```
# Packages
import csv
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.cluster import KMeans


import warnings
warnings.filterwarnings('ignore')


#Load dataset
from google.colab import drive
drive.mount('/content/drive')
train_data = pd.read_csv("/content/drive/MyDrive/Post Undergrad/CS 5262: Machine Learning/train_data.csv")
df = pd.DataFrame(train_data)
test_data = pd.read_csv("/content/drive/MyDrive/Post Undergrad/CS 5262: Machine Learning/test_data.csv")
```

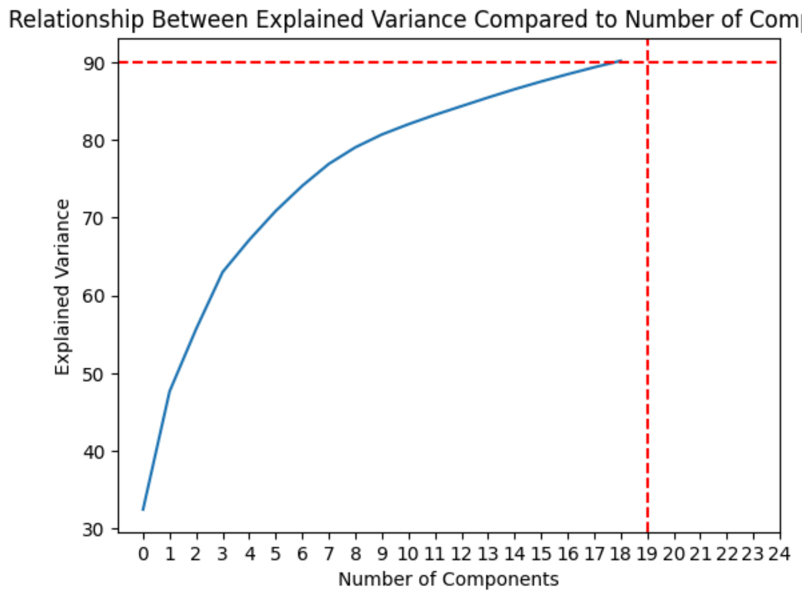
```
Drive already mounted at /content/drive; to attempt to forcibly remount, call drive.mount("/content/drive", force_remount=True).
```

```
# Edit Data file train
dfx = df.copy()
X = dfx.drop("disease_status",axis=1)
X = X.drop("File_ID",axis=1)
X = X.drop("tSNE_1",axis=1)
X = X.drop("tSNE_2",axis=1)
y = df.iloc[:,44:]
X =  pd.DataFrame(X)
y = pd.DataFrame(y)
# Edit Data file test
test_datax = test_data.copy()
Xtest = dfx.drop("disease_status",axis=1)
Xtest = Xtest.drop("File_ID",axis=1)
Xtest = Xtest.drop("tSNE_1",axis=1)
Xtest = Xtest.drop("tSNE_2",axis=1)
ytest = test_data.iloc[:,44:]
print(ytest)
```

```
       File_ID  disease_status
0           31               0
1           31               0
2           31               0
3           31               0
4           31               0
...        ...             ...
46723       27               1
46724       27               1
46725       26               1
46726       16               1
46727       17               1

[46728 rows x 2 columns]
```

```
# Complete PCA
from sklearn.decomposition import PCA
pca = PCA(n_components = .90)
X_train_pca = pca.fit_transform(X)
X_test_pca = pca.transform(test_data.iloc[:,:42])
X_train_pca = pd.DataFrame(X_train_pca)
X_test_pca = pd.DataFrame(X_test_pca)
X_train_pca.columns = ["PC1","PC2","PCA3","PCA4","PCA5","PCA6","PCA7","PCA8","PCA9","PCA10","PCA11","PCA12","PCA13","PCA14","PCA15","PCA16",
test = X_test_pca.join(test_data.iloc[:,44:])


# Display how we chose number of PC
explained_variance = pca.explained_variance_ratio_
explained_variance = pca.explained_variance_ratio_
plt.plot(np.cumsum(explained_variance)*100)
plt.axhline(y = 90, color = 'r', linestyle = '--')
plt.axvline(x = 19, color = 'r', linestyle = '--')
plt.xticks(range(0,25))
plt.xlabel('Number of Components')
plt.ylabel("Explained Variance")
plt.title("Relationship Between Explained Variance Compared to Number of Components")
```

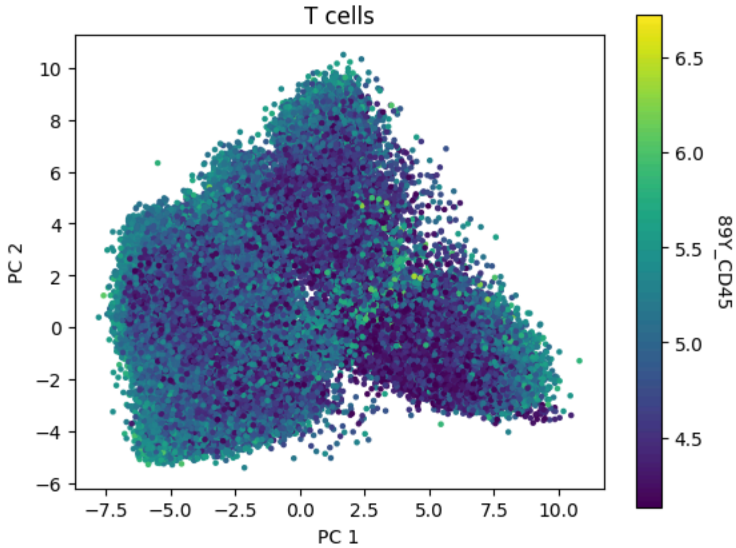
```
Text(0.5, 1.0, 'Relationship Between Explained Variance Compared to Number of Components')
```

