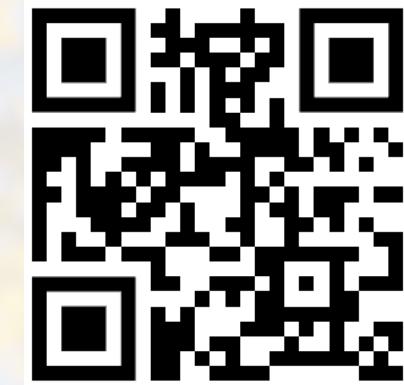


# Machine Learning Development Environment for Single-cell Sequencing Data Analyses



University of Missouri



NIH: R35-GM126985

# Challenges for Method Development

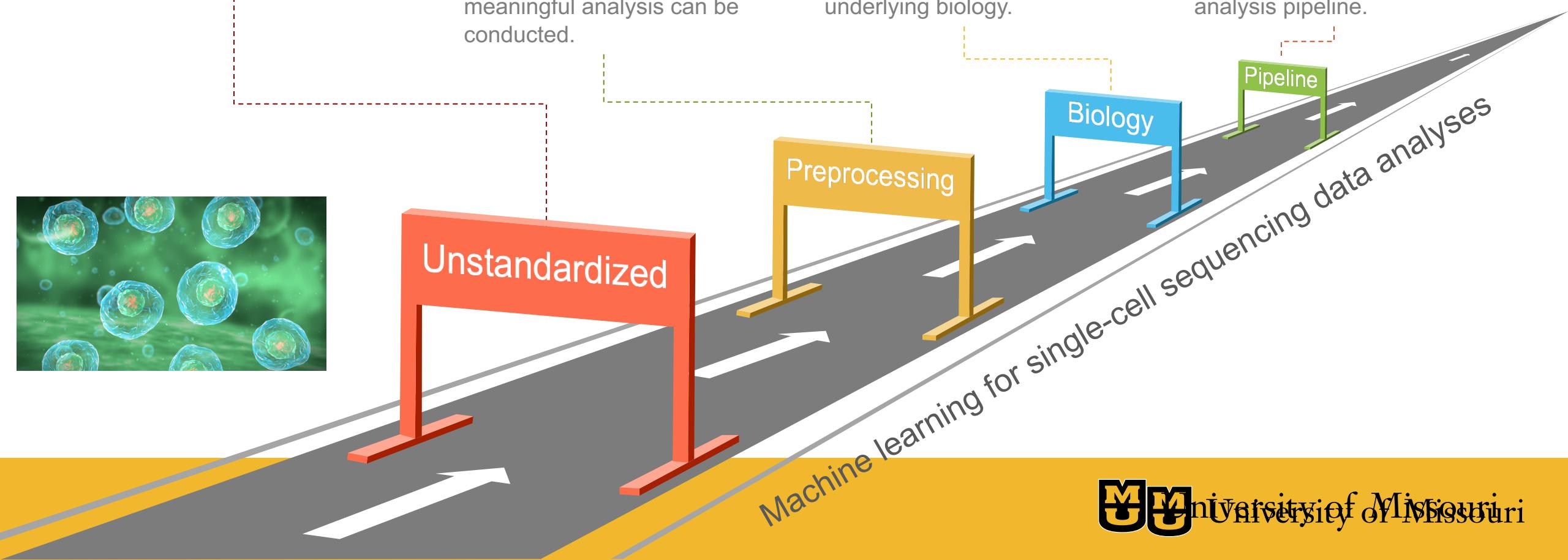
Technological complexity | biological knowledge | data quality | data formatting

Single-cell data are complex and inherently unstandardized.

Most data require preprocessing, such as quality control and normalization before any meaningful analysis can be conducted.

Problem formulations typically require domain knowledge about single-cell technologies and underlying biology.

It is hard to improve a certain method (e.g., just clustering algorithm) without mastering an entire analysis pipeline.



University of Missouri

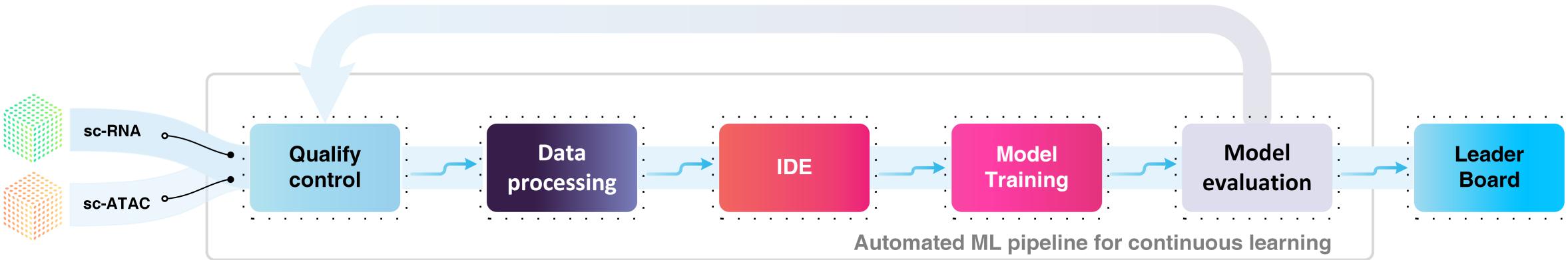
# Ecosystem for Single-Cell Data Analyses

One platform



University of Missouri

# Automated ML pipeline for continuous learning



## Public Datasets

- GEO
- SRA
- CancerSEA
- 10X Genomics
- Simulation
- ...



