Machine Learning Development Environment for Single-cell Sequencing Data Analyses

Lei Jiang

PhD student , Computer Science

University of Missouri

NIH: R35-GM126985

Challenges for Method Development

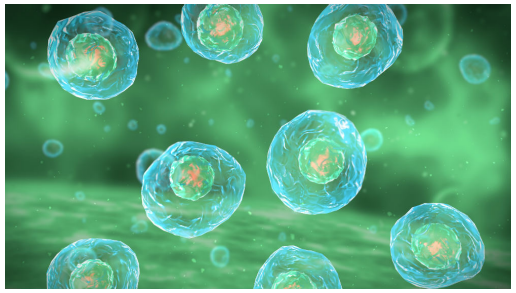
Technological complexity | biological knowledge | data quality | data formatting

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.



Ecosystem for Single-Cell Data Analyses

One platform

Workflow engine

Streamlined
Standardized
Automated
Reproducible

Data engine

AI-ready
Large-scale
Benchmarks
Findable

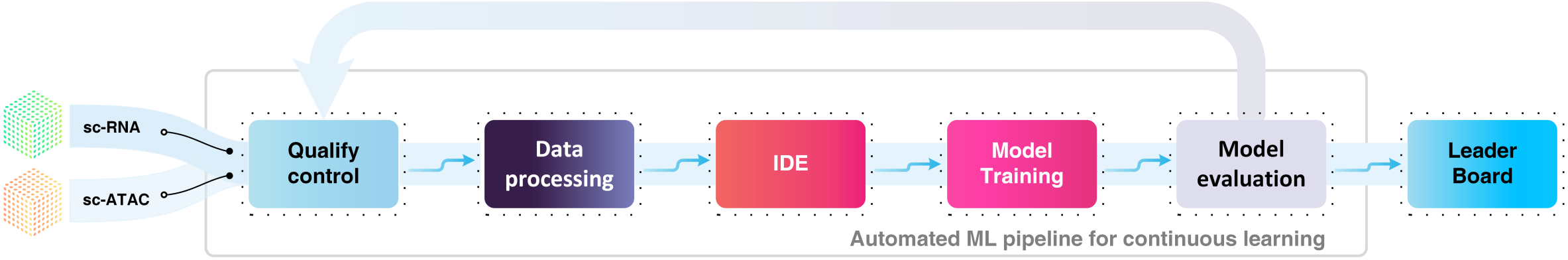
UI engine

Accessible
Customizable
User-friendly
Visualized

Orchestration engine

Scalable
Available
Sustainable
Flexible

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

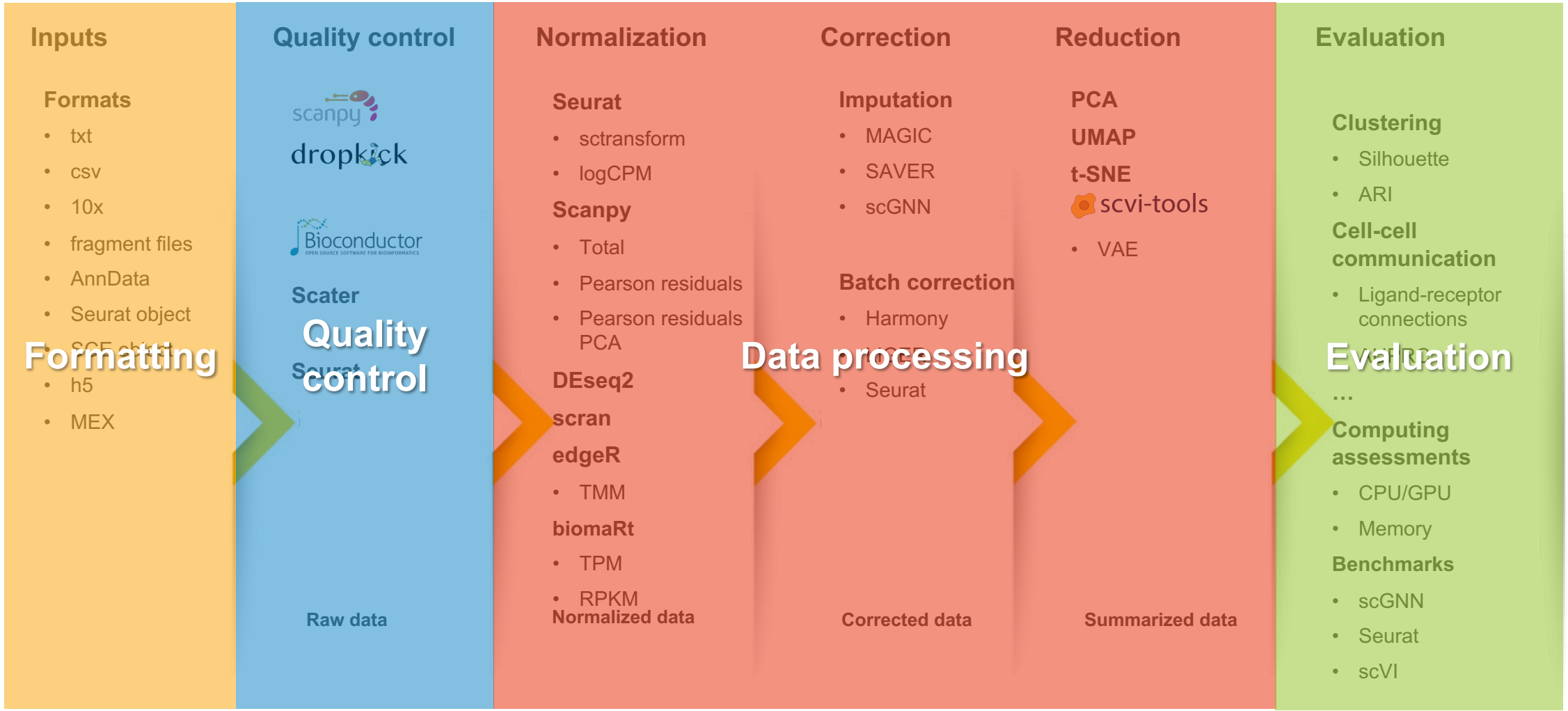
Environments

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

Normalized data

Corrected data

Summarized data

Tasks

- Differential expression
- Marker genes
- Genes over condition
- Genes over time

- Differential expression
- Marker genes
- Genes over condition
- Genes over time
- Cell-cell communication
- Gene regulatory relations

- Visual comparison of data
- Trajectory inference
- Imputation
- Multi-omic data integration

- Visualization
- Trajectory inference
- Clustering
- KNN graph inference
- Cell type identification