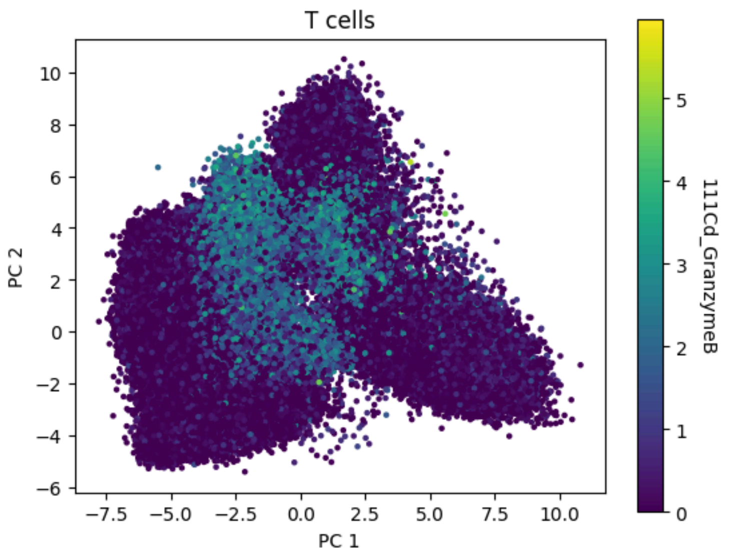


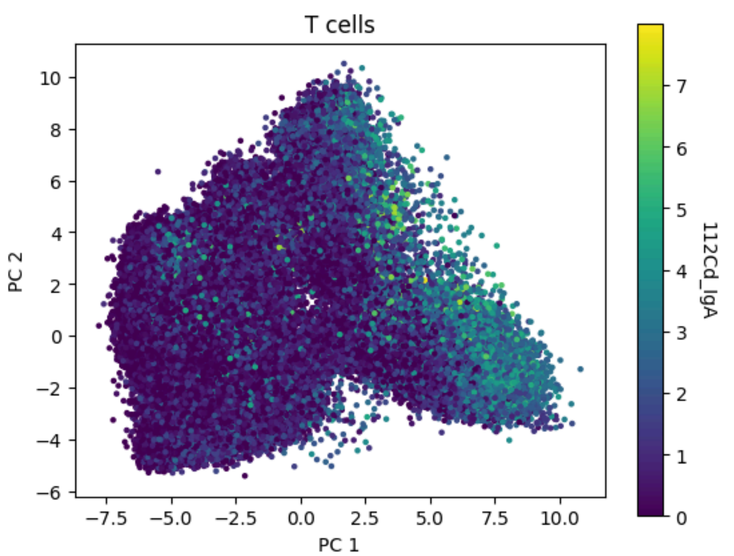

```
print(X_test_pca)
```

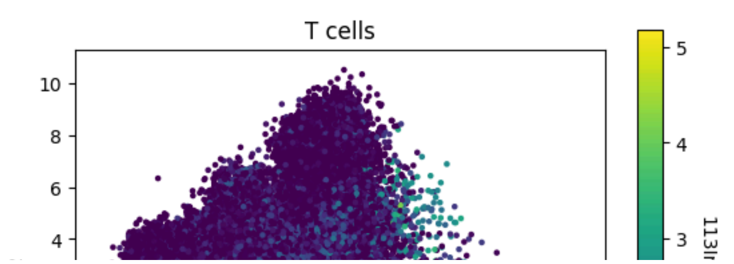
```
             0         1         2         3         4         5         6  \
0      2.065260  2.837408 -3.046193 -1.421741  0.189240 -1.898389  1.847129
1     -3.300881  1.209013  0.190181  0.706488  0.766812  1.604509 -0.497695
2     -5.592737 -1.456376  2.469224 -1.712299  0.604456 -1.147351  1.143878
3     -2.795340 -1.756877 -2.538769  1.319350  1.557766  3.924942 -3.256800
4     -4.501865 -4.112488  1.097678  1.273828 -1.522866  2.227537  1.320202
...         ...       ...       ...       ...       ...       ...       ...
46723 -3.923201 -4.490558  2.107046  0.361813 -2.443016  4.145321  1.092002
46724  6.504190 -1.620971  0.425800 -1.590910  0.029881  1.404748  0.304200
46725  1.731255  4.385363  1.066837 -2.250015 -2.062843  1.233641 -0.861078
46726  6.104070 -2.361998  1.100090 -0.555060  1.238457  1.332134  0.398946
46727 -2.323189 -2.764531 -2.336211  1.213100 -2.914122 -0.012792  1.446788

             7         8         9        10        11        12        13  \
0      1.507118 -2.436506  0.623992 -0.112060  0.303669 -0.180848  0.142228
1     -0.027334 -1.560305  1.682948  1.635736 -1.979390 -0.676116  2.053563
2      0.318788  0.296627 -0.553186  0.094893 -0.633350 -0.112551 -0.406903
3     -0.606798 -1.368937  0.029845 -0.433859 -0.843123  1.876561 -0.464940
4      0.037277 -1.718266  0.839413  0.050322 -1.092558  0.505751 -0.245638
...         ...       ...       ...       ...       ...       ...       ...
46723  5.723486  2.721470  2.075391  0.674281  0.453538  0.656315 -1.373106
46724 -2.456482 -0.341142  0.520839  0.303330 -0.079992 -1.203117 -1.303837
46725  5.163595  0.869351  2.376043  2.943963 -2.474540 -0.648952  0.368044
46726 -2.280628 -1.053748 -0.893973 -0.755296  0.542654 -0.527036  0.013577
46727 -2.354774  0.046579 -0.484763  0.128513  0.263087  0.804200 -0.166092

            14        15        16        17        18
0      0.911206 -0.052895 -0.465407 -0.401209  0.873088
1      0.623632  2.049746 -2.081984  1.560676  1.172755
2     -0.052396  0.107912 -0.013970 -0.180314 -0.273145
3      2.102752  0.375789  0.893243  0.399323 -0.699122
4     -0.946035  0.067604  0.109420  0.110448  0.180827
...         ...       ...       ...       ...       ...
46723  0.405445  0.131920 -0.965557  0.143121  0.146782
46724 -0.902980 -0.507832 -1.495366  0.617554 -0.246559
46725  0.833838  0.632201  0.700542 -0.460588 -1.100424
46726 -0.443016 -1.019344 -1.301008  0.541734 -0.666112
46727 -0.459190 -0.035664  0.615323  0.330930 -1.076969

[46728 rows x 19 columns]
```

