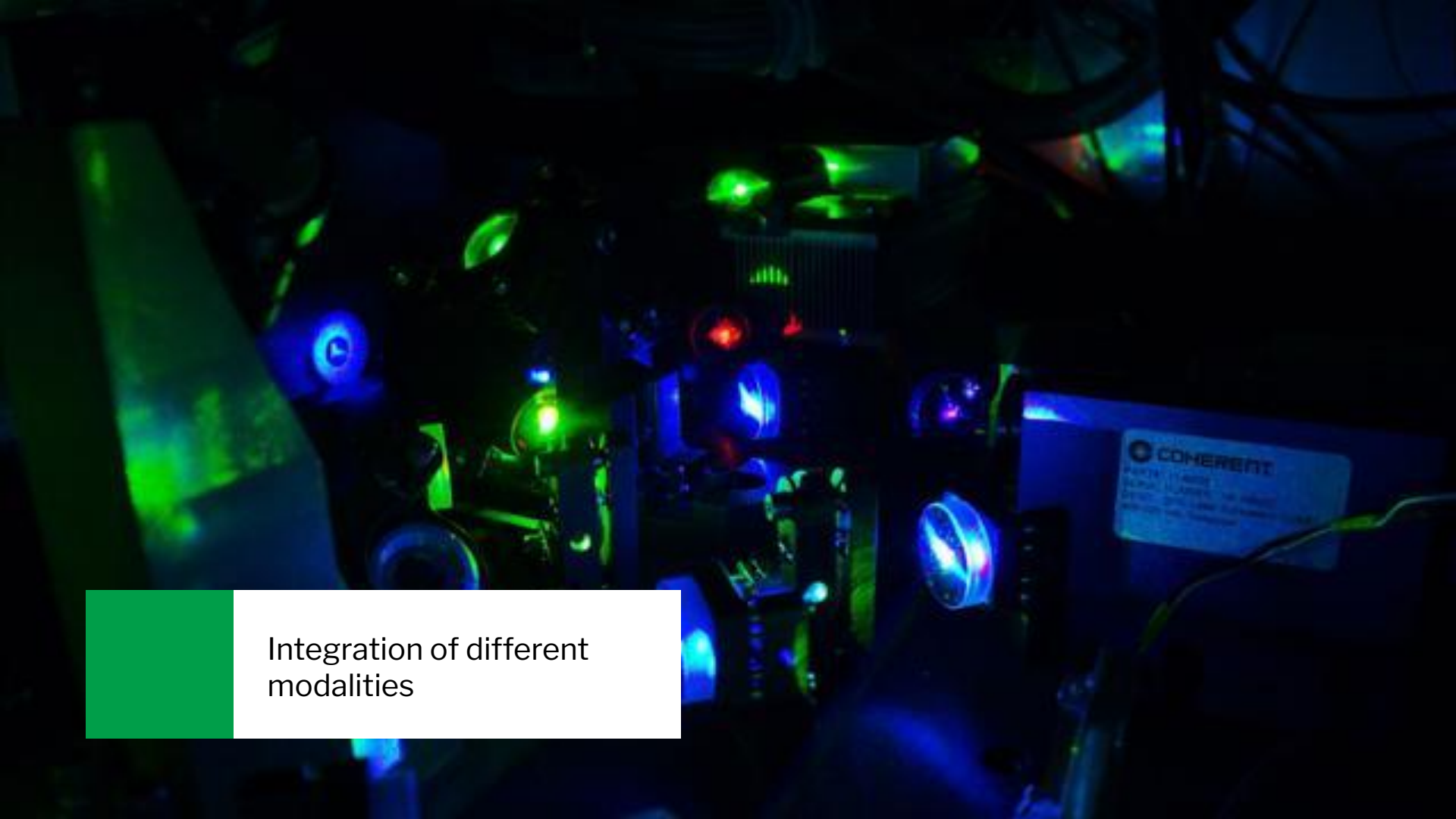
EMBL

Deep learning-based (example: mira)



