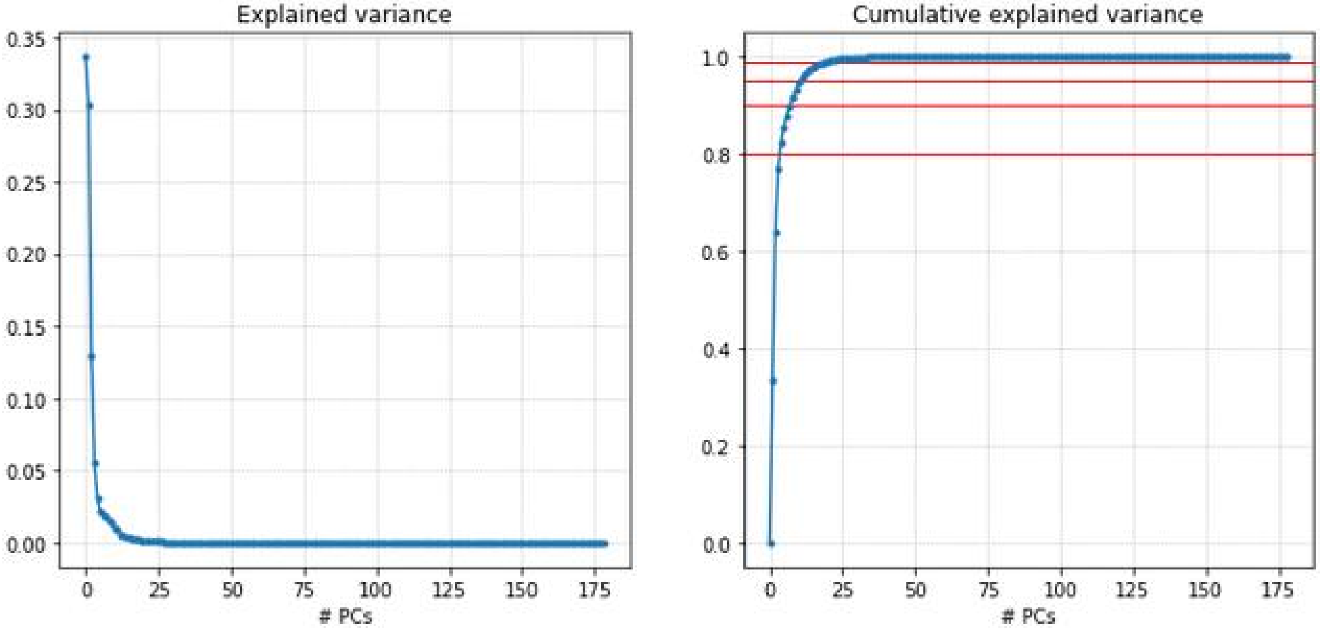
Human Microbiome Project

Databases sizes

Database	O level	F level	G level
AGP	(9511, 168)	(9511, 258)	(9511, 535)
IBD	(86,32)	(86,64)	(86, 107)
PTB	(3457, 102)	(3457, 158)	(3457, 291)
T2D	(1044, 63)	(1044, 115)	(1044, 221)
Summary	(14098, 179)	(14098, 267)	(14098, 574)

First number is the number of items (‘patients’) in each database, the second number is the number of different taxonomy types (e.g. g__Acetobacter). Taxonomies f__[Weeksellaceae] and Weeksellaceae are considered as the same ones