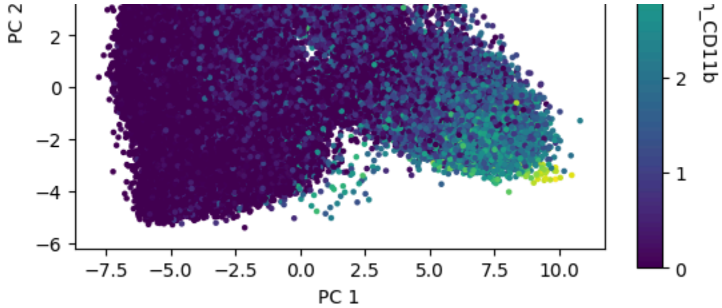
```
for i, marker in enumerate(train_data.columns):
    fig, ax = plt.subplots()
    plt.scatter(X_train_pca.PC1, X_train_pca.PC2, c=train_data.iloc[:,i], s=5)
    plt.xlabel('PC 1')
    plt.ylabel('PC 2')
    plt.title("T cells")
    ax.set_aspect('equal')
    cbar = plt.colorbar()
    cbar.ax.set_ylabel(marker, rotation=270,labelpad=15)

    plt.show()
```

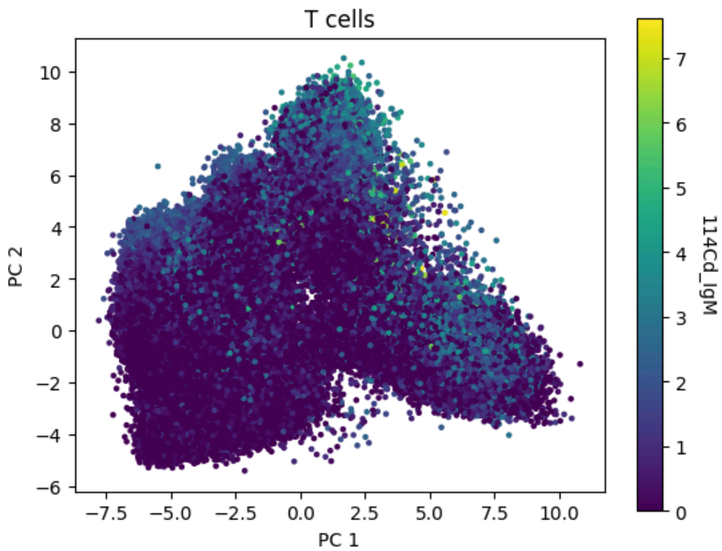

T cells
PC 2
PC 1
89Y_CD45

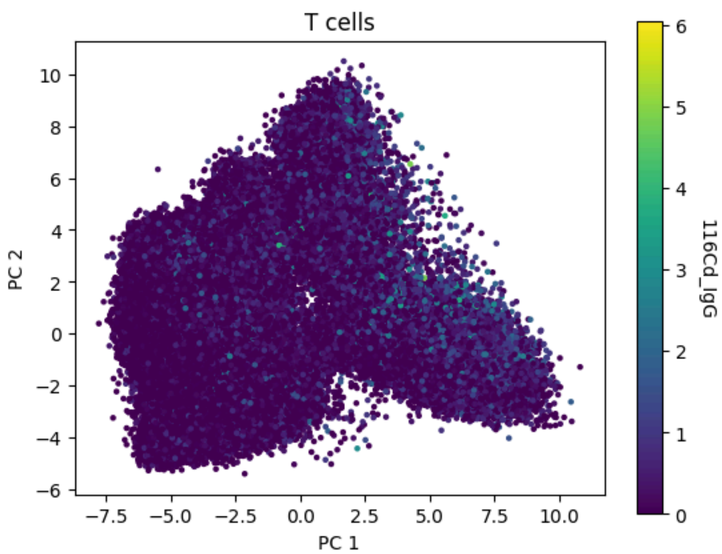

T cells
PC 2
PC 1
111Cd_GranzymeB

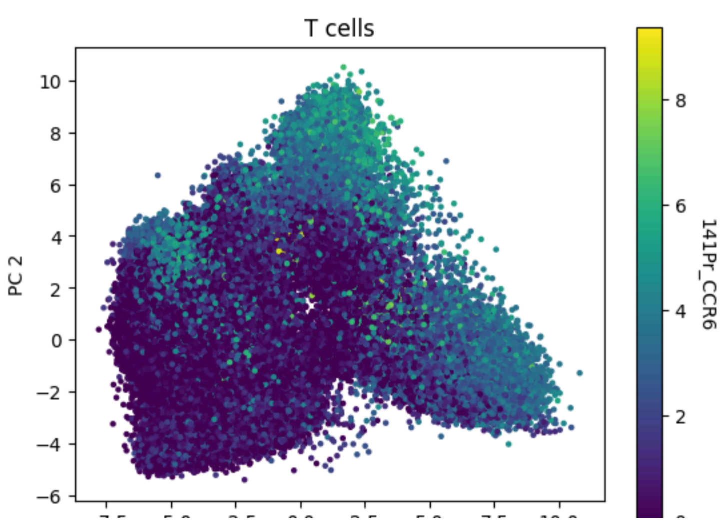

T cells
PC 2
PC 1
112Cd_IgA

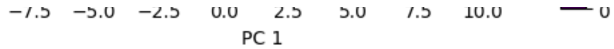

T cells

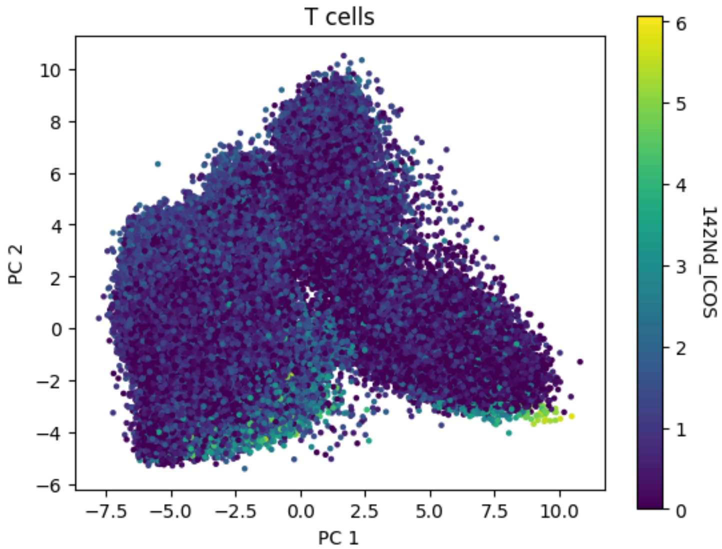

PC 2
PC 1
_CD11b

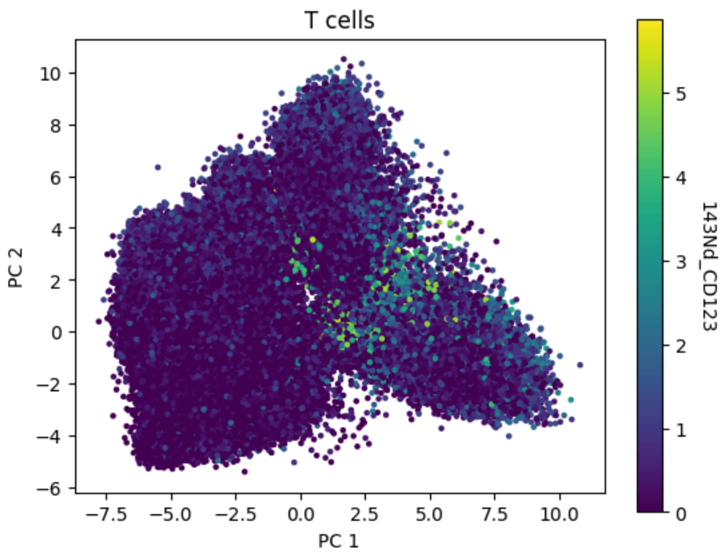

T cells
PC 2
PC 1
114Cd_IgM

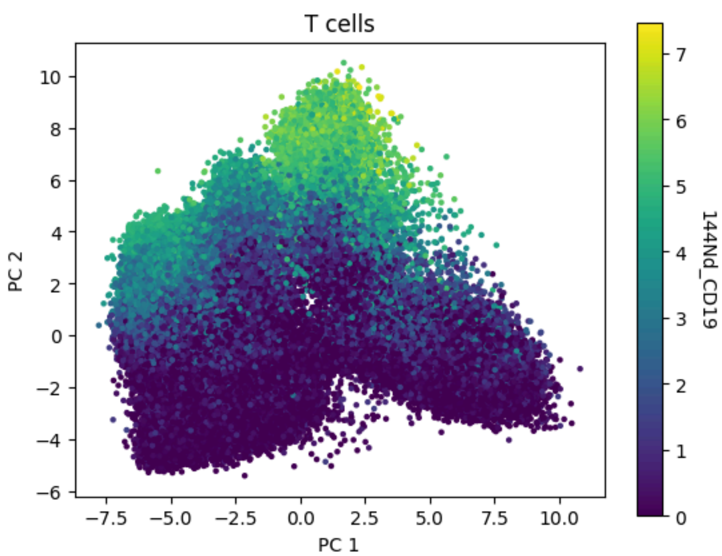