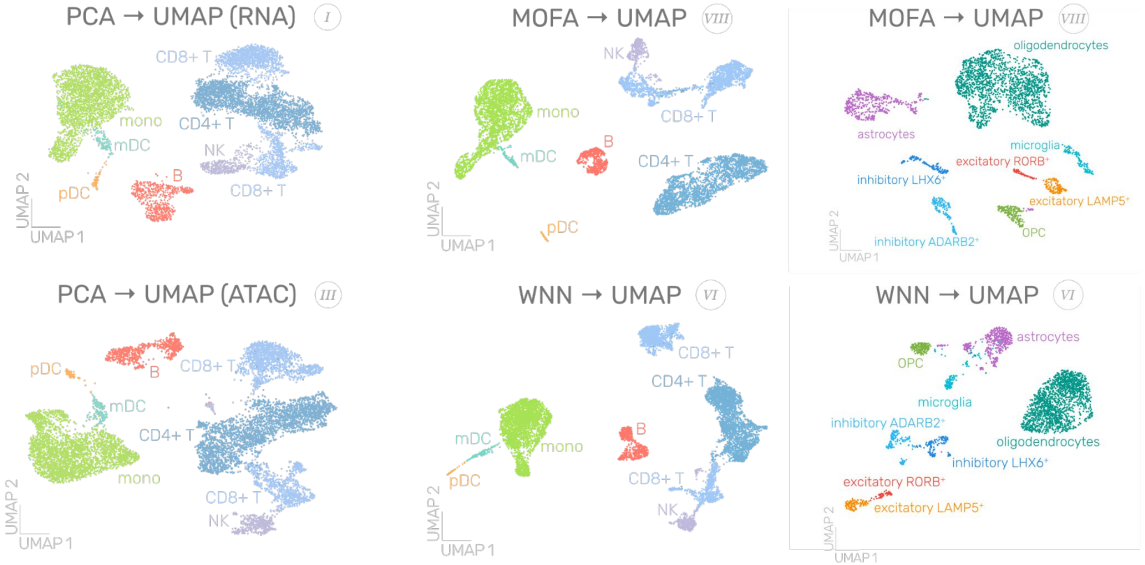
Lynch et al., 2022

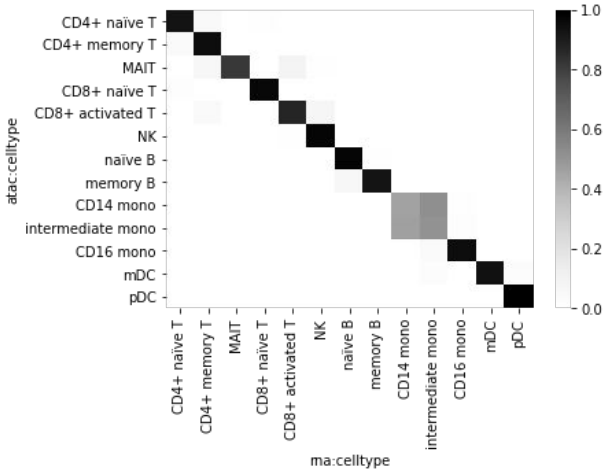
EMBL



Integration of different modalities

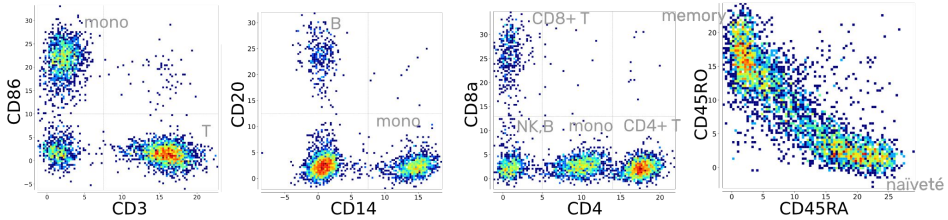
scATAC-seq + scRNA-seq integration example