University of Missouri

UI Engine

Discrete and self-contained microservices

Web

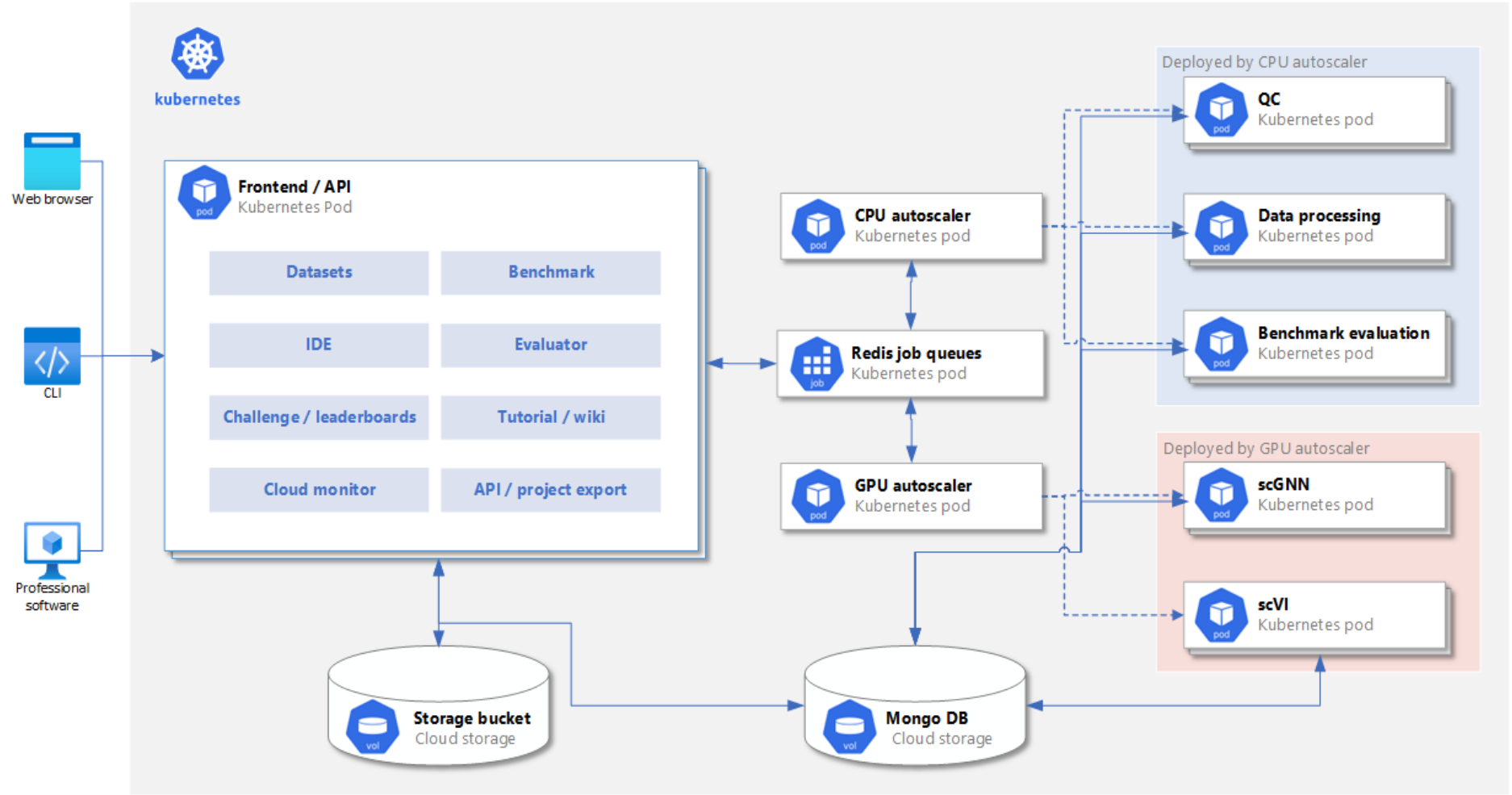
- Analysis
- Method development
- Education
- Documentation
- Visualization

CLI

- IDE
- Libraries

Professional software

- Plugin
- API



University of Missouri

Orchestration Engine

Configuration-based | code as architecture

Configuration-based development

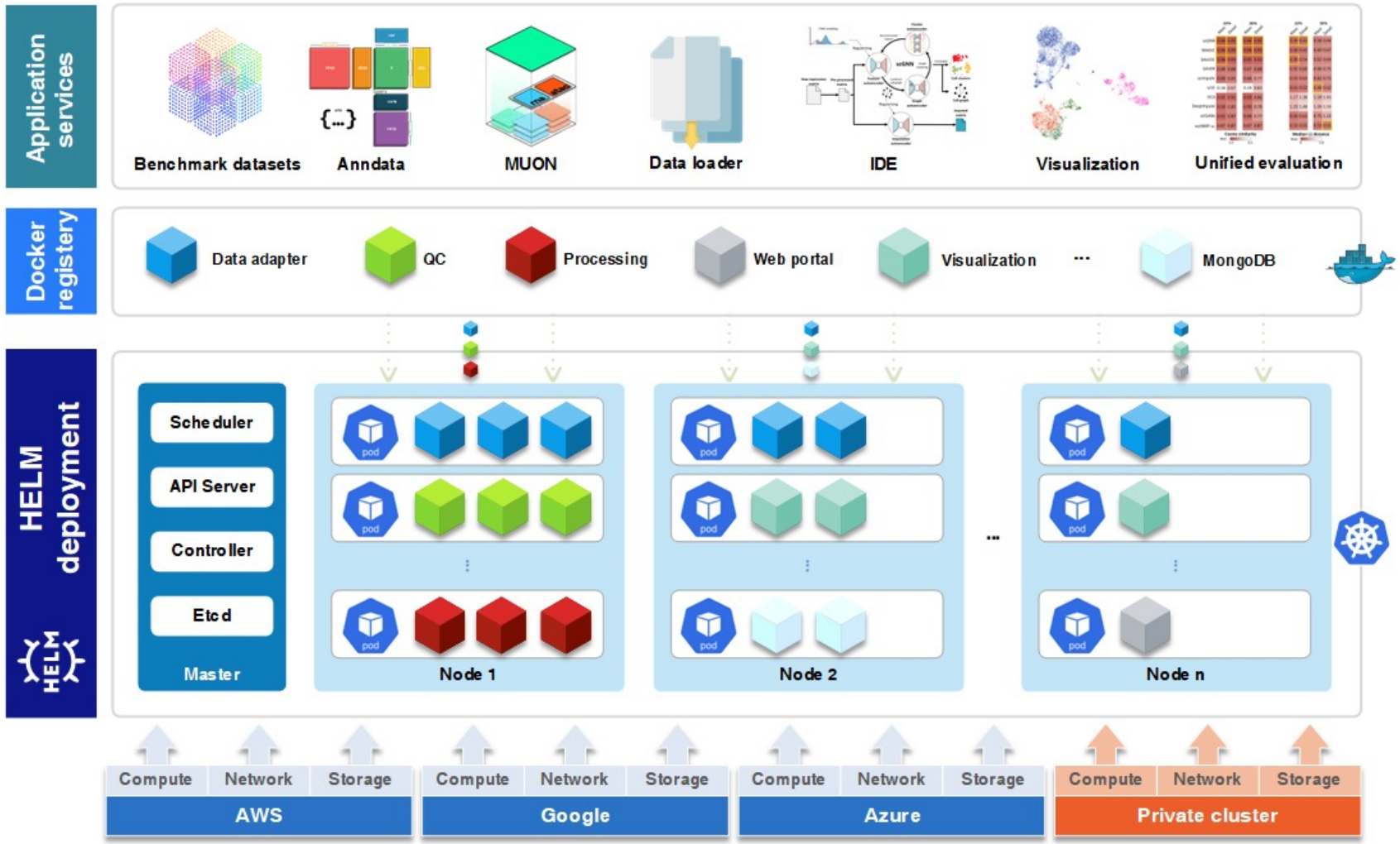
- Snakemake config
- Dockerfile
- Docker-compose
- Helm template