## Benchmarks

- Raw counts



## Feature store

- Normalized data
- Corrected data
- Summarized data



## Template Libraries



## Environments



## Performance Assessments

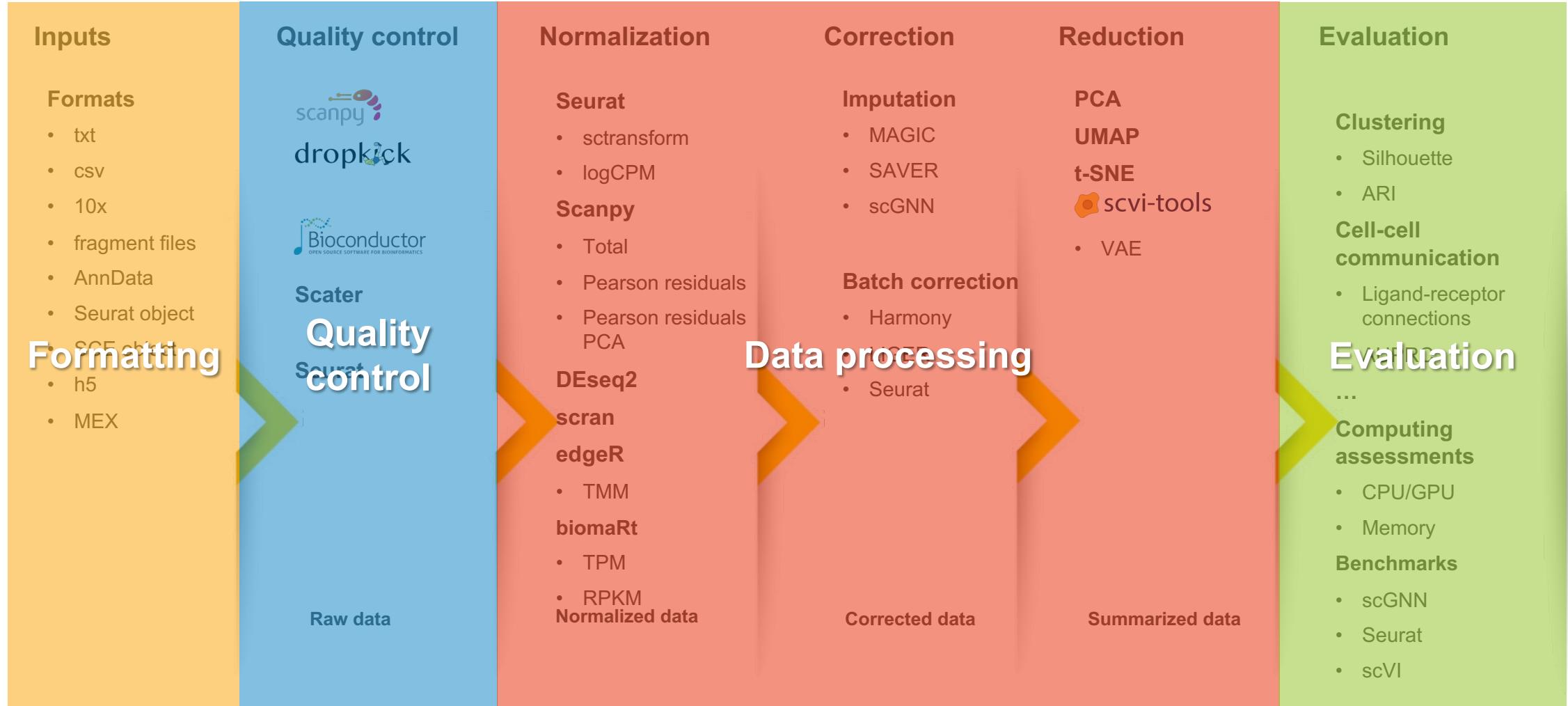


## Ranking



University of Missouri

# Workflow Engine



# Data Engine

## Large-scale AI-ready benchmarks

### Datasets

#### Sources

- GEO
- SRA
- CancerSEA
- 10X Genomics
- Simulation
- ...

#### Species

- Human
- Mouse
- Virus
- ...

#### Tissues

- Brain
- Liver
- COVID
- ...