CEV summary for o level taxonomy

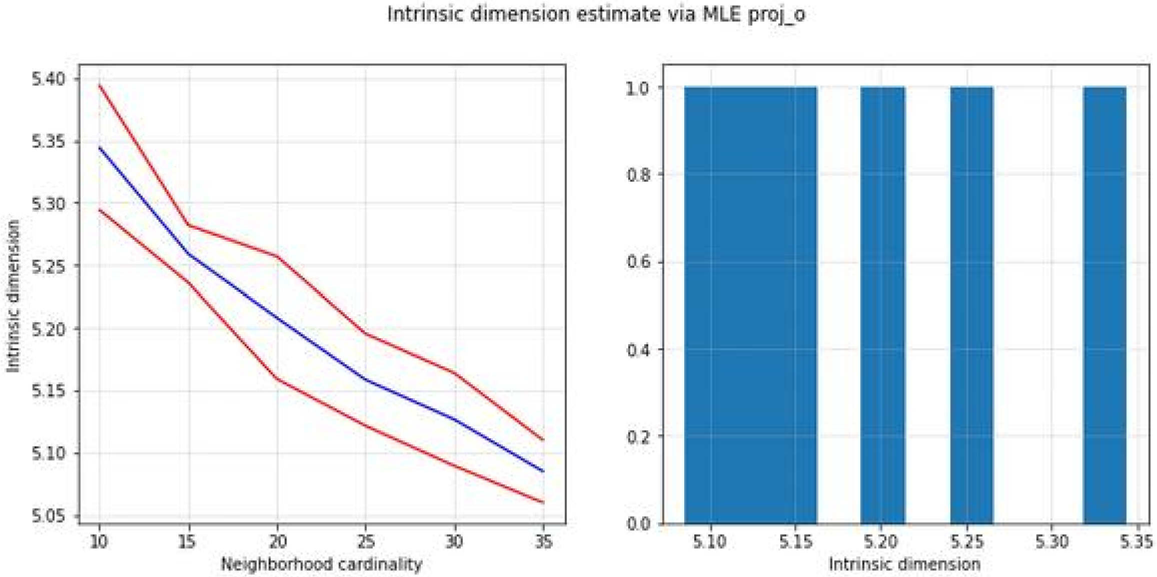


Horizontal lines are showing
0.8, 0.9, 0.95 and 0.99 part of CEV

Databases sizes after PCA (0.99 CEV)

Database	O level	F level	G level
Summary before	(14098; 179)	(14098; 267)	(14098; 574)
Summary after	(14098; 20)	(14098; 42)	(14098; 68)

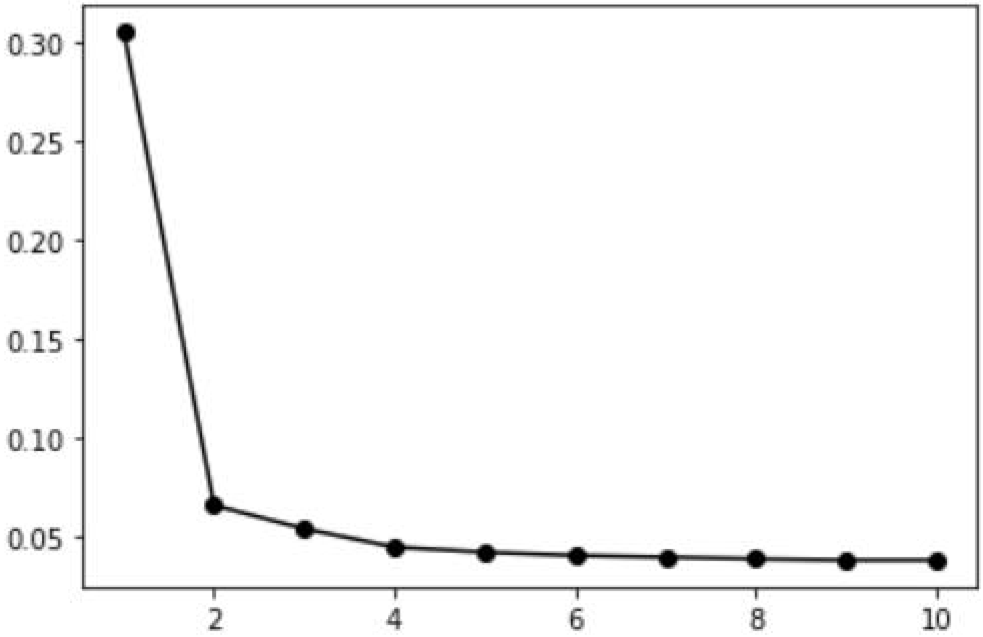
MLE estimation for O taxonomic level



Databases after ISOMAP

Database	O level	F level	G level
PCA	(14098; 20)	(14098; 42)	(14098; 68)
ISOMAP	(14098, 5)	(14098, 7)	(14098, 8)

ISOMAP reconstruction to original space through the K-NN regression to the Principal Components Space