Code as architecture

- Container
- Kubernetes
- Cluster
- Cloud

Management

- Version control
- Update
- Monitoring
- Backup

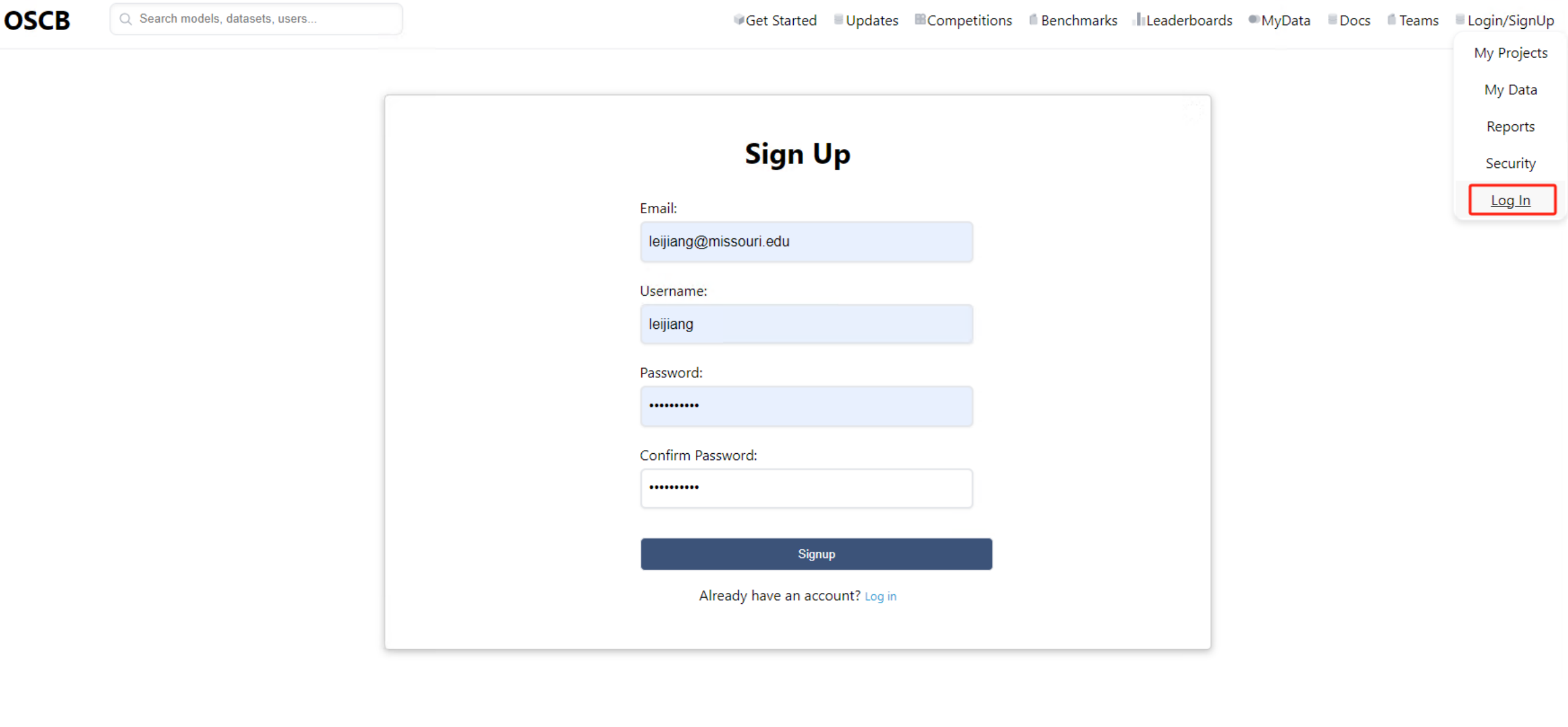


University of Missouri

URL

http://oscb.missouri.edu:3000
http://clnode159.clemson.cloudlab.us:3000
http://clnode174.clemson.cloudlab.us:3000
http://clnode143.clemson.cloudlab.us:3000
http://clnode167.clemson.cloudlab.us:3000

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

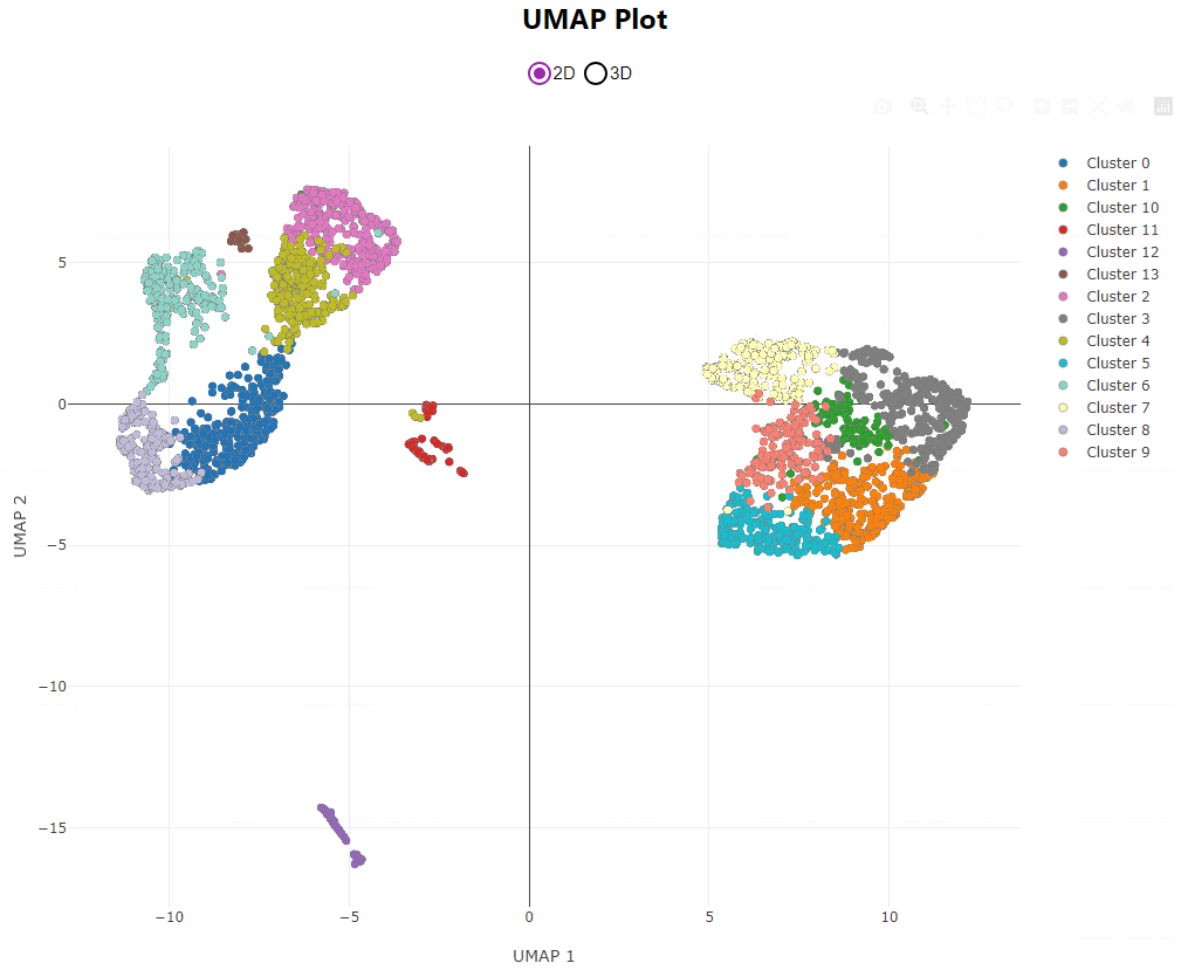
EDIT COLUMNS

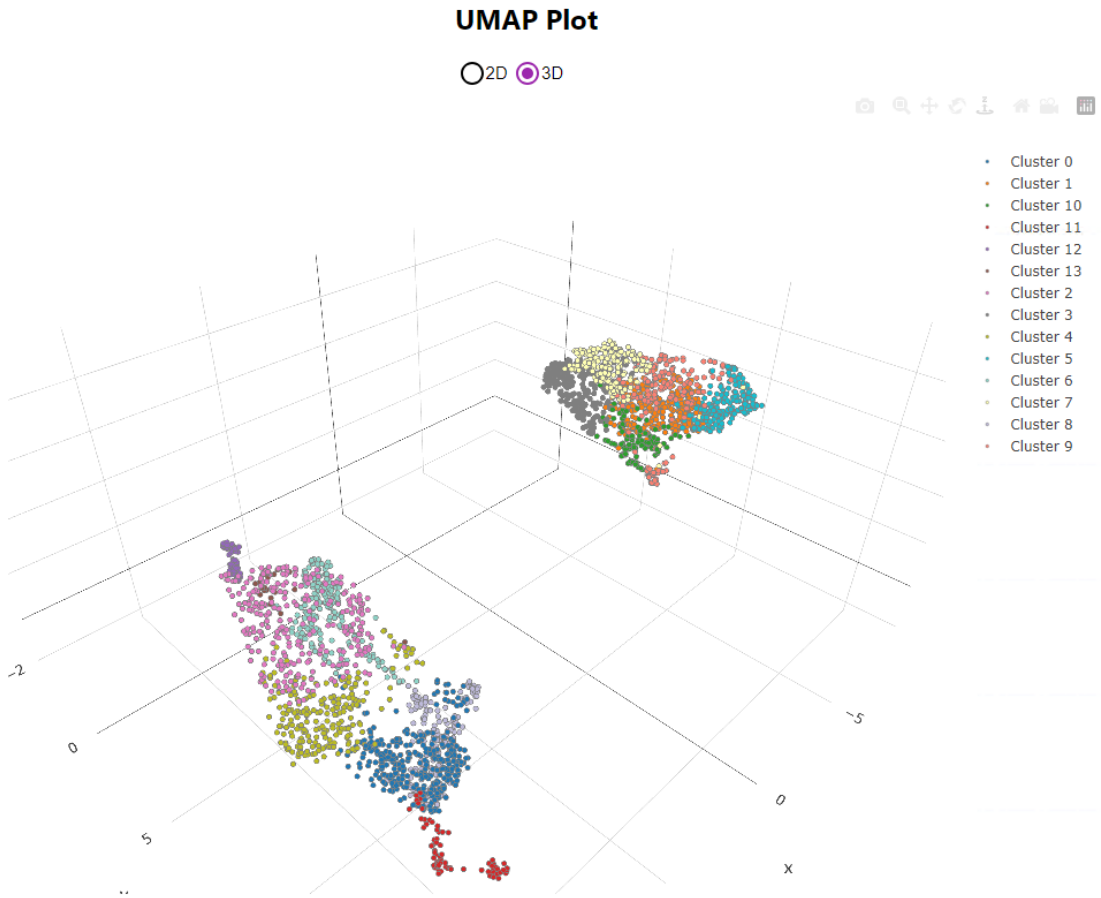
Actions	Title	Id	Category	Species	Organ Part	Cell Count Estimate
	Bladder	h-Bladder-Wang-2024	Public	human	Bladder	0