### Stage of processing

#### Raw data

### Tasks

- Differential expression
- Marker genes
- Genes over condition

- Genes over time

#### Normalized data

- Differential expression
- Marker genes
- Genes over condition

- Genes over time
- Cell-cell communication
- Gene regulatory relations

#### Corrected data

- Visual comparison of data
- Trajectory inference
- Imputation

- Multi-omic data integration

#### Summarized data

- Visualization
- Trajectory inference
- Clustering

- KNN graph inference
- Cell type identification



University of Missouri

# UI Engine

## Discrete and self-contained microservices

### Web

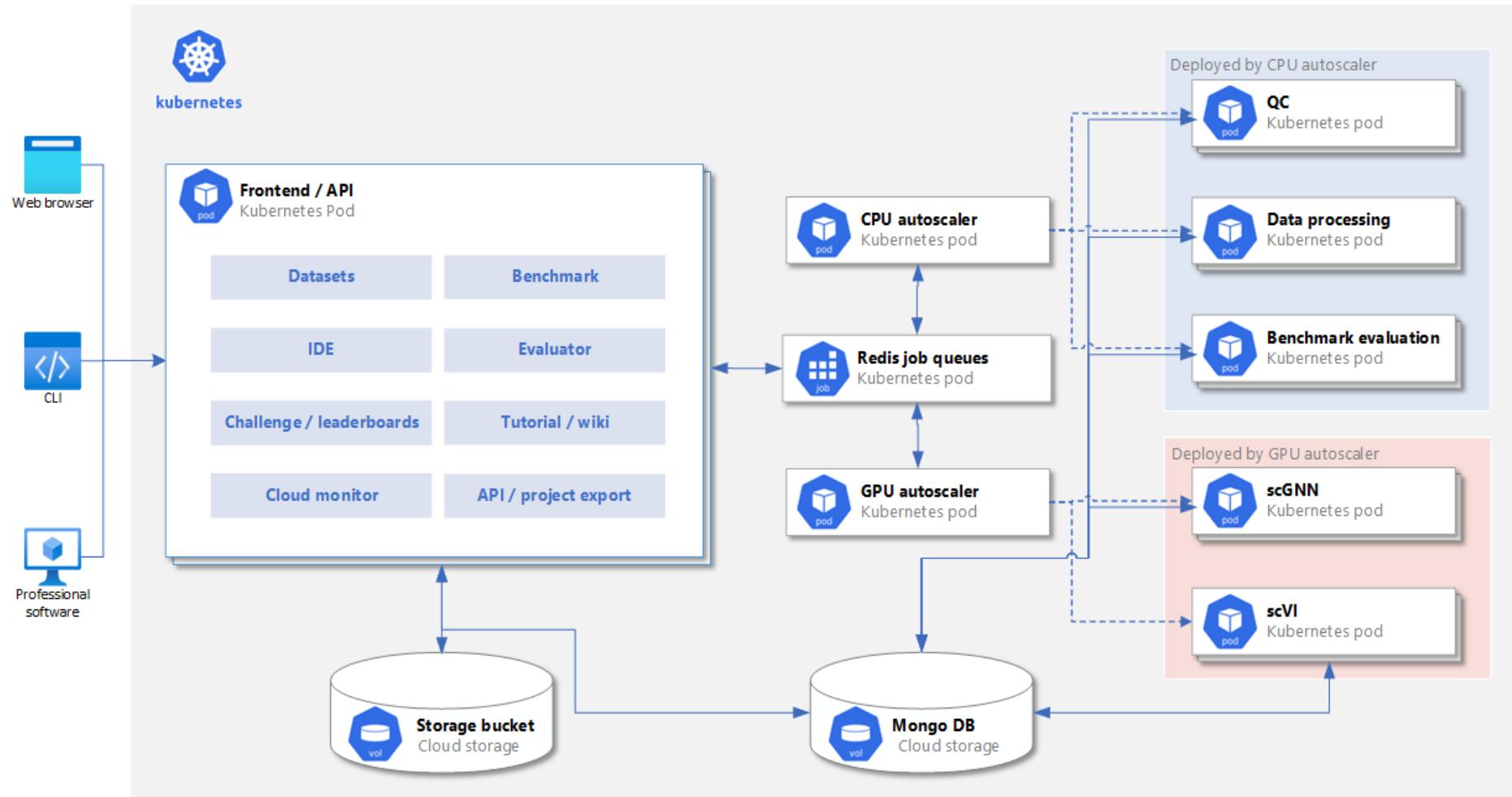
- Analysis
- Method development
- Education
- Documentation
- Visualization