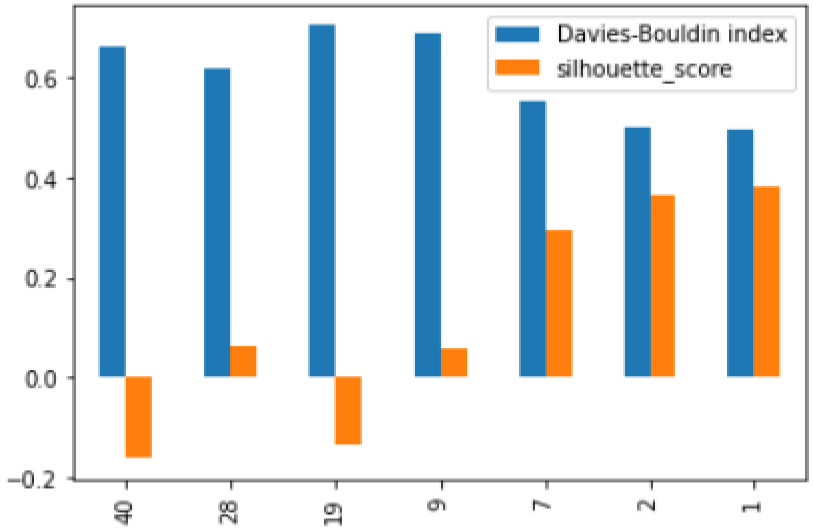


Here we see stabilization after the dimension 5. So we could say 5 is a suitable dimension for a low dimensional representation

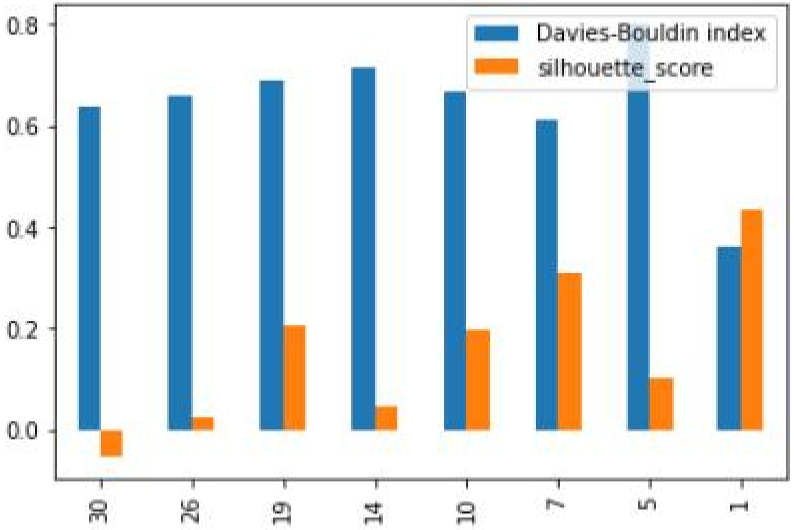
X axis is the dimension of the nonlinear reduction manifold
Y axis is the
Mean Absolute Error
MAE = mean[|q-q̂|_2 / |q|_2]
chosen to be invariant under the translations.
Here q is the true PCA-projected vector and q̂ is the predicted by the K-NN regressor vector
(hyperparameters chosen from the grid search cross validation method)

DBScan

X axis is for number of clusters, Y axis is for the scores amount.
Lower DB index and higher silhouette score are preferable



