clustering_comparison

May 18, 2024

1 Clustering Comparision

1.1 Preamble

[1]:
```
import pandas as pd
import numpy as np
import scanpy as sc
from sklearn.metrics.cluster import normalized_mutual_info_score,␣
 ↪adjusted_rand_score
from sklearn.metrics import homogeneity_score, completeness_score,␣
 ↪fowlkes_mallows_score, silhouette_score, davies_bouldin_score
from src.utils import sankey_plot
import kaleido
from sklearn.preprocessing import StandardScaler
import plotly.io as pio
```

[2]:
```
DIR = 'Data/'
DATASET_NAMES = ['PBMC1', 'PBMC2', 'PBMC3','PBMC4']
TOOLS = ['monocle', 'scanpy', 'scvi-tools', 'seurat', 'COTAN']
PARAMS_TUNING = ['default', 'celltypist', 'antibody']
```

[3]:
```
def compute_scores(dir, dataset, labels_df, labels_matched,␣
 ↪ground_truth_labels):
    scores = {}
    scores['NMI'] = {}
    scores['ARI'] = {}
    scores['homogeneity'] = {}
    scores['completeness'] = {}
    scores['fowlkes_mallows'] = {}
    for tool in TOOLS:
        scores['NMI'][tool] =␣
 ↪normalized_mutual_info_score(labels_pred=labels_df['cluster_'+tool],␣
 ↪labels_true=labels_df[f'cluster_{ground_truth_labels}'],␣
 ↪average_method='arithmetic')
        scores['ARI'][tool] =␣
 ↪adjusted_rand_score(labels_pred=labels_df['cluster_'+tool],␣
 ↪labels_true=labels_df[f'cluster_{ground_truth_labels}'])
```

```
        scores['homogeneity'][tool] =␣
↪homogeneity_score(labels_pred=labels_df['cluster_'+tool],␣
↪labels_true=labels_df[f'cluster_{ground_truth_labels}'])
        scores['completeness'][tool] =␣
↪completeness_score(labels_pred=labels_df['cluster_'+tool],␣
↪labels_true=labels_df[f'cluster_{ground_truth_labels}'])
        scores['fowlkes_mallows'][tool] =␣
↪fowlkes_mallows_score(labels_pred=labels_df['cluster_'+tool],␣
↪labels_true=labels_df[f'cluster_{ground_truth_labels}'])
    scores_df = pd.DataFrame(scores)
    scores_df.to_csv(f'{dir}{dataset}/
↪scores_{labels_matched}_{ground_truth_labels}.csv')
    scores_df.to_latex(f'{dir}{dataset}/
↪scores_{labels_matched}_{ground_truth_labels}.tex')
    display(scores_df)


def print_scores(dataset,tuning):


    # concat tools labels
    labels_df = pd.read_csv(f'{DIR}{dataset}/COTAN/{tuning}/clustering_labels.
↪csv', index_col=0)
    labels_df.rename(columns={"cluster": "cluster_COTAN"}, inplace=True)
    for tool in [t for t in TOOLS if t != 'COTAN']:
        tool_labels_df = pd.read_csv(f'{DIR}{dataset}/{tool}/{tuning}/
↪clustering_labels.csv', index_col=0)
        labels_df = labels_df.merge(tool_labels_df, how='inner', on='cell')
        labels_df.rename(columns={"cluster": f"cluster_{tool}"}, inplace=True)

    # load and concat celltypist labels
    celltypist_df = pd.read_csv(f'{DIR}{dataset}/celltypist/celltypist_labels.
↪csv', index_col=0)
    celltypist_df.index = celltypist_df.index.str[:-2]
    celltypist_df = labels_df.merge(celltypist_df, how='inner', on='cell')
    celltypist_df.rename(columns={"cluster.ids": f"cluster_celltypist"},␣
↪inplace=True)
    celltypist_mapping_df = pd.read_csv(f'{DIR}{dataset}/celltypist/
↪celltypist_mapping.csv', index_col=0)

    # load and concat protein surface labels
    antibody_df = pd.read_csv(f'{DIR}{dataset}/antibody_annotation/
↪antibody_labels.csv', index_col=0)
    antibody_df = labels_df.merge(antibody_df, how='inner', on='cell')
    antibody_df.rename(columns={"cluster.ids": f"cluster_antibody"},␣
↪inplace=True)
```

```
    antibody_mapping_df = pd.read_csv(f'{DIR}{dataset}/antibody_annotation/
↪antibody_mapping.csv', index_col=1)

    # read dataset
    adata = sc.read_10x_mtx(
        f'{DIR}{dataset}/filtered/10X/',
        var_names='gene_symbols',
        cache=False
    )
    # keep only labelled cells
    adata.var_names_make_unique()
    subset_cells = adata.obs_names.isin(labels_df.index)
    adata = adata[subset_cells, :]

    mito_genes = adata.var_names.str.startswith('MT-')
    # for each cell compute fraction of counts in mito genes vs. all genes
    # the `.A1` is only necessary as X is sparse (to transform to a dense array␣
↪after summing)
    adata.obs['percent_mito'] = np.sum(adata[:, mito_genes].X, axis=1).A1 / np.
↪sum(adata.X, axis=1).A1
    # add the total counts per cell as observations-annotation to adata
    adata.obs['n_counts'] = adata.X.sum(axis=1).A1

    sc.pp.normalize_total(adata, target_sum=1e4)
    sc.pp.log1p(adata)
    sc.pp.highly_variable_genes(adata, min_mean=0.0125, max_mean=3, min_disp=0.
↪5)
    adata.raw = adata
    adata = adata[:, adata.var.highly_variable]
    sc.pp.regress_out(adata, ['n_counts', 'percent_mito'])
    sc.pp.scale(adata, max_value=10)
    sc.tl.pca(adata, svd_solver='arpack',n_comps=20)
    pca_matrix = adata.obsm['X_pca']
    scaler = StandardScaler()
    scaled_pca_matrix = scaler.fit_transform(pca_matrix)

    #Custers number

    df = {}
    for tool in TOOLS:
        df[tool] = labels_df[f'cluster_{tool}'].unique().shape[0]
    df_size = pd.DataFrame(df, index=[0])
    display(f'{dataset} - number of clusters')
    display(df_size)

    # compute silhouette score
    silhouette = {}
```

```
  for tool in TOOLS:
      silhouette[tool] = silhouette_score(scaled_pca_matrix,␣
↪labels_df[f'cluster_{tool}'])
  if tuning=='celltypist':
      silhouette['celltypist'] = silhouette_score(scaled_pca_matrix,␣
↪celltypist_df[f'cluster_celltypist'])
  elif tuning=='antibody':
      silhouette['antibody'] = silhouette_score(scaled_pca_matrix,␣
↪antibody_df[f'cluster_antibody'])
  silhouette_df = pd.DataFrame(silhouette, index=[0])
  silhouette_df.to_csv(f'{DIR}{dataset}/{tuning}_silhouette.csv')
  silhouette_df.to_latex(f'{DIR}{dataset}/{tuning}_silhouette.tex')
  display(f'{dataset} - Silhuette (higher is better)')
  display(silhouette_df)

  #From https://evafast.github.io/blog/2019/06/28/example_content/
  davies_bouldin = {}
  for tool in TOOLS:
      davies_bouldin[tool] = davies_bouldin_score(adata.obsm['X_pca'],␣
↪labels_df[f'cluster_{tool}'])
  if tuning=='celltypist':
      davies_bouldin['celltypist'] = davies_bouldin_score(adata.
↪obsm['X_pca'], celltypist_df[f'cluster_celltypist'])
  elif tuning=='antibody':
      davies_bouldin['antibody'] = davies_bouldin_score(adata.obsm['X_pca'],␣
↪antibody_df[f'cluster_antibody'])
  davies_bouldin_df = pd.DataFrame(davies_bouldin, index=[0])
  davies_bouldin_df.to_csv(f'{DIR}{dataset}/{tuning}_davies_bouldin.csv')
  davies_bouldin_df.to_latex(f'{DIR}{dataset}/{tuning}_davies_bouldin.tex')
  display(f'{dataset} - davies_bouldin (lower is better)')
  display(davies_bouldin_df)

  display(f'{dataset} - matching {tuning} labels' if tuning != 'default' else␣
↪f'{dataset} - default labels')

  # compute scores comparing each tool labels with celltypist labels
  if tuning == 'celltypist' or tuning == 'default':
      compute_scores(DIR, dataset, celltypist_df, tuning, 'celltypist')
      labels = []
      labels_titles = []
      for tool in TOOLS:
          labels.append(celltypist_df[f'cluster_{tool}'].to_list())
          labels_titles.append(tool)
      labels.append(celltypist_df[f'cluster_celltypist'].
↪map(celltypist_mapping_df['go'].to_dict()).to_list())
      labels_titles.append('celltypist')
```

```
        title = f'{dataset} - matching {tuning} labels' if tuning != 'default'␣
↪else f'{dataset} - default labels'
        sankey_plot(labels=labels, labels_titles=labels_titles, title=title,␣
↪path=f'{DIR}{dataset}/{tuning}_celltypist.html')

    # compute scores comparing each tool labels with protein labels
    if tuning == 'antibody' or tuning == 'default':
        compute_scores(DIR, dataset, antibody_df, tuning, 'antibody')
        labels = []
        labels_titles = []
        for tool in TOOLS:
            labels.append(antibody_df[f'cluster_{tool}'].to_list())
            labels_titles.append(tool)
        labels.append(antibody_df[f'cluster_antibody'].
↪map(antibody_mapping_df['go'].to_dict()).to_list())
        labels_titles.append('antibody')
        title = f'{dataset} - matching {tuning} labels' if tuning != 'default'␣
↪else f'{dataset} - default labels'
        sankey_plot(labels=labels, labels_titles=labels_titles, title=title,␣
↪path=f'{DIR}{dataset}/{tuning}_antibody.html')
```