### CLI

- IDE
- Libraries

### Professional software

- Plugin
- API



University of Missouri

# Orchestration Engine

Configuration-based | code as architecture

## Configuration-based development

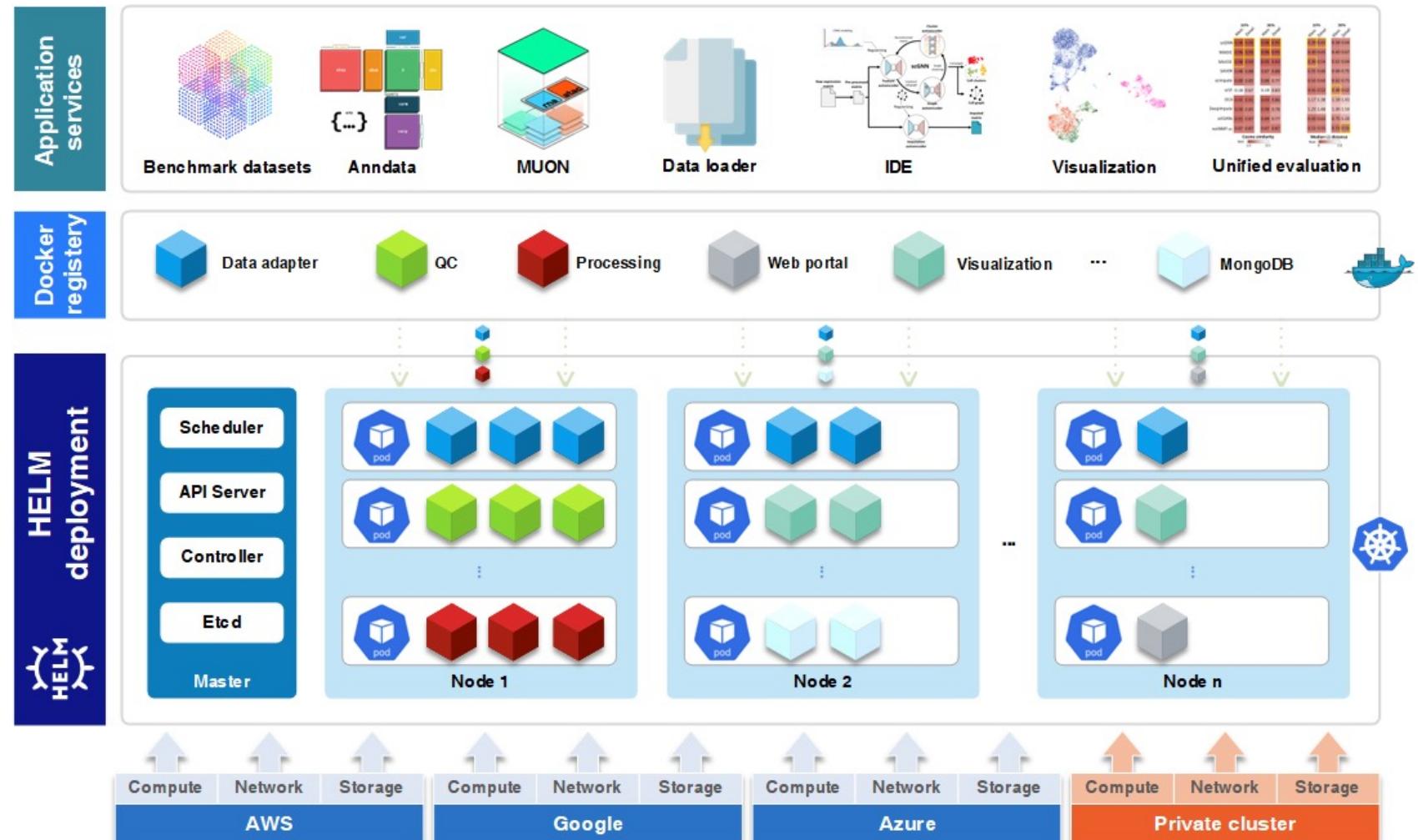
- Snakemake config
- Dockerfile
- Docker-compose
- Helm template

## Code as architecture

- Container
- Kubernetes
- Cluster
- Cloud

## Management

- Version control
- Update
- Monitoring
- Backup



University of Missouri

# Sign Up OSCB

OSCB

Search models, datasets, users...

Get Started Updates Competitions Benchmarks Leaderboards MyData Docs Teams Login/SignUp

My Projects

My Data

Reports

Security

Log In

## Sign Up

Email:

leijiang@missouri.edu

Username:

leijiang

Password:

.....

Confirm Password:

.....

Signup

Already have an account? [Log in](#)



# Public Datasets

Select Datasets Category?

My Datasets  Benchmarks Datasets  User Shared Datasets

**CREATE DATASET**

Search by filters ?

Species

Category

Author

Anatomical Entity

Organ Part

Selected Cell Types

Disease Status (Specimen)

Disease Status (Donor)

Search by text ?

Bladder

Search...

Available filters:

Bladder (1)