T cells
PC 2
PC 1
116Cd_IgG

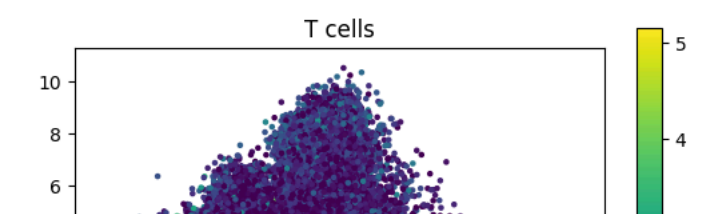

T cells
PC 2
141Pr_CCR6

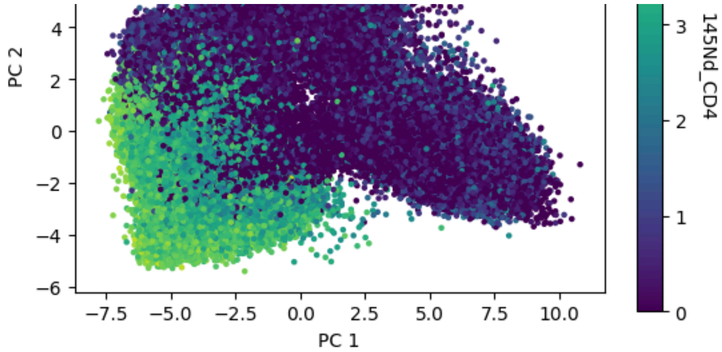

−7.5 −5.0 −2.5 0.0 2.5 5.0 7.5 10.0
PC 1
0

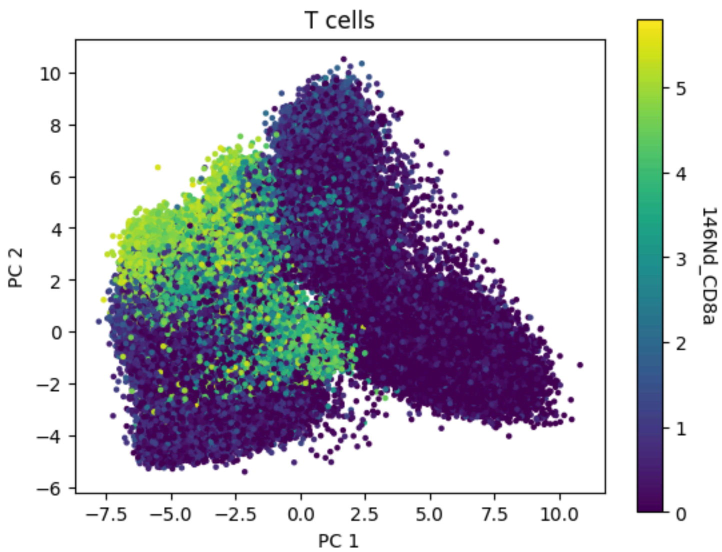

T cells
PC 2
PC 1
142Nd_ICOS

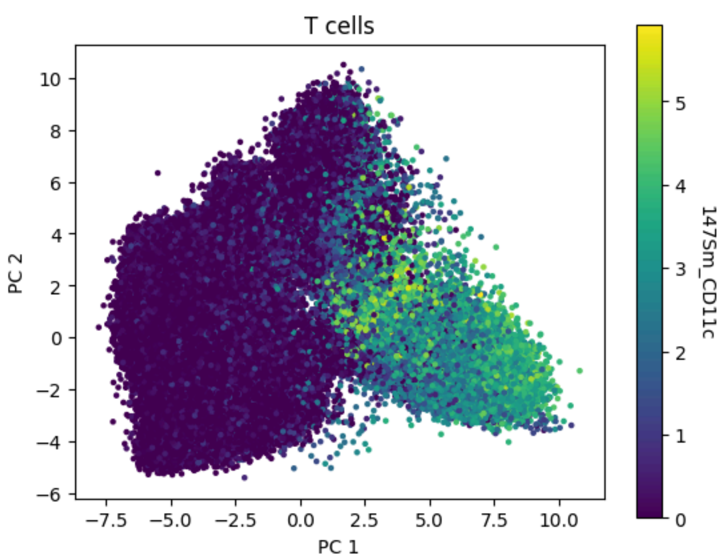

T cells
PC 2
PC 1
143Nd_CD123

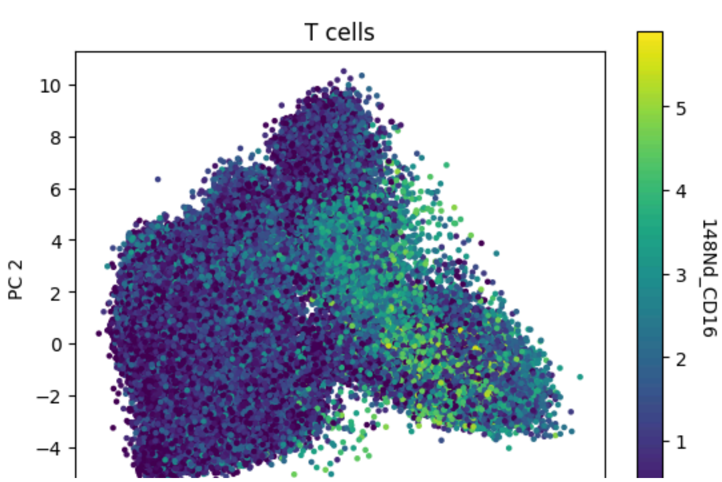

T cells
PC 2
PC 1
144Nd_CD19

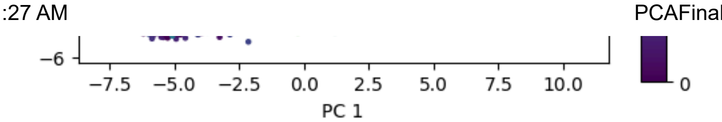