Close

Benchmarks

QC Visualization

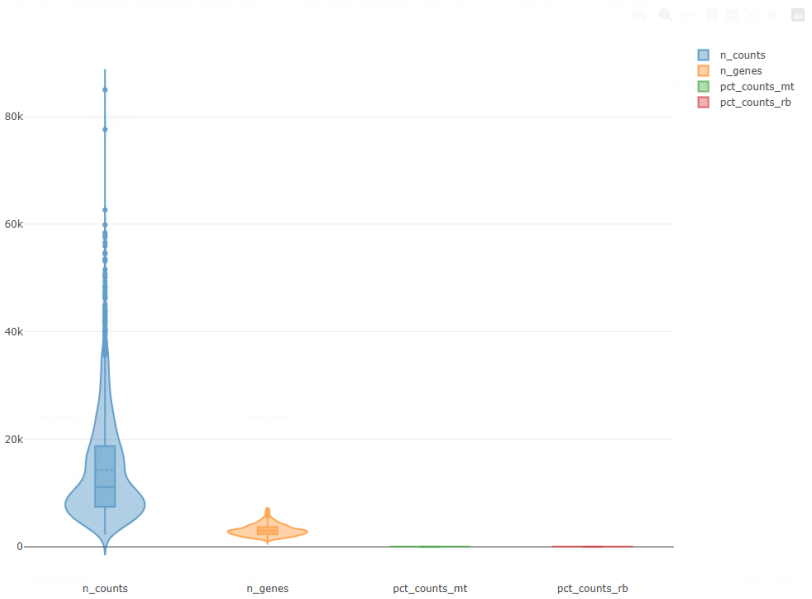




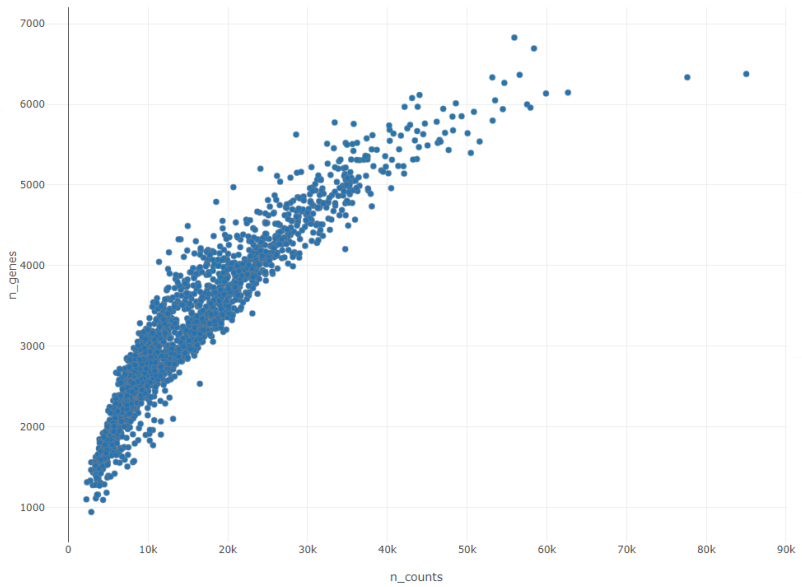
Benchmarks

QC Visualization

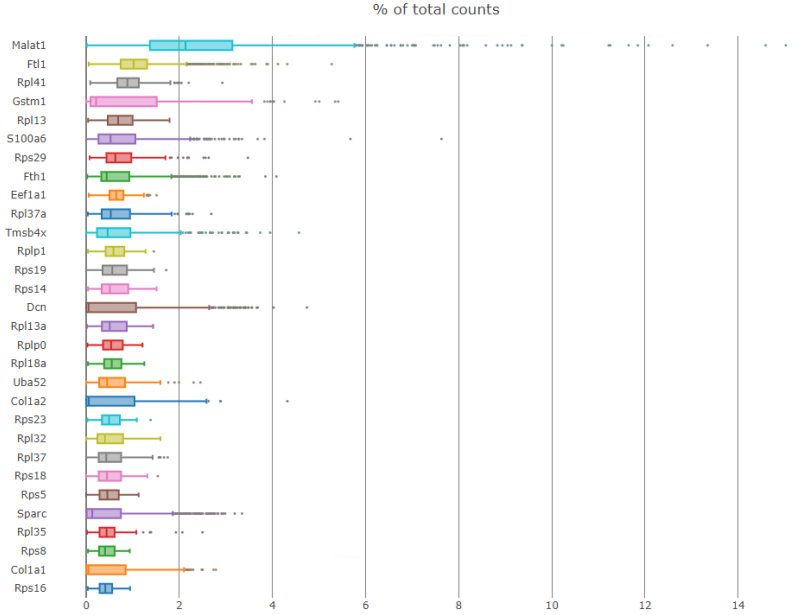
Violin Plot



Scatter Plot



Highest expression Genes Plot