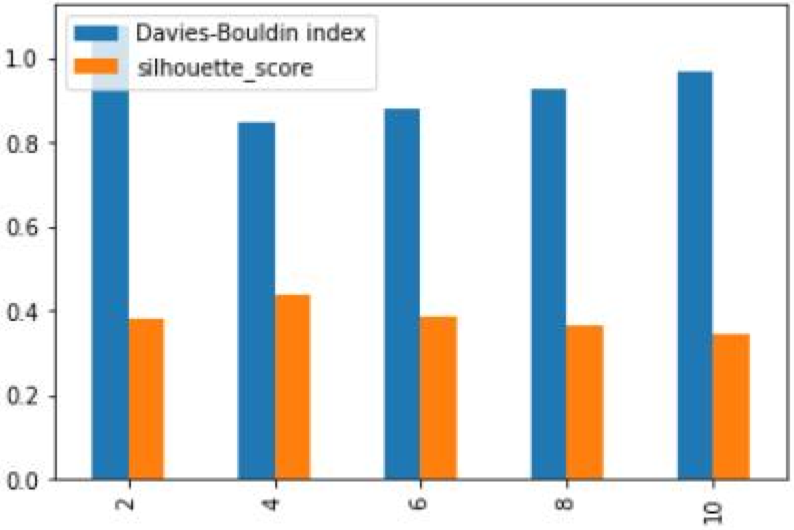
ISOMAP



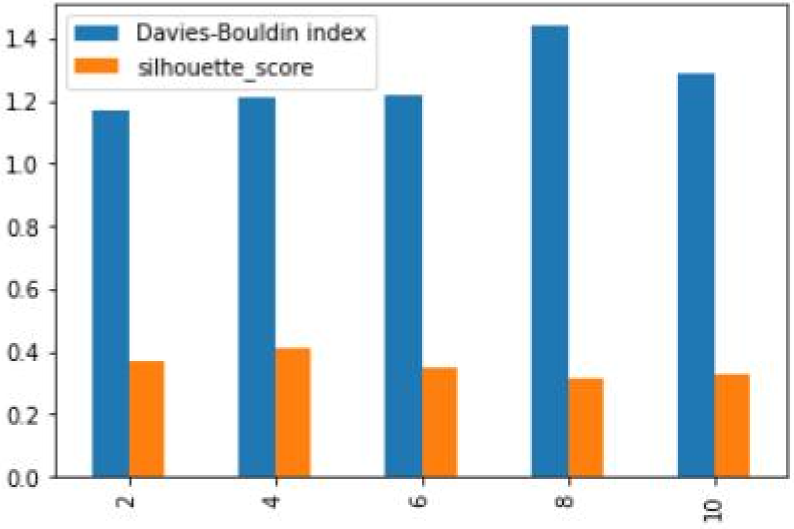
PCA

Kmeans: DBI vs silhouette

X axis is for number of clusters, Y axis is for the scores amount.
Lower DB index and higher silhouette score are preferable



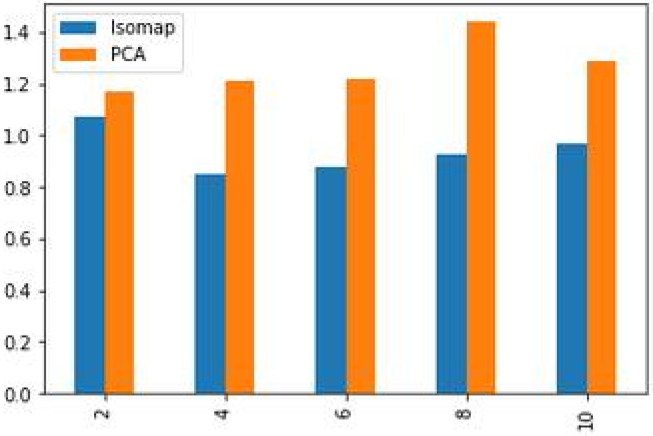
ISOMAP



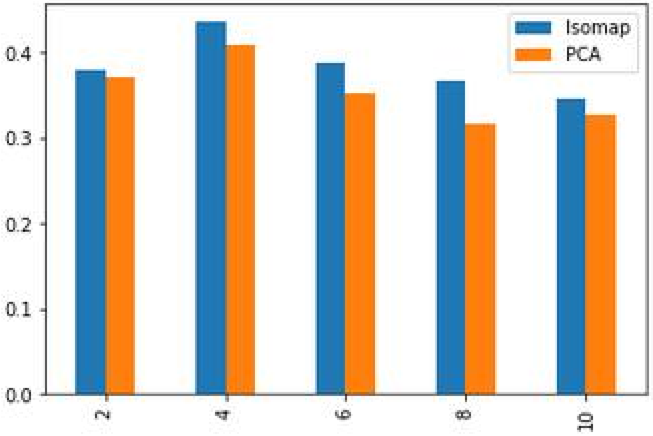
PCA

Kmeans: PCA vs Isomap

X axis is for number of clusters, Y axis is for the scores amount.
Lower DB index and higher silhouette score are preferable