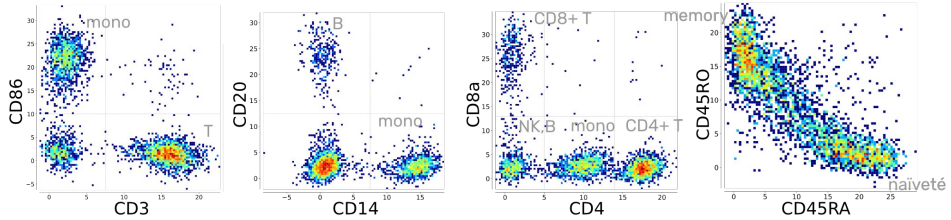


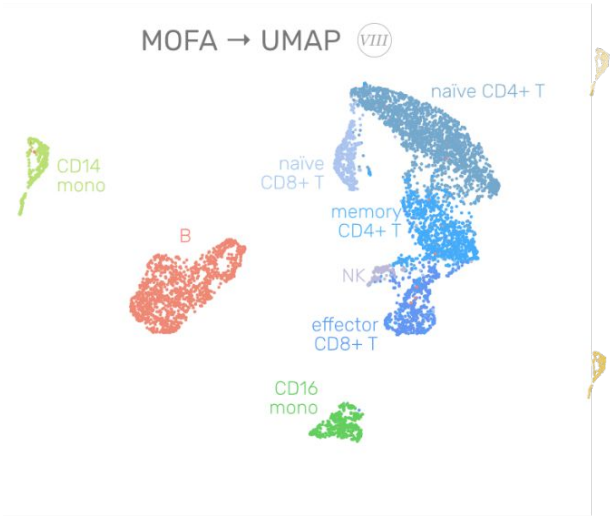
EMBL

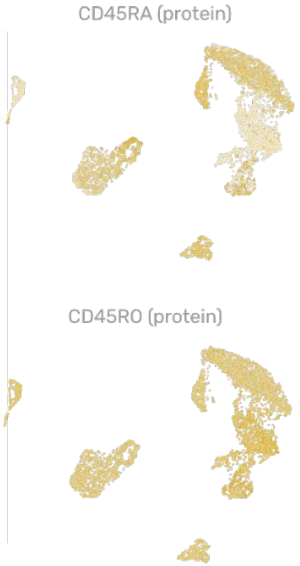
CITE-seq integration example

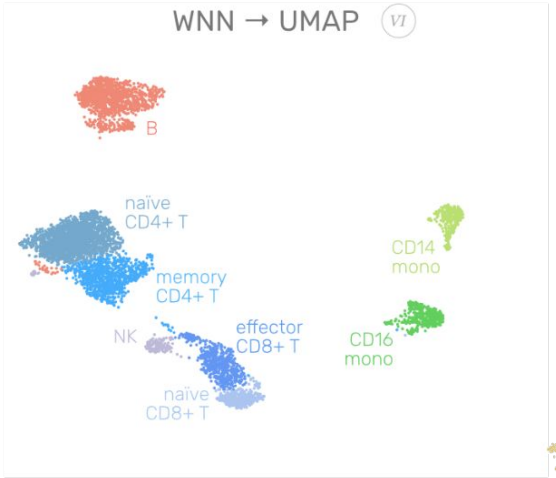


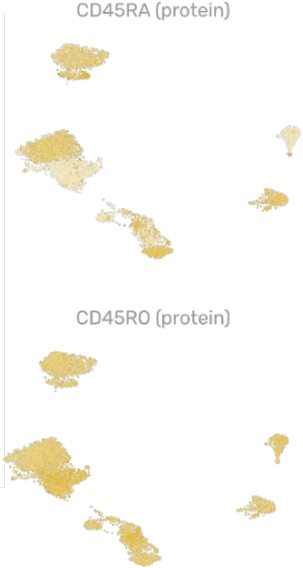
CITE-seq integration example







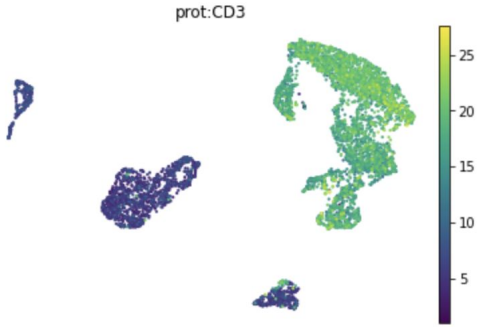




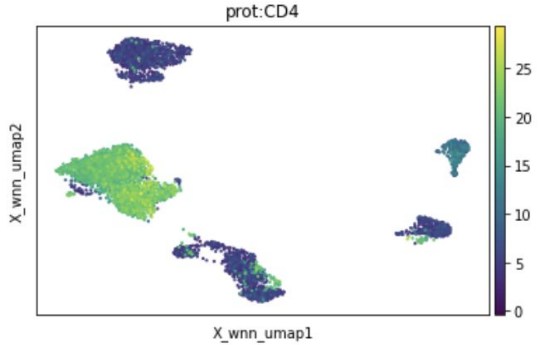
EMBL

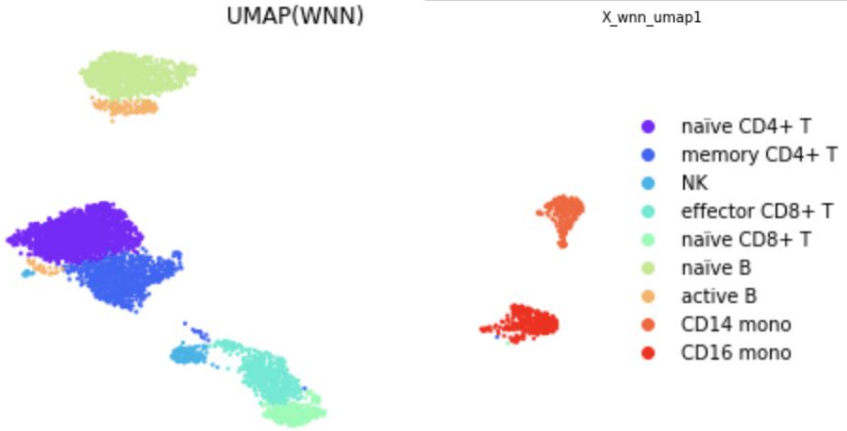
TEA-seq integration example (three modalities)

MOFA → UMAP



WNN → UMAP





EMBL

Practical session