Apply

Results 1 - 1 of 1

**EDIT COLUMNS**

Actions

Title

Id

Category

Species

Organ Part

Cell Count Estimate

Bladder

h-Bladder-Wang-2024

Public

human

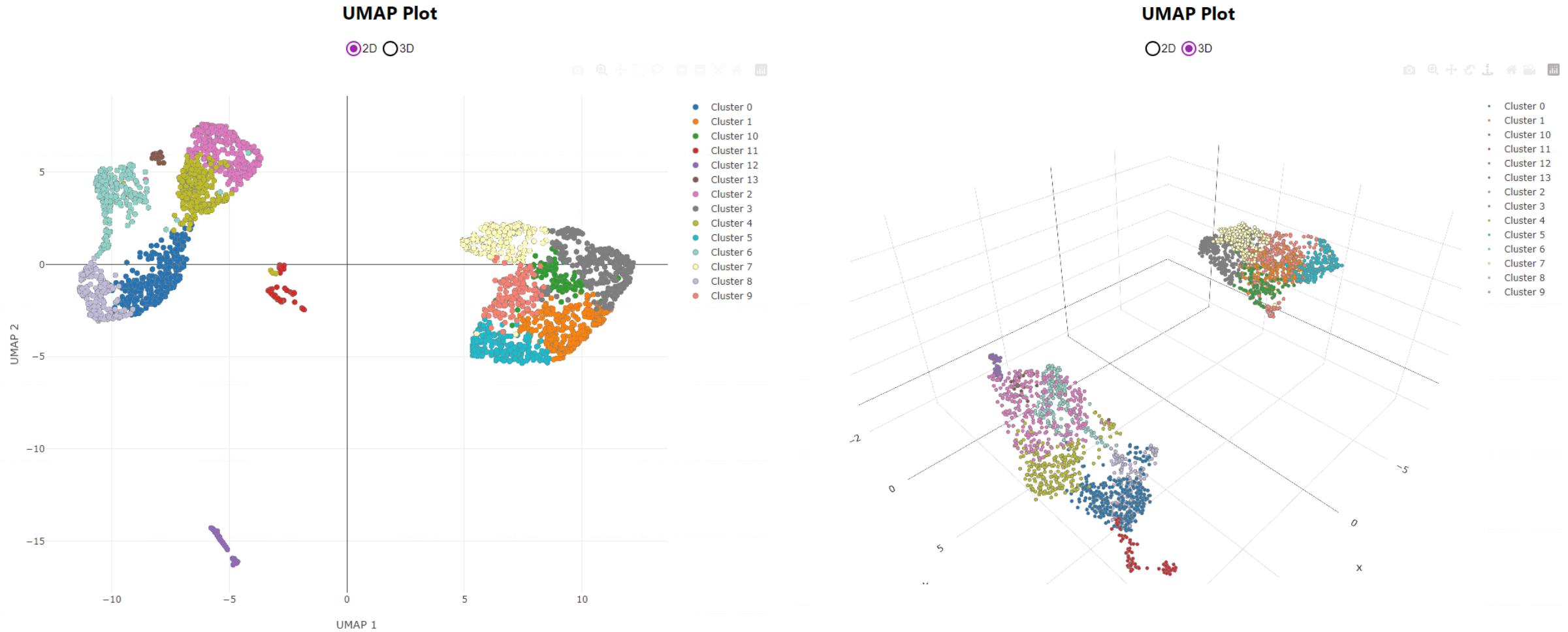
Bladder

0

Close

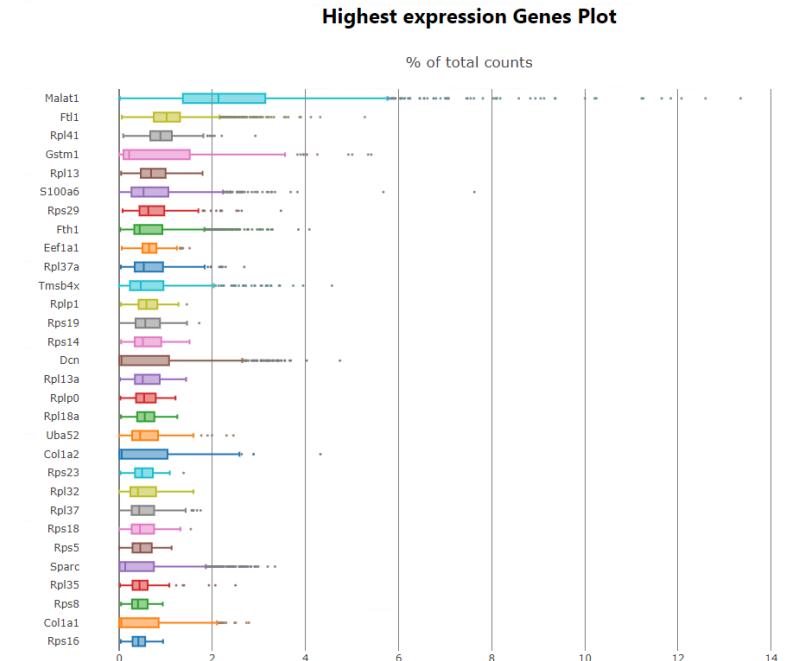
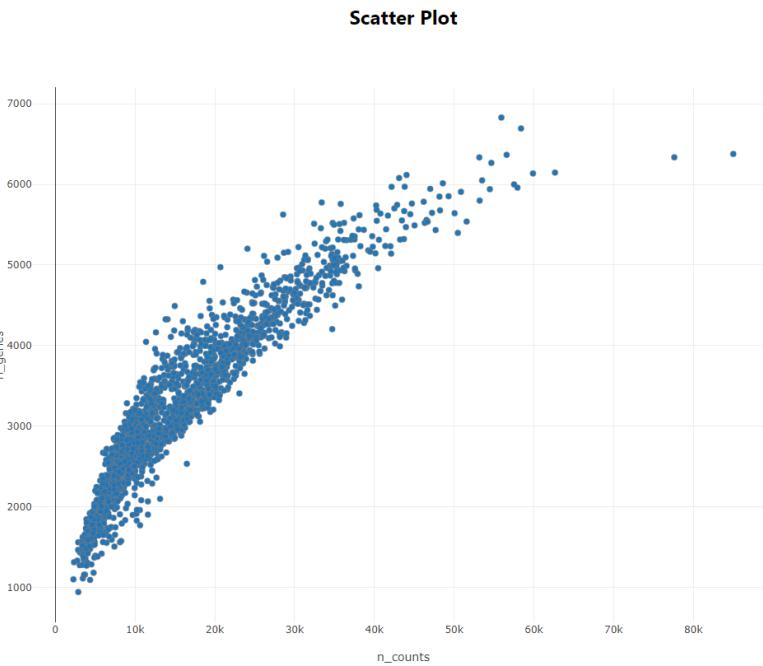
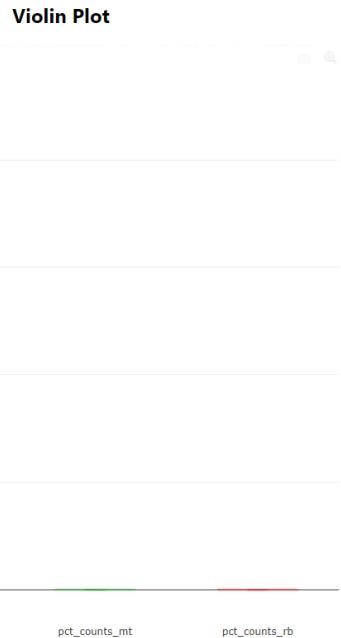
# Benchmarks