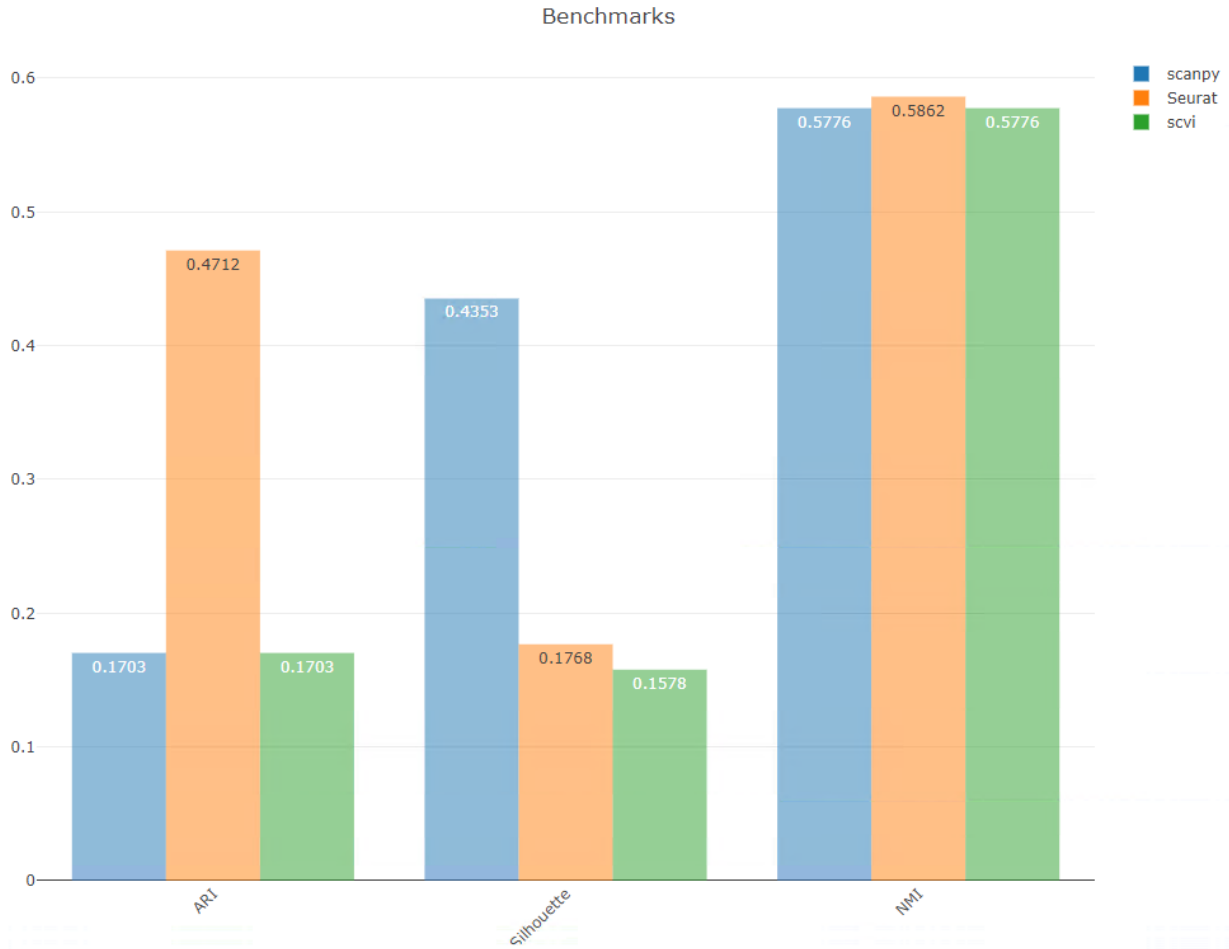


Benchmarks

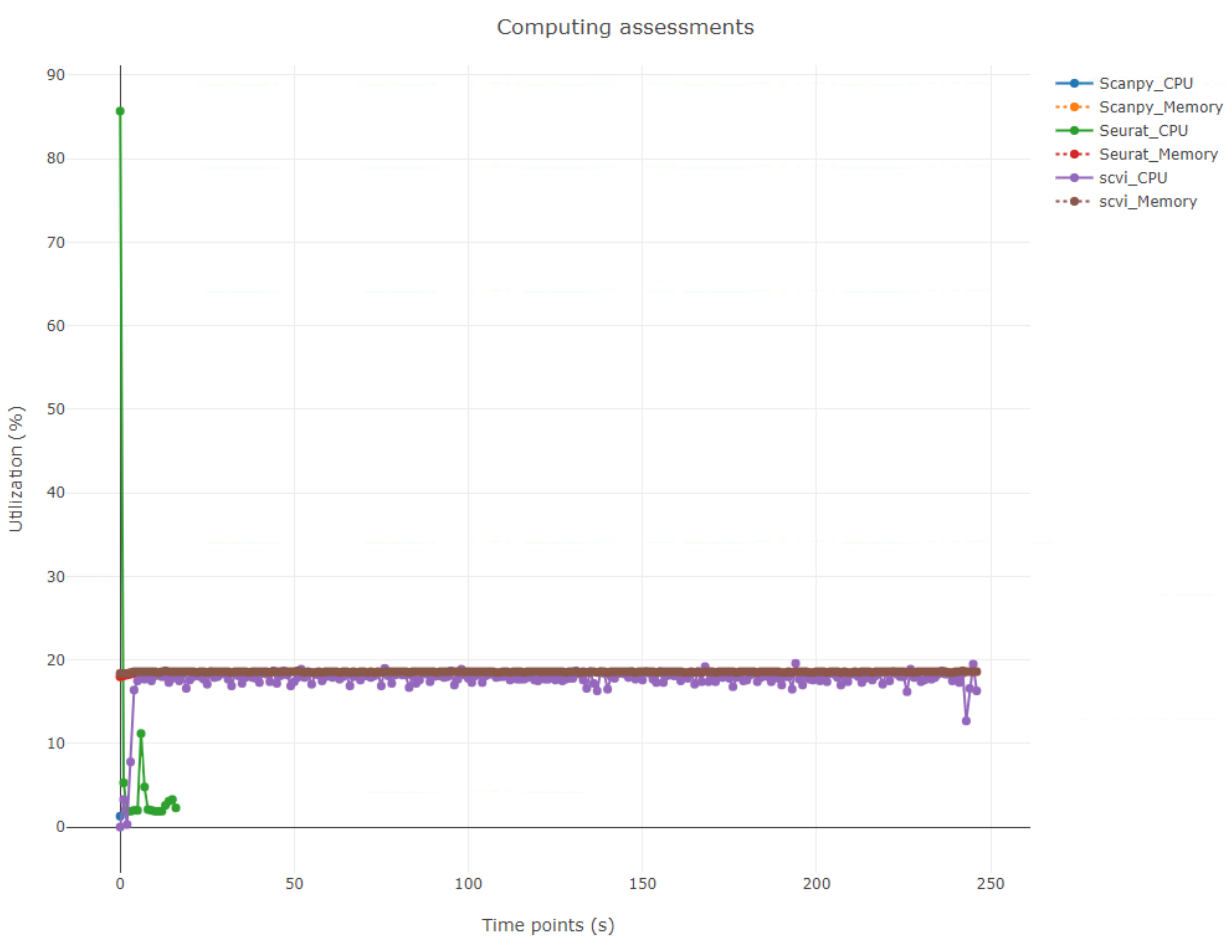
Assessment

Benchmark Results for h-Bladder-Wang-2024

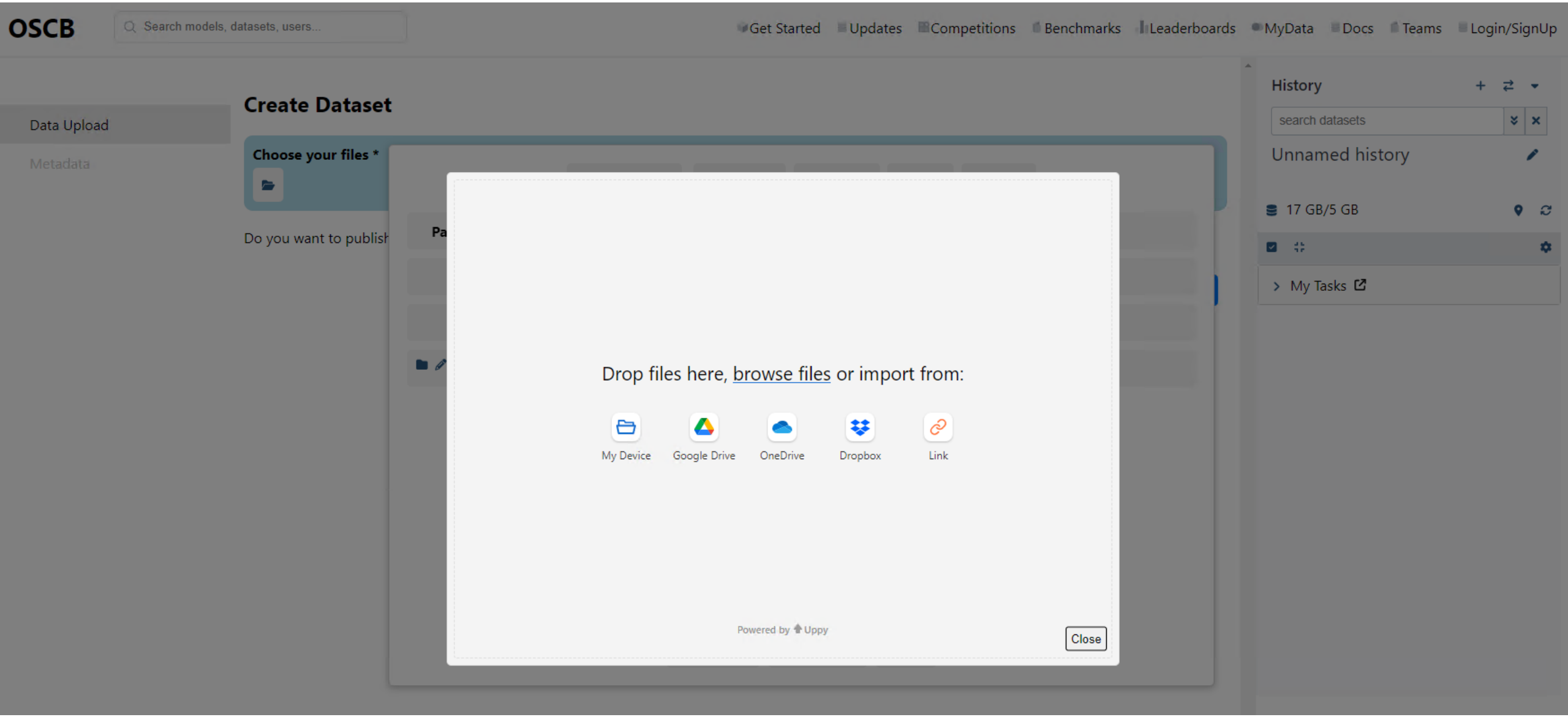
Benchmarks



Computing assessments

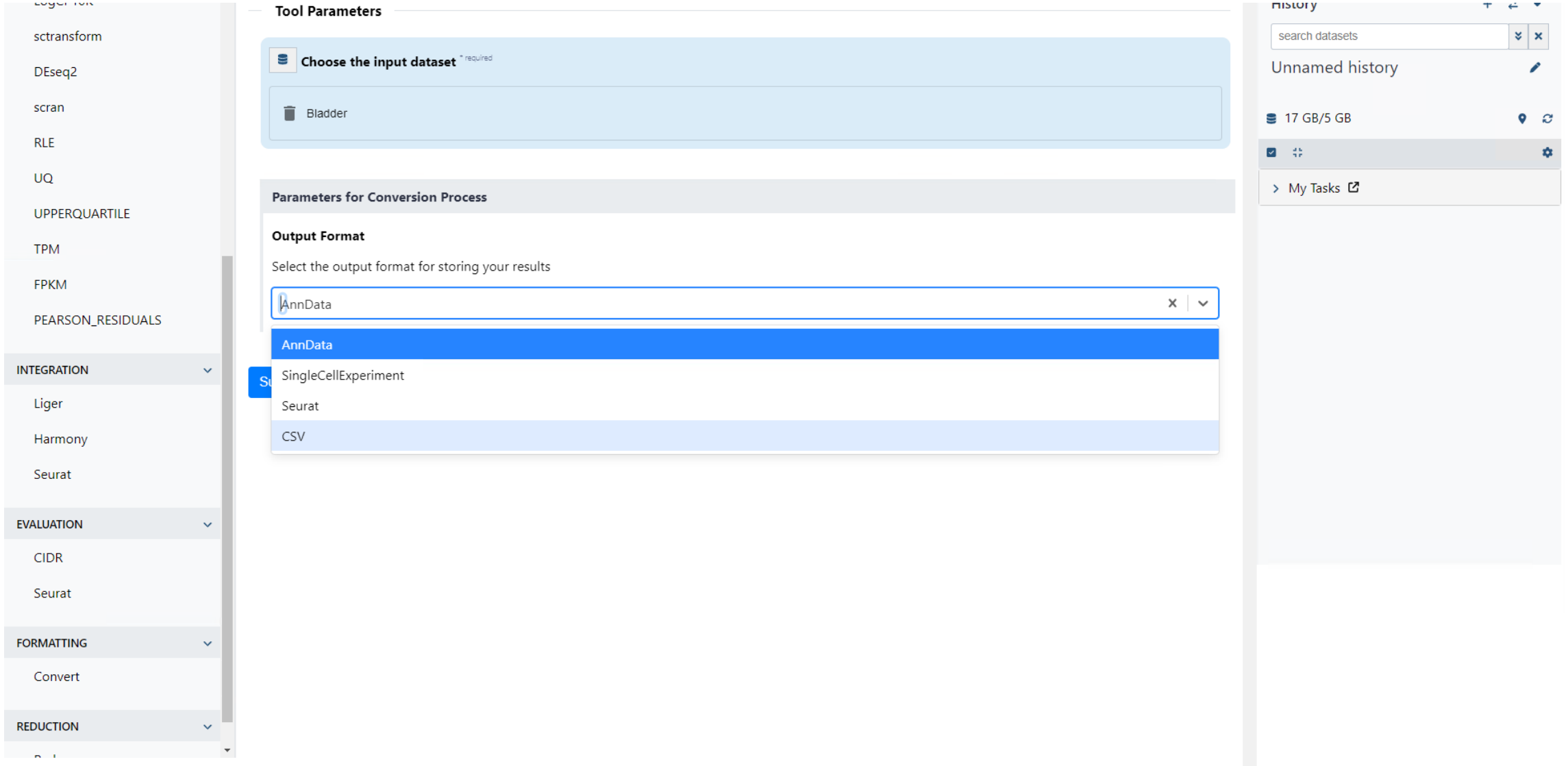