1.2 Default parameters

[169]:
```
print_scores(tuning = 'default',dataset="PBMC1")
```

```
/tmp/ipykernel_70563/2944004898.py:58: ImplicitModificationWarning:

Trying to modify attribute `.obs` of view, initializing view as actual.


'PBMC1 - number of clusters'
```

	monocle	scanpy	scvi-tools	seurat	COTAN
0	3	18	13	11	14

```
'PBMC1 - Silhuette (higher is better)'
```

	monocle	scanpy	scvi-tools	seurat	COTAN
0	0.100032	0.043761	0.065534	0.148254	0.13383

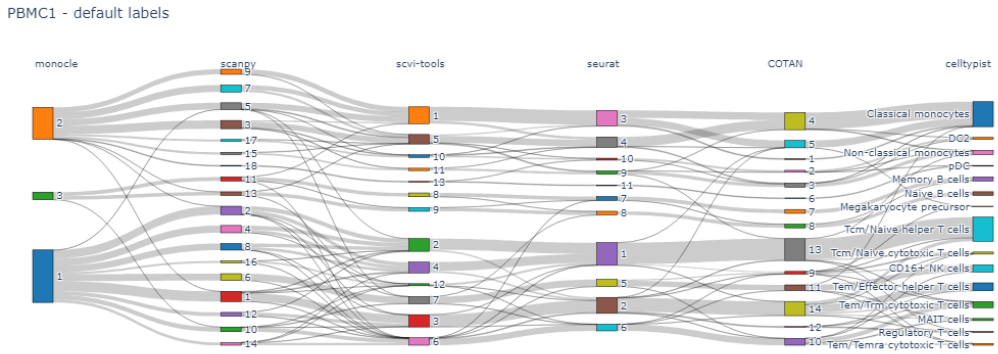
```
'PBMC1 - davies_bouldin (lower is better)'
```

	monocle	scanpy	scvi-tools	seurat	COTAN
0	0.893661	2.47358	2.574291	1.392309	1.728304