## QC Visualization



# Benchmarks

## QC Visualization

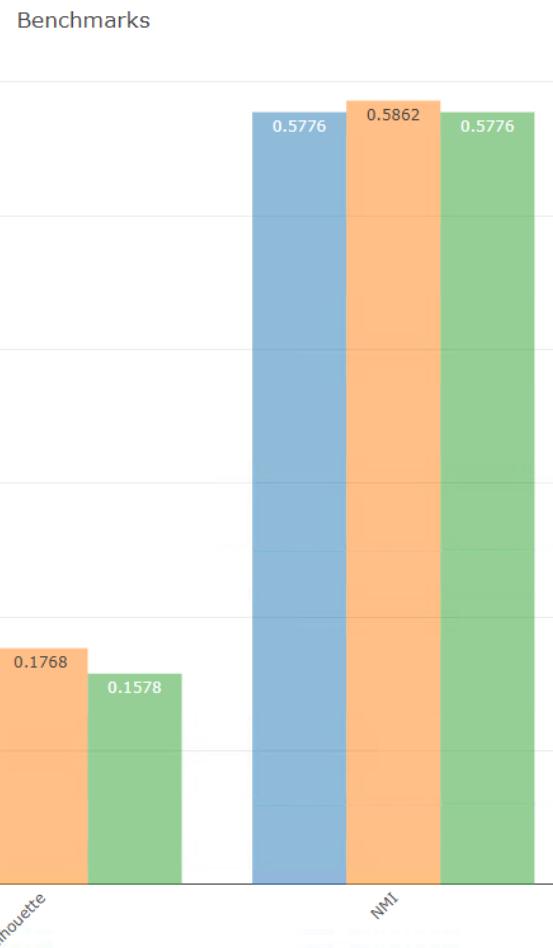


# Benchmarks

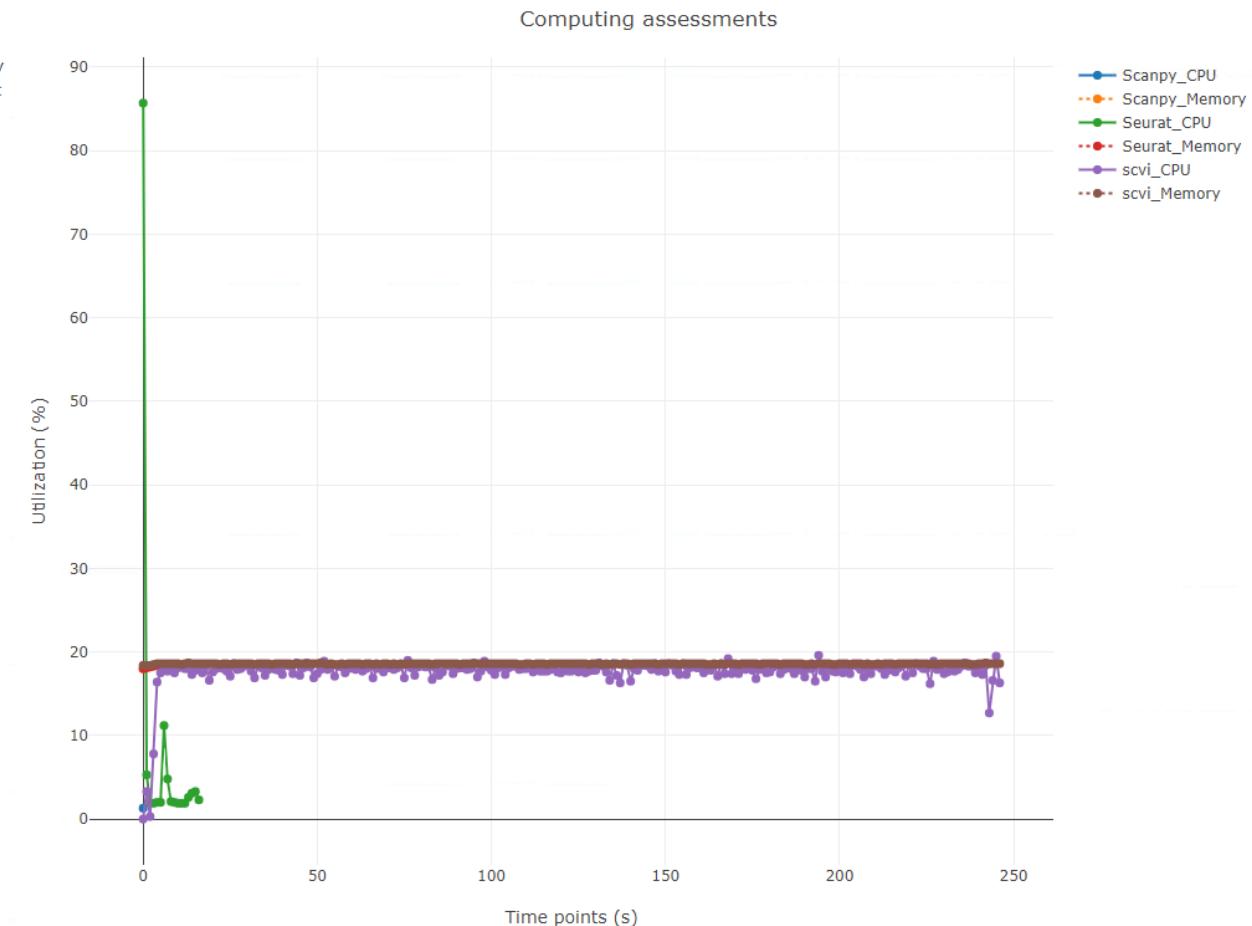
## Assessment

Benchmark Results for h-Bladder-Wang-2024

### Benchmarks



### Computing assessments



# Upload Your Own Datasets

OSCB

Search models, datasets, users...