Upload Your Own Datasets



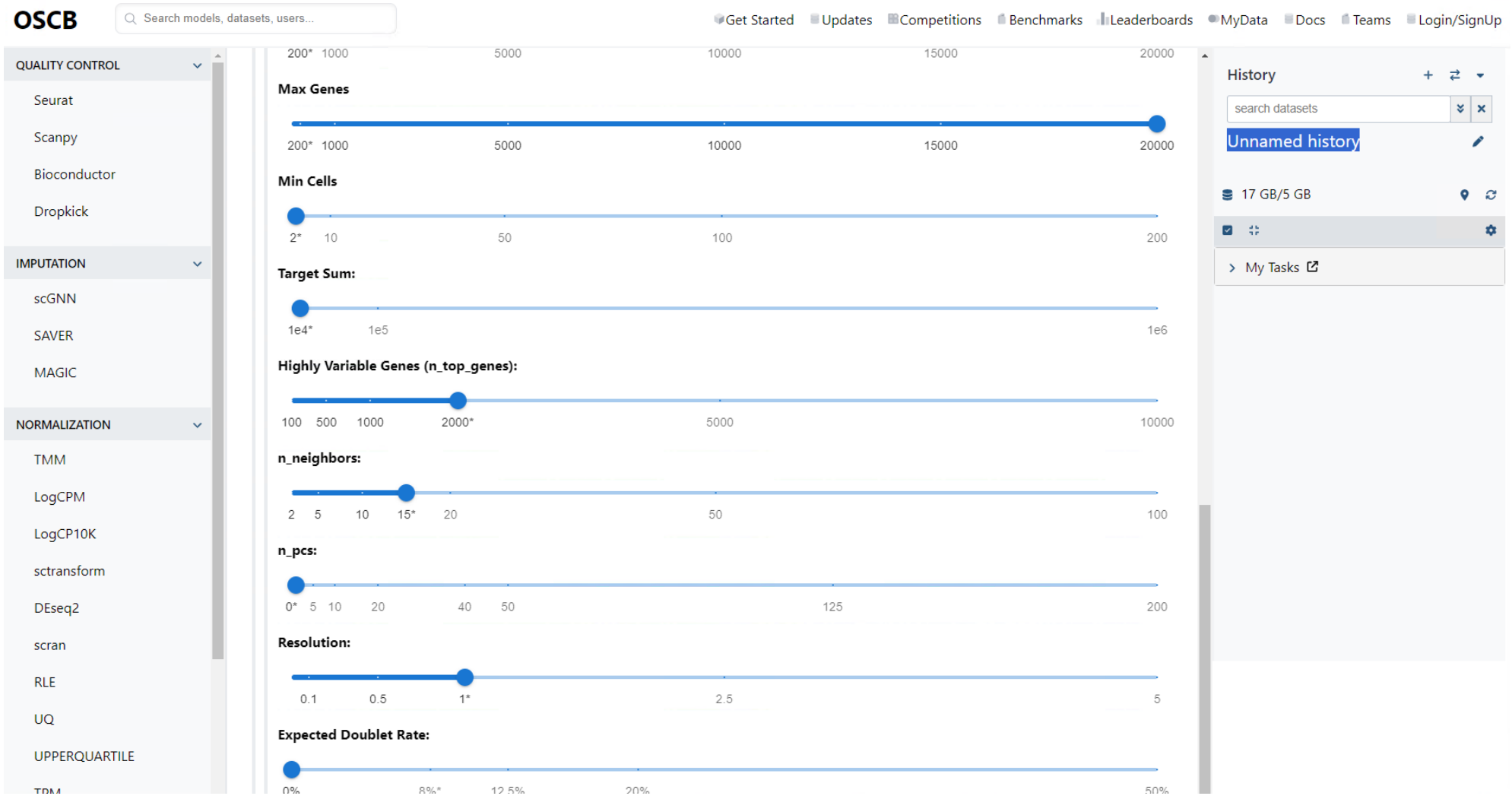
Tools

Formatting



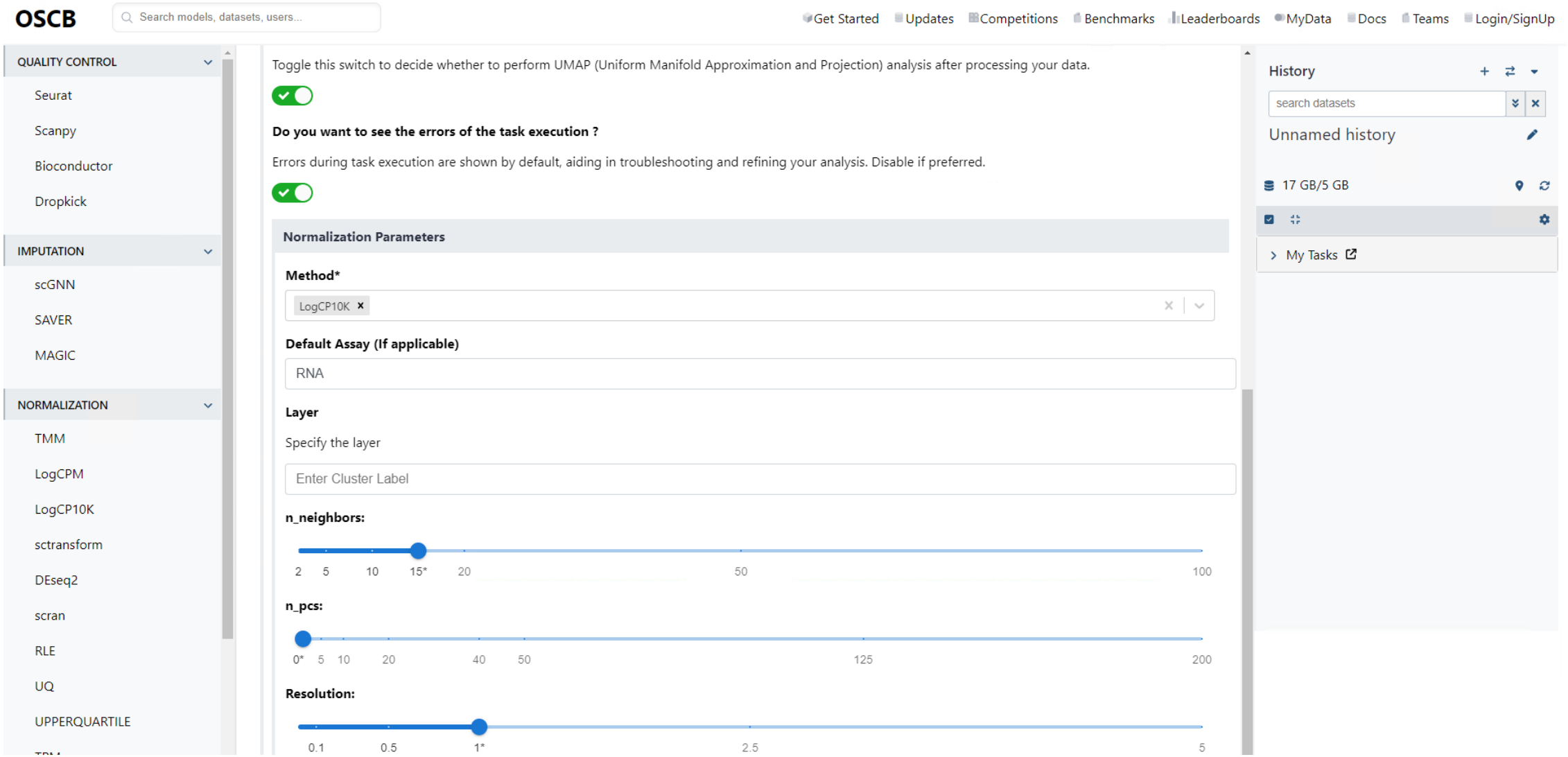
Tools

QC


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
200* 1000 5000 10000 15000 20000
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