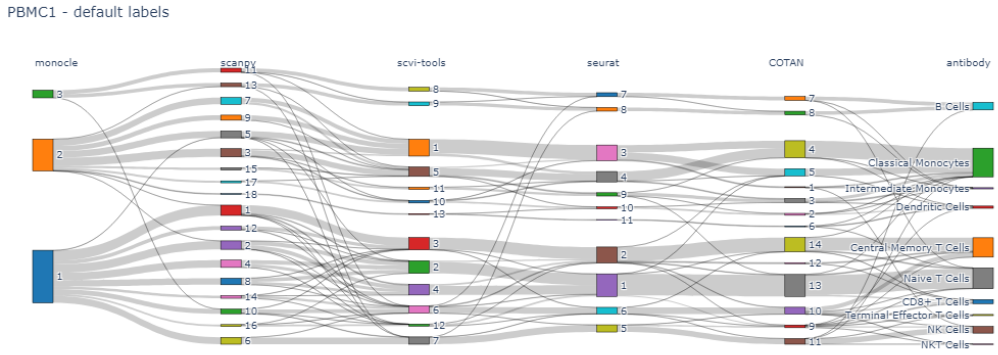
```
'PBMC1 - default labels'
```

	NMI	ARI	homogeneity	completeness	fowlkes_mallows
monocle	0.578257	0.384609	0.410140	0.979930	0.602512
scanpy	0.721042	0.404607	0.824980	0.640363	0.508176
scvi-tools	0.776232	0.599664	0.809790	0.745344	0.666244

seurat	0.793630	0.649593	0.784165	0.803327	0.705921
COTAN	0.787289	0.670392	0.803876	0.771373	0.723485



	NMI	ARI	homogeneity	completeness	fowlkes_mallows
monocle	0.611988	0.425929	0.446299	0.973344	0.635502
scanpy	0.659645	0.391203	0.795577	0.563386	0.507730
scvi-tools	0.708581	0.551051	0.776228	0.651780	0.632750
seurat	0.738344	0.643097	0.764146	0.714228	0.706018
COTAN	0.732140	0.651092	0.784252	0.686521	0.713737



[12]:
```
print_scores(tuning = 'default',dataset="PBMC2")
```

```
/tmp/ipykernel_8461/2944004898.py:58: ImplicitModificationWarning:

Trying to modify attribute `.obs` of view, initializing view as actual.
```

'PBMC2 - number of clusters'

	monocle	scanpy	scvi-tools	seurat	COTAN
0	2	18	20	14	19

'PBMC2 - Silhuette (higher is better)'

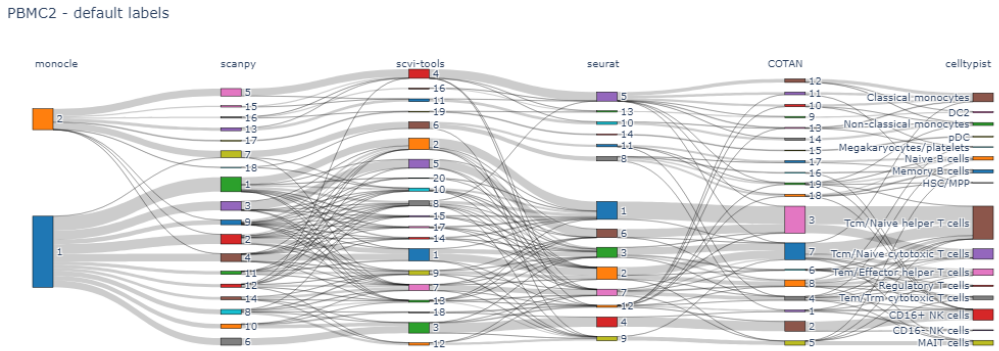
	monocle	scanpy	scvi-tools	seurat	COTAN
0	0.224441	0.059225	0.000832	0.111509	0.101869

'PBMC2 - davies_bouldin (lower is better)'

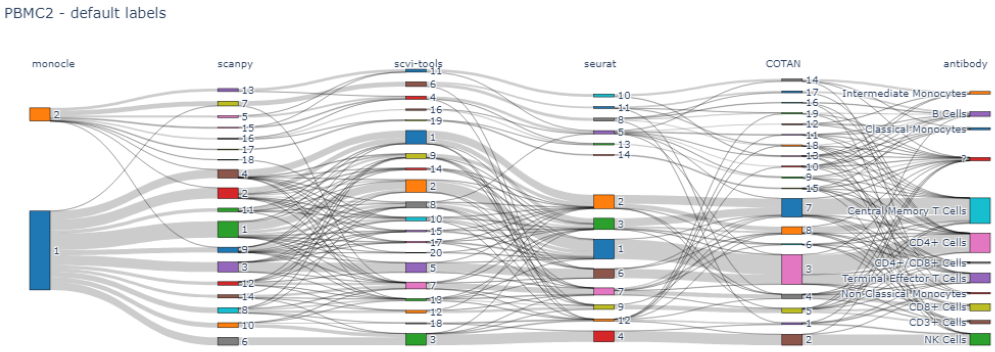
	monocle	scanpy	scvi-tools	seurat	COTAN
0	1.866736	2.073818	3.935702	1.630557	2.241596

'PBMC2 - default labels'

	NMI	ARI	homogeneity	completeness	fowlkes_mallows
monocle	0.393166	0.207180	0.245998	0.978626	0.521364
scanpy	0.718820	0.457213	0.804000	0.649960	0.556684
scvi-tools	0.699788	0.424696	0.785920	0.630670	0.525031
seurat	0.775988	0.562430	0.819560	0.736815	0.640108
COTAN	0.689037	0.441866	0.724779	0.656654	0.533910



	NMI	ARI	homogeneity	completeness	fowlkes_mallows
monocle	0.277314	0.107135	0.165534	0.853977	0.450594
scanpy	0.682604	0.524109	0.759388	0.619922	0.602311
scvi-tools	0.652891	0.485961	0.734303	0.587729	0.567847
seurat	0.743681	0.679941	0.777650	0.712555	0.730603
COTAN	0.653065	0.544334	0.662016	0.644353	0.621068



[10]:
```
print_scores(tuning = 'default',dataset="PBMC3")
```

```
/tmp/ipykernel_8461/2944004898.py:58: ImplicitModificationWarning:

Trying to modify attribute `.obs` of view, initializing view as actual.
```

'PBMC3 - number of clusters'

	monocle	scanpy	scvi-tools	seurat	COTAN
0	3	22	17	18	32

'PBMC3 - Silhuette (higher is better)'