Get Started Updates Competitions Benchmarks Leaderboards MyData Docs Teams Login/SignUp

Data Upload

Metadata

## Create Dataset

Choose your files \*



Do you want to publish

Drop files here, [browse files](#) or import from:



My Device



Google Drive



OneDrive



Dropbox



Link

Powered by Uppy

[Close](#)

History



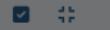
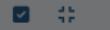
search datasets



Unnamed history



17 GB/5 GB



> My Tasks



# Tools

## Formatting

The screenshot shows the Cell Ranger software interface with the following components:

- Left Sidebar:** A vertical sidebar with a tree view of available tools and datasets.
  - Root Categories:** sctransform, DEseq2, scran, RLE, UQ, UPPERQUARTILE, TPM, FPKM, PEARSON\_RESIDUALS.
  - INTEGRATION:** Liger, Harmony, Seurat.
  - EVALUATION:** CIDR, Seurat.
  - FORMATTING:** Convert.
  - REDUCTION:** (partially visible).
- Tool Parameters Panel:** The main panel for the "Convert" tool.
  - Tool Parameters:** "Choose the input dataset" (required). A dropdown menu shows "Bladder".
  - Parameters for Conversion Process:**
    - Output Format:** A dropdown menu with the following options:
      - AnnData (selected)
      - SingleCellExperiment
      - Seurat
      - CSV
- Right Sidebar:** A vertical sidebar titled "HISTORY".
  - Search bar: "search datasets".
  - History section: "Unnamed history" (17 GB/5 GB).
  - Tasks section: "My Tasks".

# Tools

## QC

**OSCB**

Get Started Updates Competitions Benchmarks Leaderboards MyData Docs Teams Login/SignUp

**QUALITY CONTROL**

- Seurat
- Scanpy
- Bioconductor
- Dropkick

**IMPUTATION**

- scGNN
- SAVER
- MAGIC

**NORMALIZATION**

- TMM
- LogCPM
- LogCP10K
- sctransform
- DEseq2
- scran
- RLE
- UQ
- UPPERQUARTILE
- TDM

**Max Genes:** 200\* 1000 5000 10000 15000 20000

**Min Cells:** 2\* 10 50 100 200

**Target Sum:** 1e4\* 1e5 1e6

**Highly Variable Genes (n\_top\_genes):** 100 500 1000 2000\* 5000 10000

**n\_neighbors:** 2 5 10 15\* 20 50 100

**n\_pcs:** 0\* 5 10 20 40 50 125 200

**Resolution:** 0.1 0.5 1\* 2.5 5

**Expected Doublet Rate:** 0% 8%\* 12.5% 20% 50%

**History**

search datasets

Unnamed history

17 GB/5 GB

My Tasks

# Tools

## Normalization

**OSCB**  Search models, datasets, users...

Get Started Updates Competitions Benchmarks Leaderboards MyData Docs Teams Login/SignUp

**QUALITY CONTROL**

- Seurat
- Scanpy
- Bioconductor
- Dropkick

**IMPUTATION**

- scGNN
- SAVER
- MAGIC

**NORMALIZATION**

- TMM
- LogCPM
- LogCP10K
- sctransform
- DEseq2
- scran
- RLE
- UQ
- UPPERQUARTILE
- TMM

Toggle this switch to decide whether to perform UMAP (Uniform Manifold Approximation and Projection) analysis after processing your data.

**Do you want to see the errors of the task execution ?**

Errors during task execution are shown by default, aiding in troubleshooting and refining your analysis. Disable if preferred.

**Normalization Parameters**

**Method\***

LogCP10K

**Default Assay (If applicable)**

RNA

**Layer**

Specify the layer

Enter Cluster Label

**n\_neighbors:**

2 5 10 15\* 20 50 100

**n\_pcs:**

0\* 5 10 20 40 50 125 200

**Resolution:**

0.1 0.5 1\* 2.5 5

**History**

search datasets

Unnamed history

17 GB/5 GB

My Tasks

# Tools

## Imputation

The screenshot shows the 'Tools' section for 'Imputation'. On the left, a sidebar lists various tools: Seurat, Scanpy, Bioconductor, Dropkick, IMPUTATION (selected), scGNN, SAVER, MAGIC, NORMALIZATION (selected), TMM, LogCPM, LogCP10K, sctransform, DEseq2, scran, RLE, UQ, UPPERQUARTILE, TPM, and FPKM.