T cells

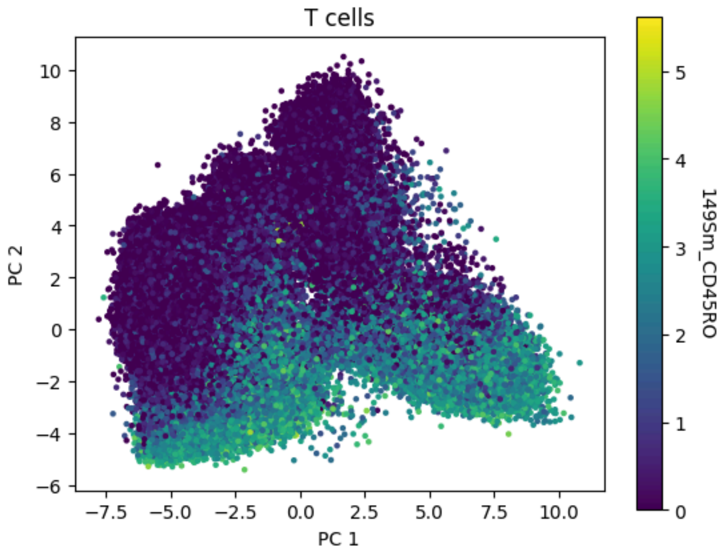

PC 2
PC 1
145Nd_CD4

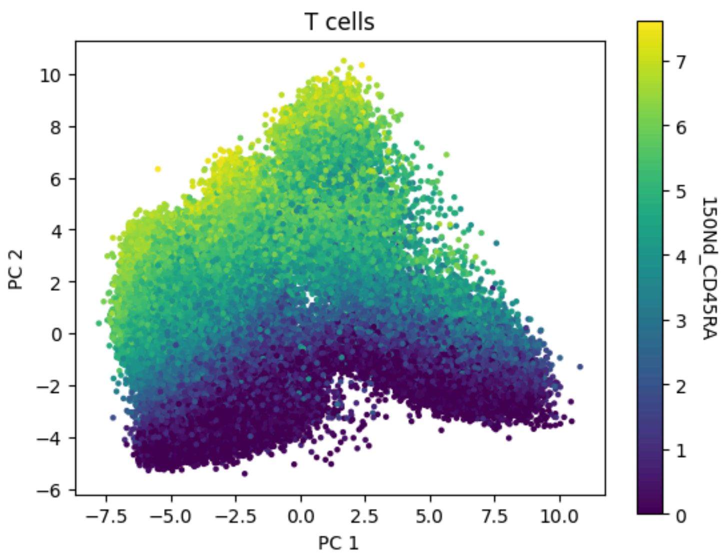

T cells
PC 2
PC 1
146Nd_CD8a

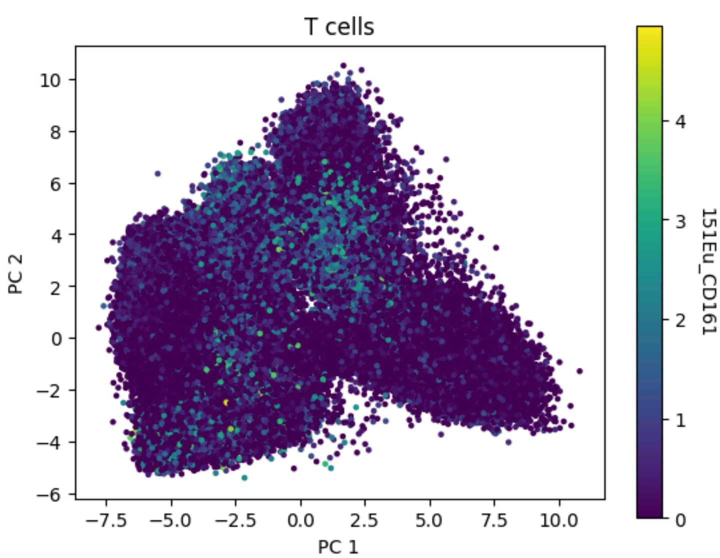

T cells
PC 2
PC 1
147Sm_CD11c

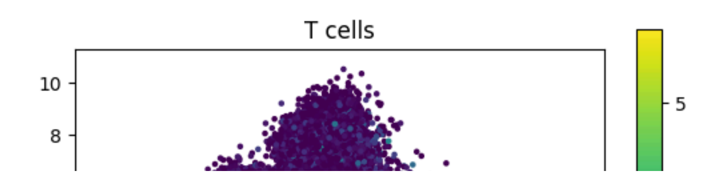

T cells
PC 2
148Nd_CD16

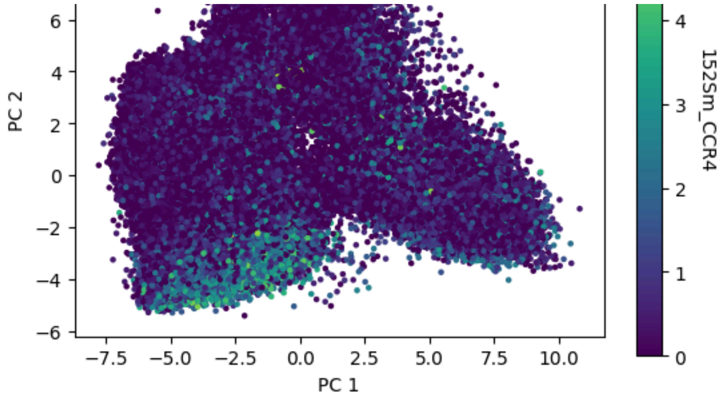

PC 1

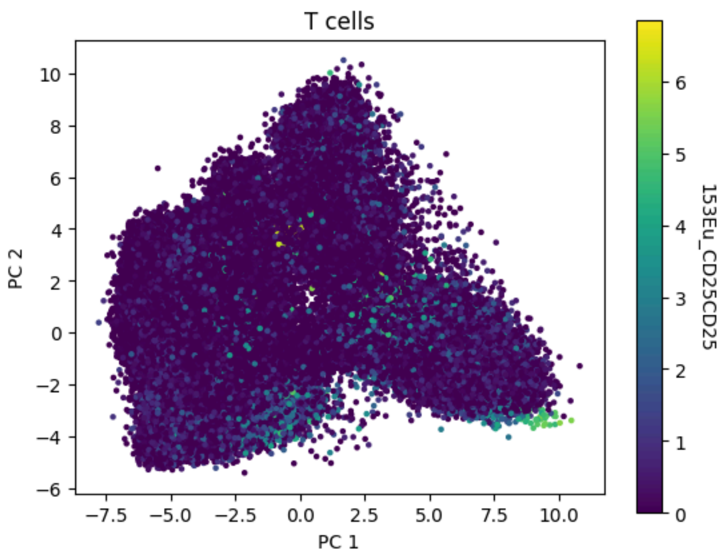

T cells
PC 2
PC 1
149Sm_CD45RO