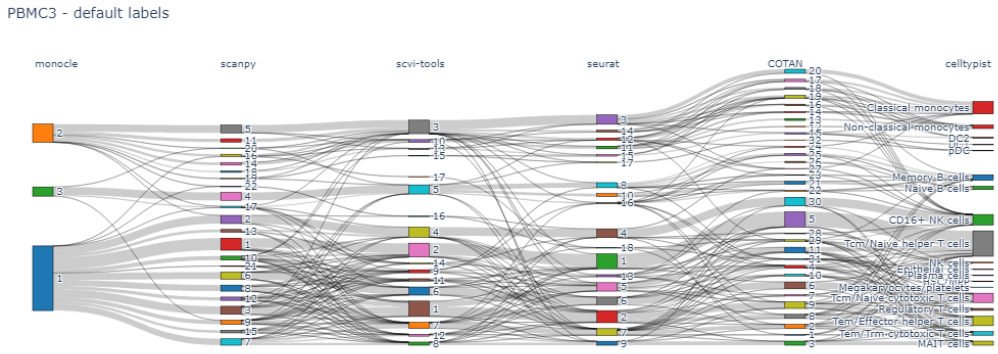
	monocle	scanpy	scvi-tools	seurat	COTAN
0	0.171584	0.007616	0.055559	0.111834	0.092445

'PBMC3 - davies_bouldin (lower is better)'

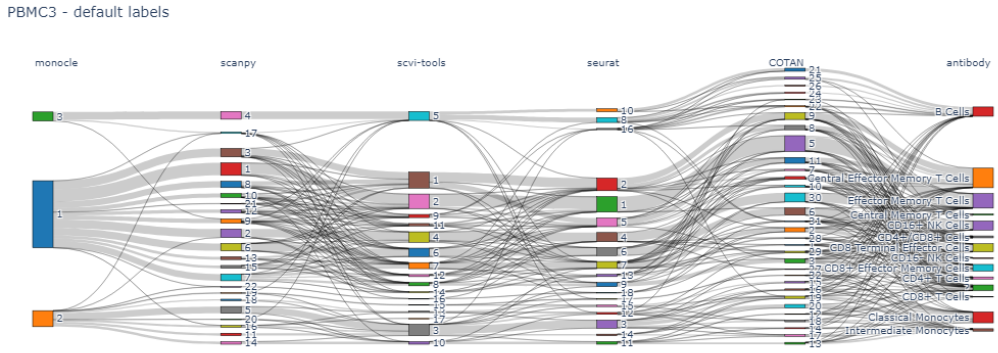
	monocle	scanpy	scvi-tools	seurat	COTAN
0	0.999779	2.376915	2.058343	1.698551	2.281481

'PBMC3 - default labels'

	NMI	ARI	homogeneity	completeness	fowlkes_mallows
monocle	0.500696	0.233560	0.338609	0.960446	0.500077
scanpy	0.685919	0.462762	0.763719	0.622505	0.541286
scvi-tools	0.738418	0.579677	0.757237	0.720511	0.635237
seurat	0.770512	0.585110	0.821173	0.725738	0.644073
COTAN	0.723833	0.527470	0.849029	0.630815	0.609217



	NMI	ARI	homogeneity	completeness	fowlkes_mallows
monocle	0.429744	0.168276	0.280823	0.914939	0.437511
scanpy	0.664567	0.542630	0.702416	0.630588	0.596647
scvi-tools	0.691391	0.620339	0.677930	0.705398	0.662580
seurat	0.735217	0.664188	0.744324	0.726330	0.701375
COTAN	0.670484	0.541727	0.743155	0.610759	0.598370



[4]:
```
print_scores(tuning = 'default',dataset="PBMC4")
```

```
/tmp/ipykernel_8461/2944004898.py:58: ImplicitModificationWarning:

Trying to modify attribute `.obs` of view, initializing view as actual.

'PBMC4 - number of clusters'

   monocle  scanpy  scvi-tools  seurat  COTAN
```

```
0        3      22          16      19     24
```

```
'PBMC4 - Silhuette (higher is better)'
```

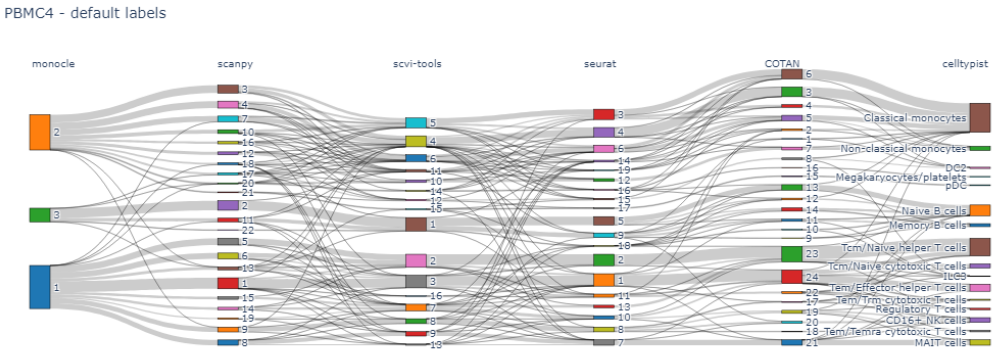
```
    monocle    scanpy  scvi-tools    seurat     COTAN
0  0.081765  0.050853    0.061618  0.112255  0.103418
```

```
'PBMC4 - davies_bouldin (lower is better)'
```

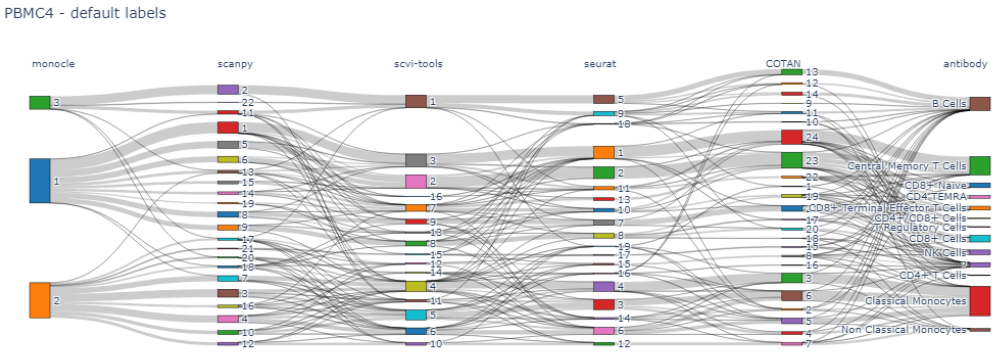
```
    monocle    scanpy  scvi-tools    seurat     COTAN
0  0.954689  2.100956    2.087707  1.442075  1.823095
```

```
'PBMC4 - default labels'
```