The main panel has the following sections:

- SELECT THE OUTPUT FORMAT FOR STOREING YOUR RESULTS**: AnnData
- Species Type**: human
- ID Type**: Choose the ID Type to specify the gene identifier format for your analysis. A dropdown menu shows 'Select...'.
- Cluster Label**: Enter the Cluster Label to identify or categorize your data points within the AnnData object. An input field shows 'Enter Cluster Label'.
- Do you want to analyse umap after processing ?**: A toggle switch is turned on (green).
- Do you want to see the errors of the task execution ?**: A toggle switch is turned on (green).
- Imputation Parameters**
  - Method\***: Magic
  - Default Assay (If applicable)**: RNA
  - Layer**: (empty input field)

On the right, there is a sidebar with a search bar ('search datasets'), an 'Unnamed history' section (17 GB/5 GB), and a 'My Tasks' section.

# Tools

## Integration

**OSCB**

Get Started Updates Competitions Benchmarks Leaderboards MyData Docs Teams Login/SignUp

**NORMALIZATION**

- TMM
- LogCPM
- LogCP10K
- sctransform
- DEseq2
- scran
- RLE
- UQ
- UPPERQUARTILE
- TPM
- FPKM
- PEARSON\_RESIDUALS

**INTEGRATION**

- Liger
- Harmony
- Seurat

**EVALUATION**

**FORMATTING**

**Tool Parameters**

**Choose the input dataset** \* required  
Bladder

**Parameters for Integration Process**

**Output Format**  
Select the output format for storing your results  
AnnData

**Method\***  
Harmony

**Integration Parameters**

**Default Assay (If applicable)**  
RNA

**n\_pcs:**  
 0\* 5 10 20 40 50 125 200

**Layer**  
Specify the layer  
Enter Cluster Label

**History**

search datasets

Unnamed history

17 GB/5 GB

My Tasks

# Tools

## Reduction and Visualization

OSCB

Get Started Updates Competitions Benchmarks Leaderboards MyData Docs Teams Login/SignUp

sctransform  Toggle this switch to decide whether to perform UMAP (Uniform Manifold Approximation and Projection) analysis after processing your data.

DEseq2

scran

RLE

UQ

UPPERQUARTILE

TPM

FPKM

PEARSON\_RESIDUALS

**INTEGRATION**

Liger

Harmony

Seurat

**EVALUATION**

CIDR

Seurat

**FORMATTING**

Convert

**REDUCTION**

Do you want to see the errors of the task execution ?  
Errors during task execution are shown by default, aiding in troubleshooting and refining your analysis. Disable if preferred.

**Dimension Reduction Parameters**

**Default Assay (If applicable)**: RNA

**Layer**: Specify the layer  
Enter Cluster Label

**n\_neighbors:** 15\*

**n\_pcs:** 40

**Resolution:** 1\*

**Use Default**

History

search datasets

Unnamed history

17 GB/5 GB

My Tasks

# Task Monitor

## Task Details for Task ID: f6298e10-d22b-4388-80ad-b4466f6240c1

### Dataset Information

**Dataset Title:** Bladder

**Dataset URL:** [/usr/src/app/storage/leijiang/projects/Bladder\\_1717462919380/droplet\\_Bladder\\_s](http://usr/src/app/storage/leijiang/projects/Bladder_1717462919380/droplet_Bladder_s)

### Execution Details

**Tool:** Quality Control      **Method:** Scanpy

**Status:** SUCCESS

### Live Logs

```
INFO | Total number of cells: 2500
INFO | Number of cells after filtering of low quality cells: 2441
INFO | Normalizing dataset usig log10000.
INFO | Finding highly variable genes.
INFO | Scanpy Quality Control is completed.
INFO | Computing PCA, neighborhood graph, tSNE, UMAP, and 3D UMAP
INFO | Clustering the neighborhood graph.
INFO | Retrieving metadata and embeddings from AnnData object.
INFO | Saving AnnData object.
INFO | AnnData object with n_obs x n_vars = 2441 x 2000 obs: 'orig.ident', 'n_counts', 'n_genes', 'channel', 'tissue', 'subtissue', 'mouse.sex', 'mouse.id', 'percent.ercc', 'percent.ribo', 'free_annotation', 'cell_ontology_class', 'res.0.4', 'previous_free_annotation', 'previous_cell_ontology_class', 'cluster.ids', 'cell_ontology_id', 'n_genes_by_counts', 'log1p_n_genes_by_counts', 'total_counts', 'log1p_total_counts', 'pct_counts_in_top_20_genes', 'total_counts_mt', 'log1p_total_counts_mt', 'pct_counts_mt', 'total_counts_ribo', 'log1p_total_counts_ribo', 'pct_counts_ribo', 'total_counts_hb', 'log1p_total_counts_hb', 'pct_counts_hb', 'outlier', 'mt_outlier', 'leiden', 'louvain' var: 'name', 'n_cells', 'mt', 'ribo', 'hb', 'n_cells_by_counts', 'mean_counts', 'log1p_mean_counts', 'pct_dropout_by_counts', 'total_counts', 'log1p_total_counts', 'highly_variable', 'means', 'dispersions', 'dispersions_norm', 'mean', 'std' uns: 'log1p', 'hvg', 'pca', 'neighbors', 'tsne', 'umap', 'leiden', 'louvain' obsm: 'X_pca', 'X_tsne', 'X_umap', 'X_umap_3D' varm: 'PCs' layers: 'log10k' obsp: 'distances', 'connectivities'
```