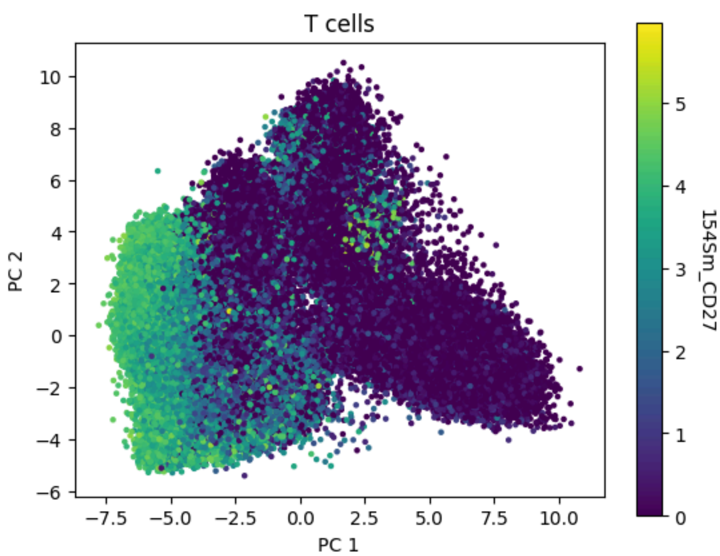

T cells
PC 2
PC 1
150Nd_CD45RA

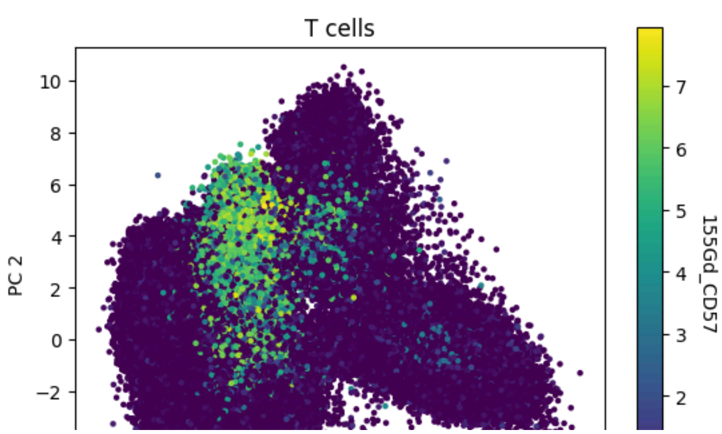

T cells
PC 2
PC 1
151Eu_CD161

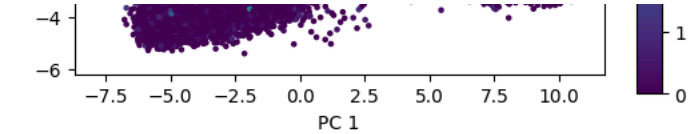

T cells

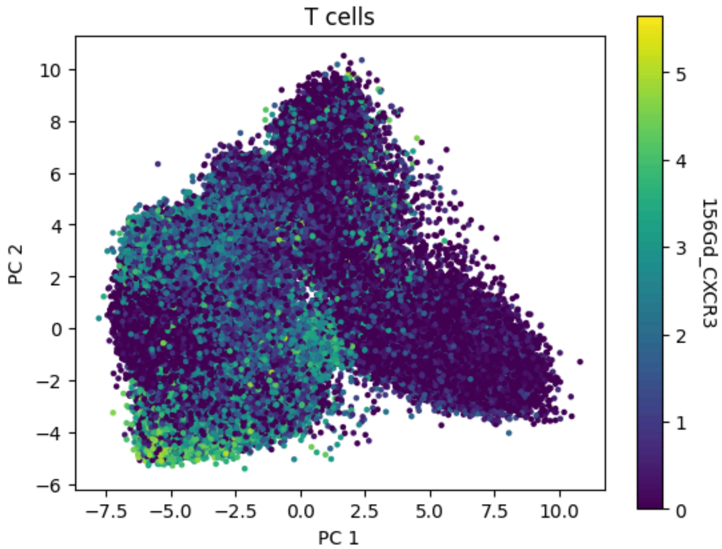

PC 2
PC 1
152Sm_CCR4

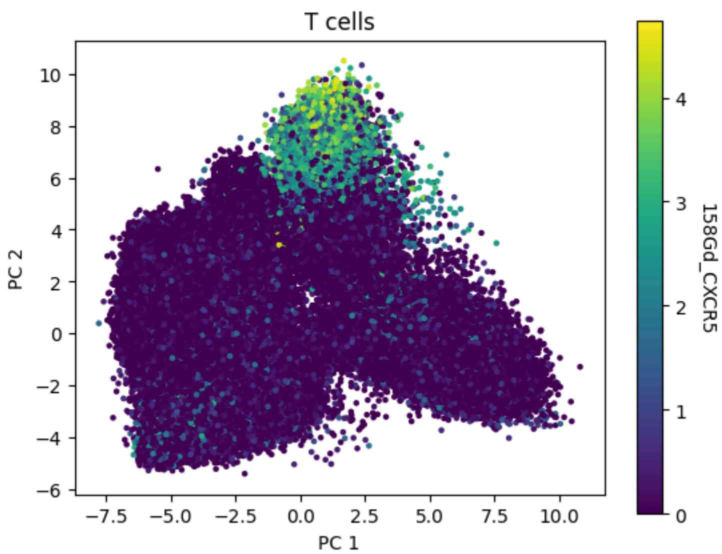

T cells
PC 2
PC 1
153Eu_CD25CD25

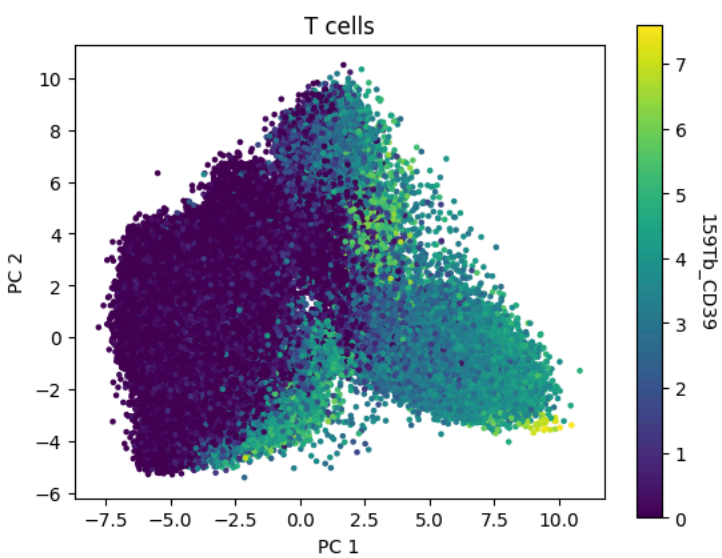

T cells
PC 2
PC 1
154Sm_CD27

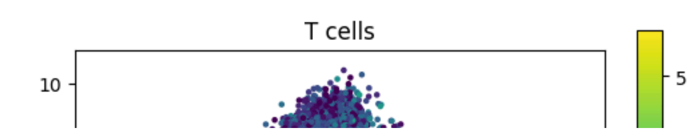

T cells
PC 2
155Gd_CD57


−4
−6
−7.5 −5.0 −2.5 0.0 2.5 5.0 7.5 10.0
PC 1
1
0


T cells
10
8
6
4
2
0
−2
−4
−6
PC 2
−7.5 −5.0 −2.5 0.0 2.5 5.0 7.5 10.0
PC 1
5
4
3
2
1
0
156Gd_CXCR3


T cells
10
8
6
4
2
0
−2
−4
−6
PC 2
−7.5 −5.0 −2.5 0.0 2.5 5.0 7.5 10.0
PC 1
4
3
2
1
0
158Gd_CXCR5


T cells
10
8
6
4
2
0
−2
−4
−6
PC 2
−7.5 −5.0 −2.5 0.0 2.5 5.0 7.5 10.0
PC 1
7
6
5
4
3
2
1
0
159Tb_CD39


T cells
10
5