	NMI	ARI	homogeneity	completeness	fowlkes_mallows
monocle	0.617025	0.470070	0.453383	0.965513	0.647279
scanpy	0.701228	0.380357	0.819943	0.612541	0.487560
scvi-tools	0.739299	0.504966	0.788229	0.696088	0.584900
seurat	0.760207	0.494746	0.847372	0.689301	0.583823
COTAN	0.716555	0.435917	0.808134	0.643618	0.526063



	NMI	ARI	homogeneity	completeness	fowlkes_mallows
monocle	0.536861	0.325029	0.372515	0.960701	0.532810
scanpy	0.622945	0.371655	0.659143	0.590516	0.439575
scvi-tools	0.651550	0.425369	0.634107	0.669980	0.487767
seurat	0.669274	0.436706	0.676741	0.661971	0.496402
COTAN	0.622356	0.367136	0.635352	0.609880	0.433784



1.3 Matching cellTypist clusters number

[173]:
```
print_scores(tuning = 'celltypist',dataset="PBMC1")
```

```
/tmp/ipykernel_70563/2944004898.py:58: ImplicitModificationWarning:

Trying to modify attribute `.obs` of view, initializing view as actual.
```

```
'PBMC1 - number of clusters'
```

	monocle	scanpy	scvi-tools	seurat	COTAN
0	18	17	20	21	14

```
'PBMC1 - Silhuette (higher is better)'
```

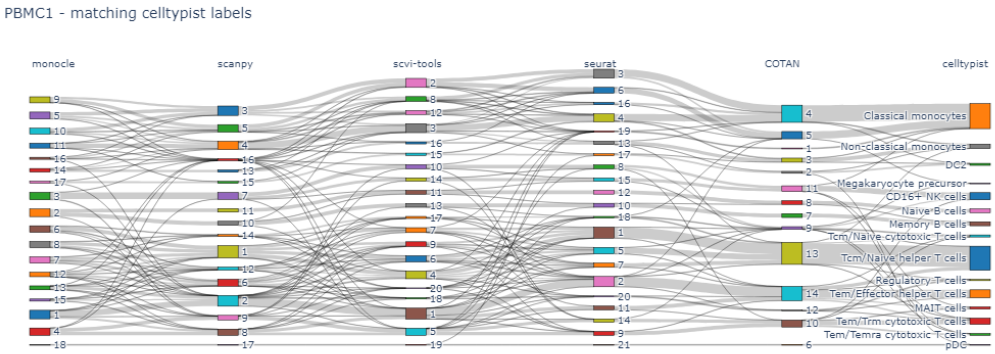
	monocle	scanpy	scvi-tools	seurat	COTAN	celltypist
0	0.005501	0.050843	0.063871	0.088209	0.13383	0.090989

```
'PBMC1 - davies_bouldin (lower is better)'
```

	monocle	scanpy	scvi-tools	seurat	COTAN	celltypist
0	2.83457	2.286672	2.797123	1.984746	1.728304	1.491801

```
'PBMC1 - matching celltypist labels'
```

	NMI	ARI	homogeneity	completeness	fowlkes_mallows
monocle	0.658065	0.341945	0.757164	0.581903	0.448601
scanpy	0.735830	0.459735	0.822412	0.665742	0.553086
scvi-tools	0.699950	0.375082	0.808964	0.616828	0.479652
seurat	0.730468	0.423069	0.849278	0.640820	0.527386
COTAN	0.787289	0.670392	0.803876	0.771373	0.723485



[174]:
```
print_scores(tuning = 'celltypist',dataset="PBMC2")
```

```
/tmp/ipykernel_70563/2944004898.py:58: ImplicitModificationWarning:

Trying to modify attribute `.obs` of view, initializing view as actual.
```

```
'PBMC2 - number of clusters'
```

	monocle	scanpy	scvi-tools	seurat	COTAN
0	18	20	19	20	17

```
'PBMC2 - Silhuette (higher is better)'
```

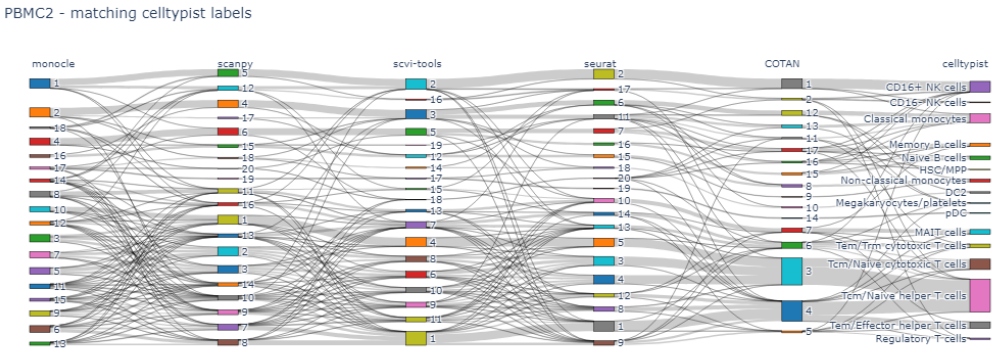
	monocle	scanpy	scvi-tools	seurat	COTAN	celltypist
0	-0.03374	0.025173	0.030773	0.060901	0.12581	0.131097

```
'PBMC2 - davies_bouldin (lower is better)'
```

	monocle	scanpy	scvi-tools	seurat	COTAN	celltypist
0	3.167255	2.376025	3.431314	2.147939	1.855098	1.231923

```
'PBMC2 - matching celltypist labels'
```

	NMI	ARI	homogeneity	completeness	fowlkes_mallows
monocle	0.605942	0.312112	0.699644	0.534375	0.425421
scanpy	0.697287	0.377675	0.809335	0.612491	0.492889
scvi-tools	0.709450	0.398779	0.791930	0.642531	0.500807
seurat	0.738307	0.418942	0.850535	0.652244	0.529176
COTAN	0.729355	0.472800	0.745550	0.713848	0.562480



[11]:
```
print_scores(tuning = 'celltypist',dataset="PBMC3")
```

```
/tmp/ipykernel_8461/2944004898.py:58: ImplicitModificationWarning:

Trying to modify attribute `.obs` of view, initializing view as actual.
```