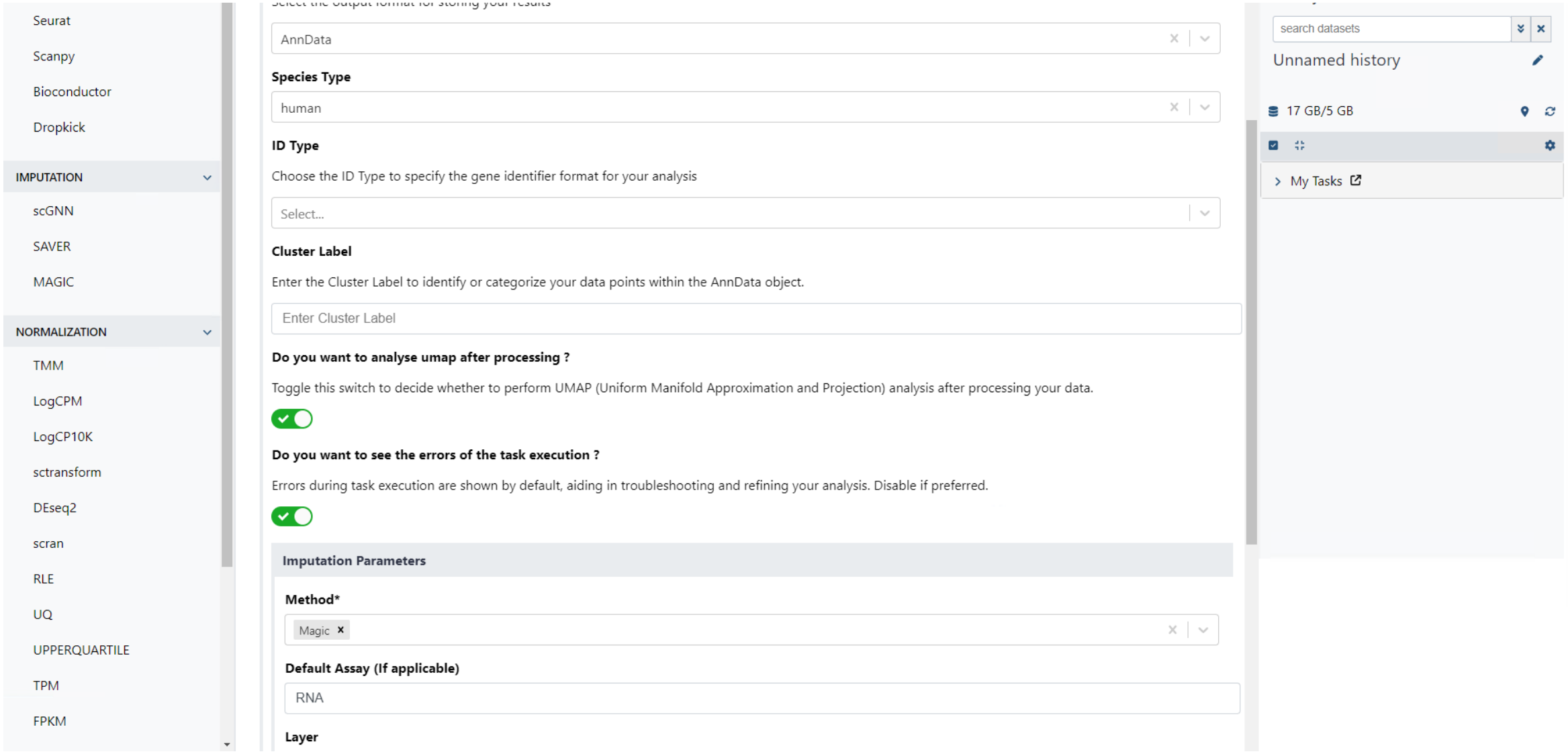
Tools

Normalization



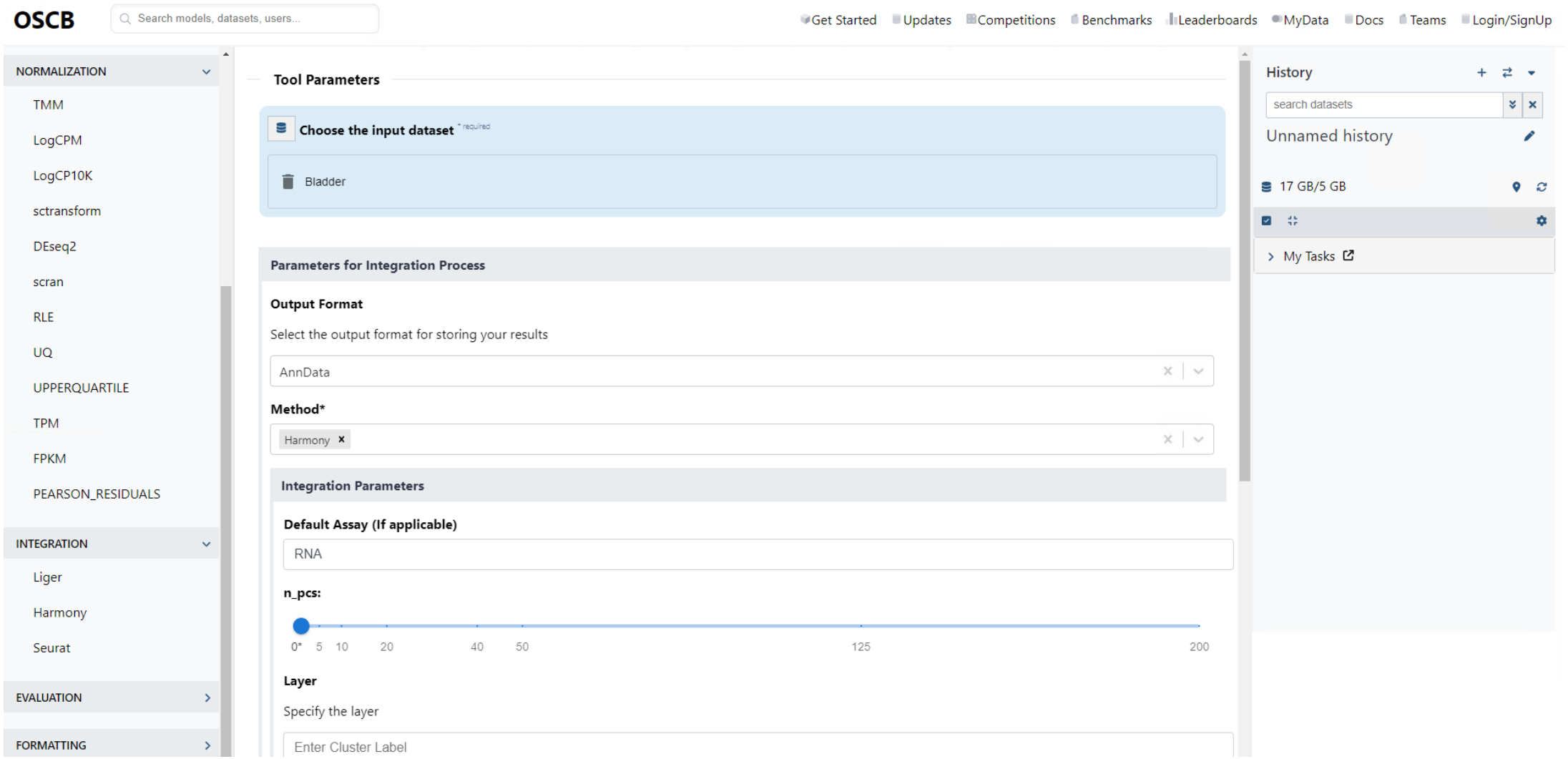
Tools

Imputation



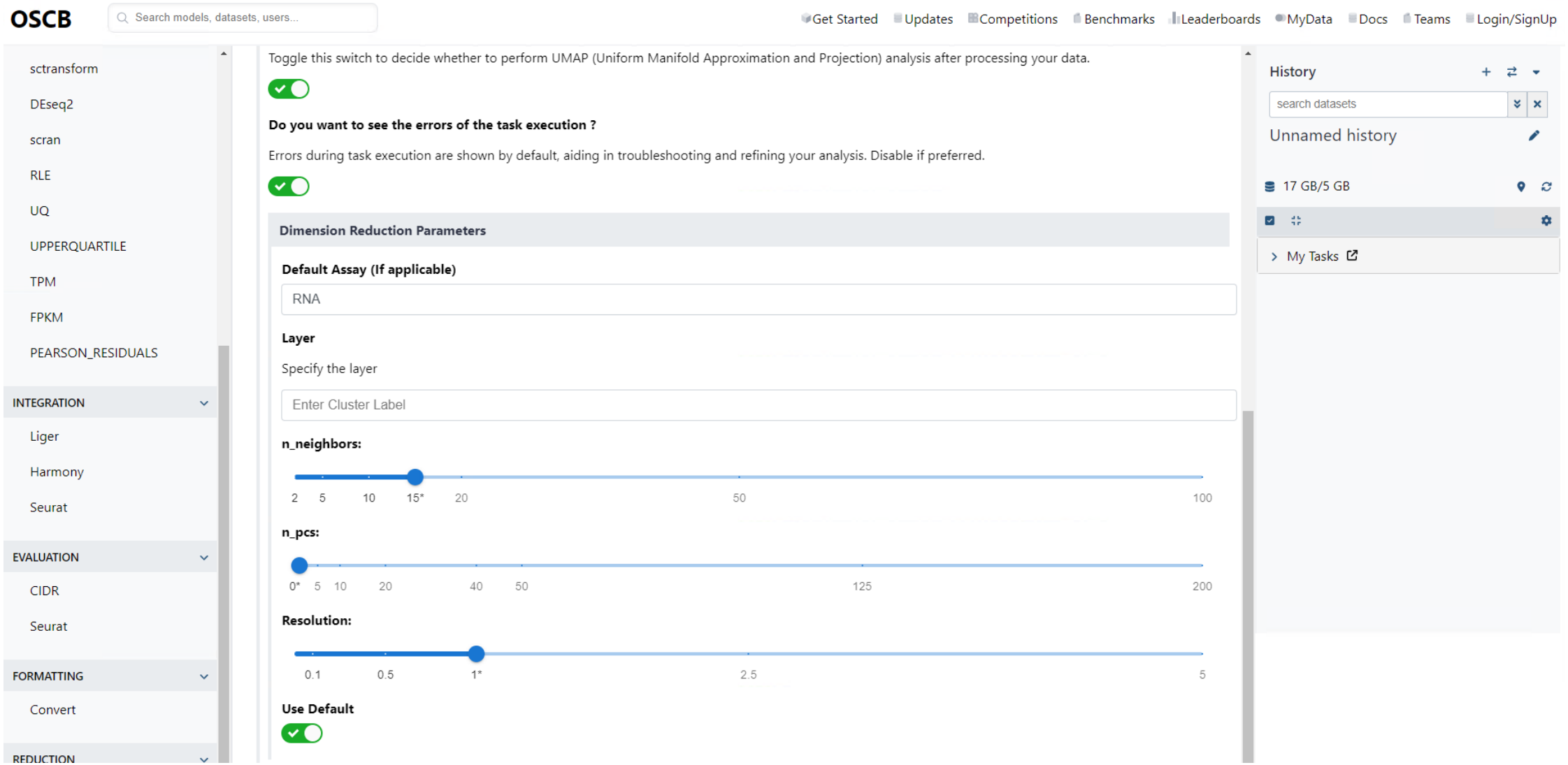
Tools

Integration



Tools

Reduction and Visualization



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

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 × n_vars = 2441 × 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'
```