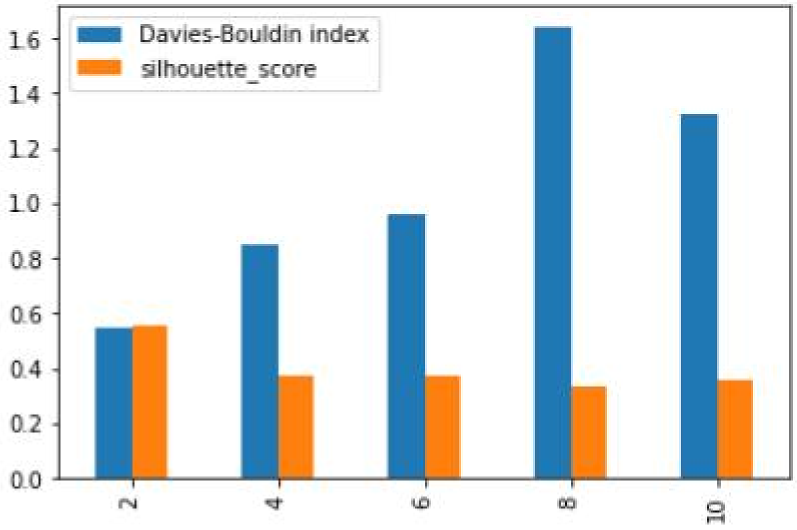
DBI



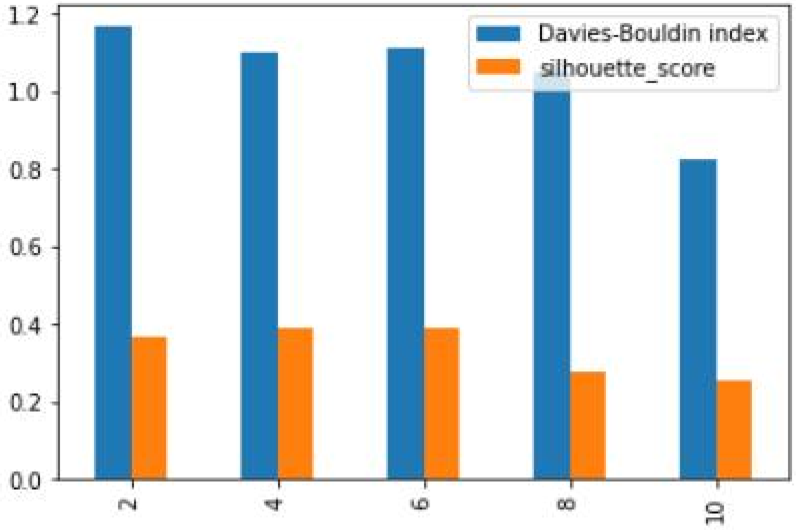
silhouette

Spectral Clustering: DBI vs silhouette

X axis is for number of clusters, Y axis is for the scores amount.
Lower DB index and higher silhouette score are preferable