'PBMC3 - number of clusters'

	monocle	scanpy	scvi-tools	seurat	COTAN
0	17	18	20	18	21

'PBMC3 - Silhuette (higher is better)'

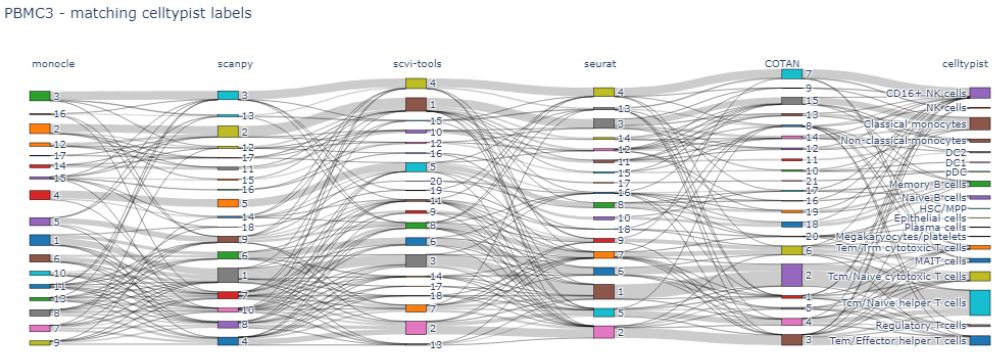
	monocle	scanpy	scvi-tools	seurat	COTAN	celltypist
0	-0.039293	0.040964	0.003634	0.112264	0.047585	0.130032

'PBMC3 - davies_bouldin (lower is better)'

	monocle	scanpy	scvi-tools	seurat	COTAN	celltypist
0	3.778924	1.889045	2.221019	1.698481	2.344713	1.140713

'PBMC3 - matching celltypist labels'

	NMI	ARI	homogeneity	completeness	fowlkes_mallows
monocle	0.593459	0.350206	0.643738	0.550465	0.432058
scanpy	0.712344	0.545918	0.758076	0.671816	0.609354
scvi-tools	0.735127	0.565025	0.767444	0.705423	0.623277
seurat	0.771047	0.586941	0.821567	0.726381	0.645653
COTAN	0.677393	0.470051	0.717462	0.641563	0.538494



[5]:
```
print_scores(tuning = 'celltypist',dataset="PBMC4")
```

```
/tmp/ipykernel_8461/2944004898.py:58: ImplicitModificationWarning:

Trying to modify attribute `.obs` of view, initializing view as actual.
```

'PBMC4 - number of clusters'

	monocle	scanpy	scvi-tools	seurat	COTAN
0	16	18	18	19	19

'PBMC4 - Silhuette (higher is better)'

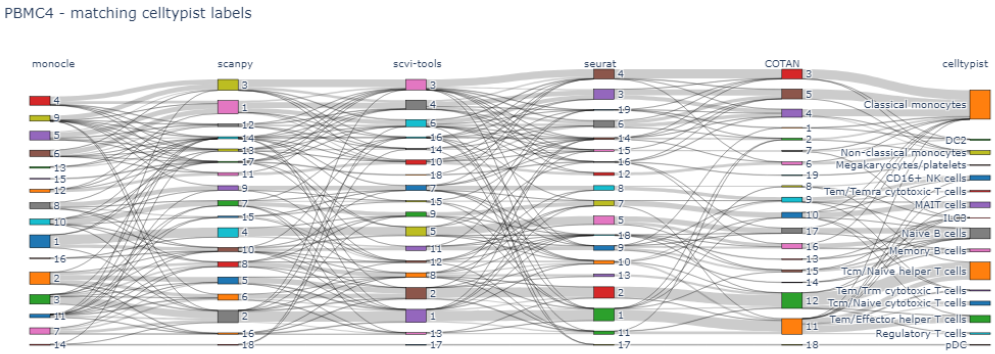
	monocle	scanpy	scvi-tools	seurat	COTAN	celltypist
0	0.022385	0.048061	0.107921	0.111704	0.098387	0.081772

'PBMC4 - davies_bouldin (lower is better)'

	monocle	scanpy	scvi-tools	seurat	COTAN	celltypist
0	2.007518	1.898663	1.612564	1.442525	1.969481	1.194603

'PBMC4 - matching celltypist labels'

	NMI	ARI	homogeneity	completeness	fowlkes_mallows
monocle	0.686019	0.421166	0.747399	0.633956	0.512728
scanpy	0.730100	0.473433	0.810168	0.664434	0.562407
scvi-tools	0.752718	0.500477	0.831022	0.687899	0.587099
seurat	0.759495	0.492528	0.846776	0.688525	0.581840
COTAN	0.724052	0.447103	0.782942	0.673402	0.532939



1.4 Matching antibody clusters number

[177]:
```
print_scores(tuning = 'antibody',dataset="PBMC1")
```

```
/tmp/ipykernel_70563/2944004898.py:58: ImplicitModificationWarning:

Trying to modify attribute `.obs` of view, initializing view as actual.
```

```
'PBMC1 - number of clusters'
```

	monocle	scanpy	scvi-tools	seurat	COTAN
0	9	11	10	11	11

```
'PBMC1 - Silhuette (higher is better)'
```

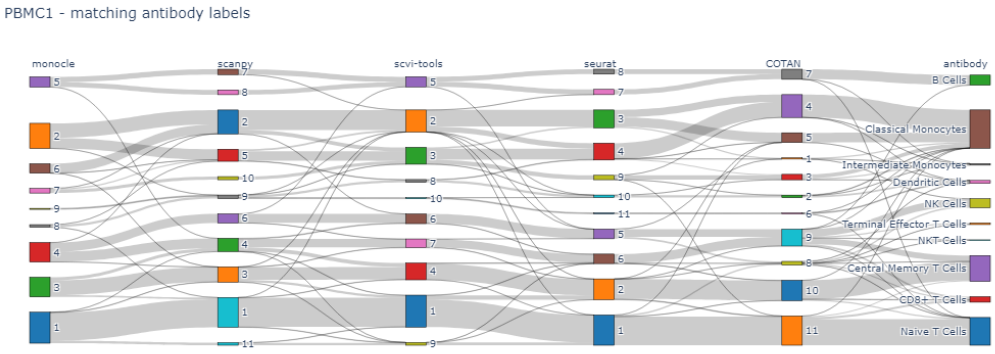
	monocle	scanpy	scvi-tools	seurat	COTAN	antibody
0	0.099643	0.073115	0.069687	0.150193	0.090531	0.042617

```
'PBMC1 - davies_bouldin (lower is better)'
```

	monocle	scanpy	scvi-tools	seurat	COTAN	antibody
0	1.498221	1.689467	1.560514	1.392406	2.067883	1.721174

```
'PBMC1 - matching antibody labels'
```

	NMI	ARI	homogeneity	completeness	fowlkes_mallows
monocle	0.715777	0.633217	0.707203	0.724563	0.698798
scanpy	0.736466	0.645073	0.769273	0.706342	0.707927
scvi-tools	0.746355	0.650567	0.757864	0.735189	0.712214
seurat	0.739813	0.640616	0.767621	0.713949	0.704048
COTAN	0.721550	0.649977	0.737446	0.706326	0.711698



[178]:
```
print_scores(tuning = 'antibody',dataset="PBMC2")
```

```
/tmp/ipykernel_70563/2944004898.py:58: ImplicitModificationWarning:

Trying to modify attribute `.obs` of view, initializing view as actual.
```

```
'PBMC2 - number of clusters'
```

	monocle	scanpy	scvi-tools	seurat	COTAN
0	11	10	12	12	12

```
'PBMC2 - Silhuette (higher is better)'
```

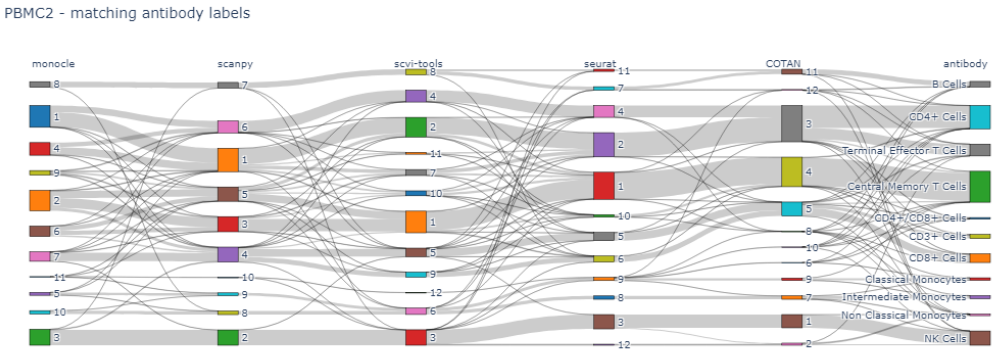
	monocle	scanpy	scvi-tools	seurat	COTAN	antibody
0	-0.041184	0.037491	-0.021431	0.091472	0.074925	0.04853

```
'PBMC2 - davies_bouldin (lower is better)'
```

	monocle	scanpy	scvi-tools	seurat	COTAN	antibody
0	3.22287	1.702317	4.327977	1.463008	2.077179	2.208843

```
'PBMC2 - matching antibody labels'
```

	NMI	ARI	homogeneity	completeness	fowlkes_mallows
monocle	0.585175	0.443933	0.599091	0.571891	0.531050
scanpy	0.739559	0.636764	0.742414	0.736726	0.695610
scvi-tools	0.666315	0.562966	0.694832	0.640047	0.632909
seurat	0.763984	0.764614	0.773986	0.754238	0.803575
COTAN	0.740643	0.676510	0.686522	0.804027	0.746725



[185]:
```
print_scores(tuning = 'antibody',dataset="PBMC3")
```

```
/tmp/ipykernel_70563/2944004898.py:58: ImplicitModificationWarning:

Trying to modify attribute `.obs` of view, initializing view as actual.
```

```
'PBMC3 - number of clusters'
```

	monocle	scanpy	scvi-tools	seurat	COTAN
0	12	14	13	14	12

```
'PBMC3 - Silhuette (higher is better)'
```

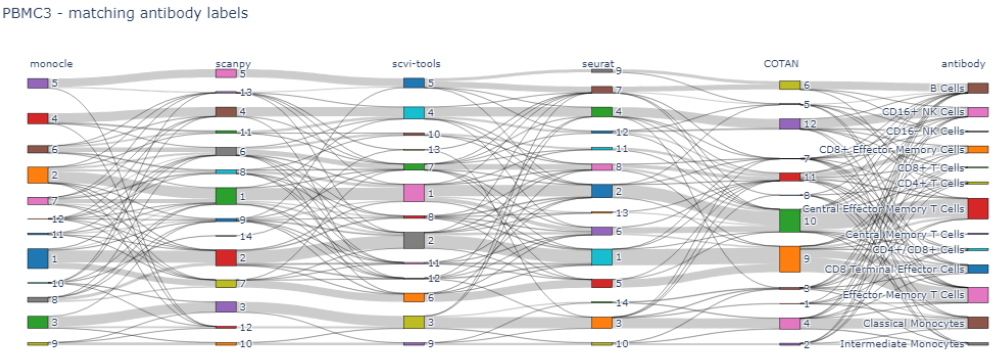
	monocle	scanpy	scvi-tools	seurat	COTAN	antibody
0	-0.048747	0.024448	-0.007328	0.063501	0.05861	0.031805

```
'PBMC3 - davies_bouldin (lower is better)'
```

	monocle	scanpy	scvi-tools	seurat	COTAN	antibody
0	3.082757	2.048357	2.967632	1.617082	1.813171	3.287867

```
'PBMC3 - matching antibody labels'
```

	NMI	ARI	homogeneity	completeness	fowlkes_mallows
monocle	0.633333	0.503352	0.609373	0.659256	0.567586
scanpy	0.737439	0.695008	0.736412	0.738468	0.732379
scvi-tools	0.712286	0.638310	0.692394	0.733355	0.684119
seurat	0.760367	0.695131	0.769539	0.751412	0.732928
COTAN	0.701448	0.613050	0.634888	0.783598	0.683175



[6]:
```
print_scores(tuning = 'antibody',dataset="PBMC4")
```

```
/tmp/ipykernel_8461/2944004898.py:58: ImplicitModificationWarning:

Trying to modify attribute `.obs` of view, initializing view as actual.
```

'PBMC4 - number of clusters'

	monocle	scanpy	scvi-tools	seurat	COTAN
0	13	11	11	13	12

'PBMC4 - Silhuette (higher is better)'

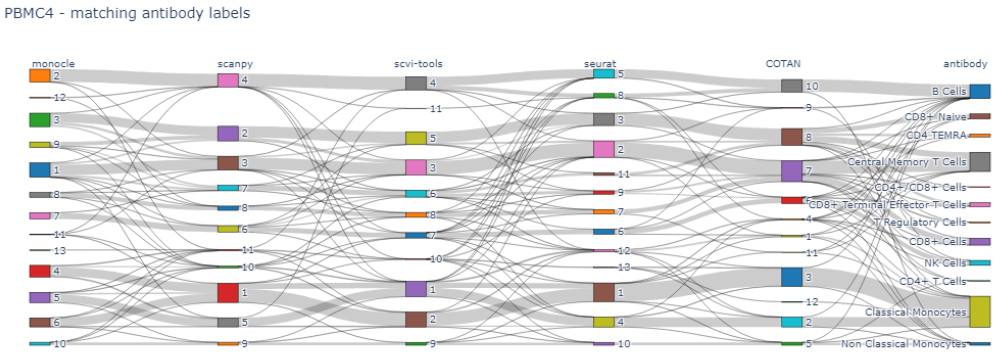
	monocle	scanpy	scvi-tools	seurat	COTAN	antibody
0	-0.009349	0.036425	0.038929	0.059369	0.03044	-0.038177

'PBMC4 - davies_bouldin (lower is better)'

	monocle	scanpy	scvi-tools	seurat	COTAN	antibody
0	2.412446	1.784603	1.670777	1.621226	1.822676	9.337289

'PBMC4 - matching antibody labels'

	NMI	ARI	homogeneity	completeness	fowlkes_mallows
monocle	0.606252	0.394823	0.583563	0.630776	0.467547
scanpy	0.684939	0.495584	0.645785	0.729147	0.560952
scvi-tools	0.673692	0.484543	0.629449	0.724624	0.551156
seurat	0.687849	0.514398	0.665613	0.711622	0.574629
COTAN	0.648815	0.441054	0.584713	0.728703	0.525102



1.5 Check cellTypist vs Antibody

[181]:
```
def compute_clustering_scores(celltypist_df, antibody_df, output_dir, dataset):
    # Merge the dataframes on the common 'cell' column
    #cotan_df = pd.read_csv(f'{DIR}{dataset}/COTAN/antibody/clustering_labels.
 ↪csv', index_col=0)
    #display("Cotan clusters objetc dimension ",cotan_df.shape)
    #display("----------------------------------------")

    celltypist_df = pd.read_csv(f'{DIR}{dataset}/celltypist/celltypist_labels.
 ↪csv', index_col=0)
    celltypist_df.index = celltypist_df.index.str[:-2]
    antibody_df = pd.read_csv(f'{DIR}{dataset}/antibody_annotation/
 ↪antibody_labels.csv', index_col=0)
    #antibody_df = labels_df.merge(antibody_df, how='inner', on='cell')
    #all_in_antibody = celltypist_df.index.isin(antibody_df.index).all()
    #all_in_celltypist = antibody_df.index.isin(celltypist_df.index).all()

    #display("All celltypist indices in antibody: ",all_in_antibody,␣
 ↪celltypist_df.index.isin(antibody_df.index).sum(),celltypist_df.shape)
    #display("All antibody indices in cellTypist:", all_in_celltypist)

    #display("----------------------------------------")

    merged_df = celltypist_df.merge(antibody_df, how='inner',left_index=True,␣
 ↪right_index=True)# on='cell')

    merged_df.columns = ['cluster_celltypist','cluster_antibody']

    # Initialize scores dictionary
    scores = {
```

```
        'NMI': normalized_mutual_info_score(merged_df['cluster_celltypist'],␣
↪merged_df['cluster_antibody'], average_method='arithmetic'),
        'ARI': adjusted_rand_score(merged_df['cluster_celltypist'],␣
↪merged_df['cluster_antibody']),
        'Homogeneity': homogeneity_score(merged_df['cluster_celltypist'],␣
↪merged_df['cluster_antibody']),
        'Completeness': completeness_score(merged_df['cluster_celltypist'],␣
↪merged_df['cluster_antibody']),
        'Fowlkes_Mallows':␣
↪fowlkes_mallows_score(merged_df['cluster_celltypist'],␣
↪merged_df['cluster_antibody'])
    }

    # Convert scores to DataFrame
    scores_df = pd.DataFrame([scores])

    # Save scores to CSV and LaTeX
    #scores_df.to_csv(f'{output_dir}{dataset}/clustering_comparison_scores.csv')
    #scores_df.to_latex(f'{output_dir}{dataset}/clustering_comparison_scores.
↪tex')

    # Display scores DataFrame
    display(scores_df)
```

[182]:
```
for dataset in DATASET_NAMES:
    #display('------------------------------')
    display(f'{dataset} - Clustering Comparison between CellTypist and␣
↪Antibody')

    # Assuming celltypist_df and antibody_df are defined elsewhere and␣
↪available here
    compute_clustering_scores(celltypist_df, antibody_df, DIR, dataset)
```

```
'PBMC1 - Clustering Comparison between CellTypist and Antibody'
```

	NMI	ARI	Homogeneity	Completeness	Fowlkes_Mallows
0	0.742338	0.713159	0.709696	0.778128	0.764437

```
'PBMC2 - Clustering Comparison between CellTypist and Antibody'
```

	NMI	ARI	Homogeneity	Completeness	Fowlkes_Mallows
0	0.660455	0.48564	0.673828	0.647602	0.582832

```
'PBMC3 - Clustering Comparison between CellTypist and Antibody'
```

	NMI	ARI	Homogeneity	Completeness	Fowlkes_Mallows
0	0.665993	0.513896	0.698774	0.636149	0.57817

```
'PBMC4 - Clustering Comparison between CellTypist and Antibody'
```

	NMI	ARI	Homogeneity	Completeness	Fowlkes_Mallows
0	0.690424	0.505103	0.758068	0.633863	0.582394

[1]:
```
!export PATH=/Library/TeX/texbin:$PATH
```