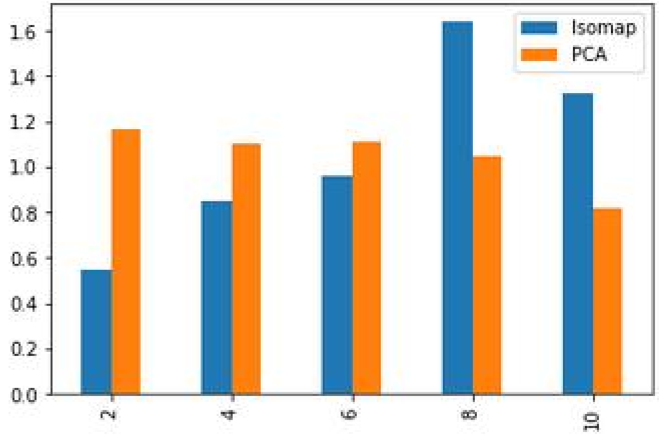
ISOMAP



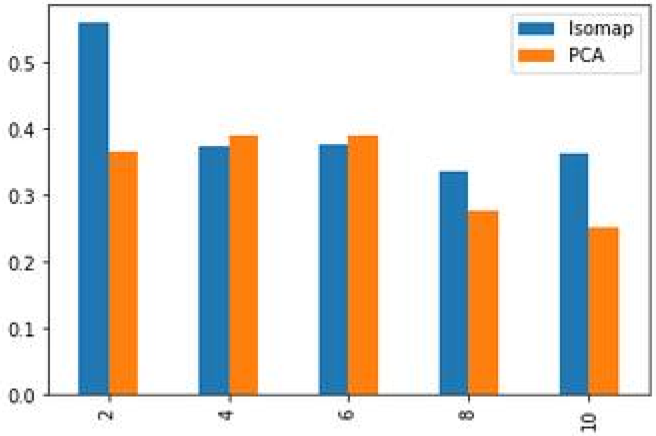
PCA

Spectral Clustering: DBI vs silhouette

X axis is for number of clusters, Y axis is for the scores amount.
Lower DB index and higher silhouette score are preferable

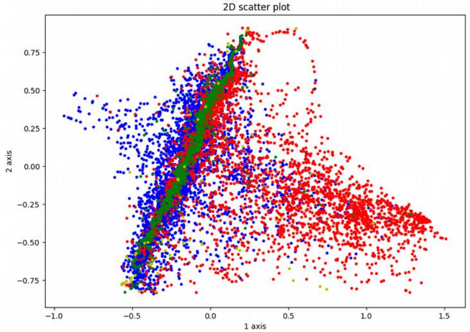


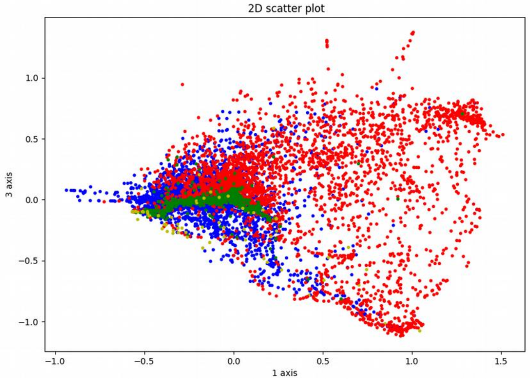
DBI

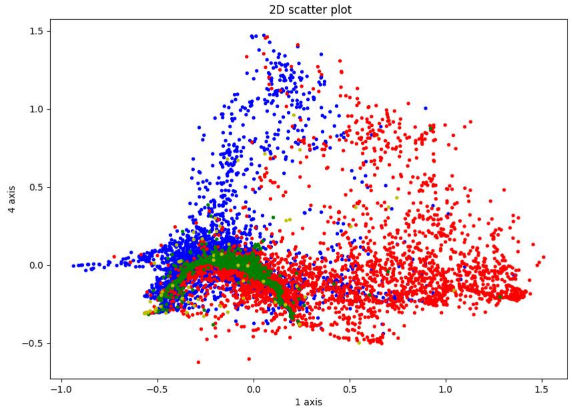


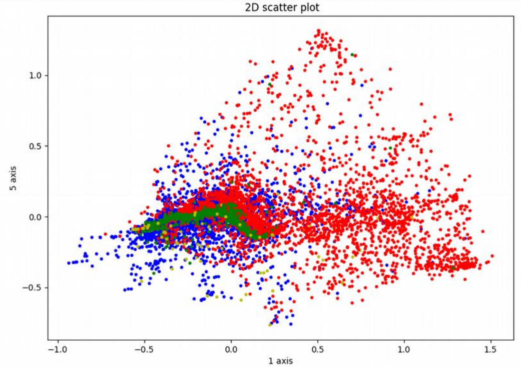
silhouette

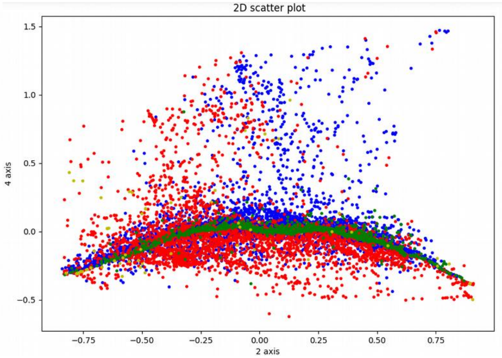
2d projections isomap

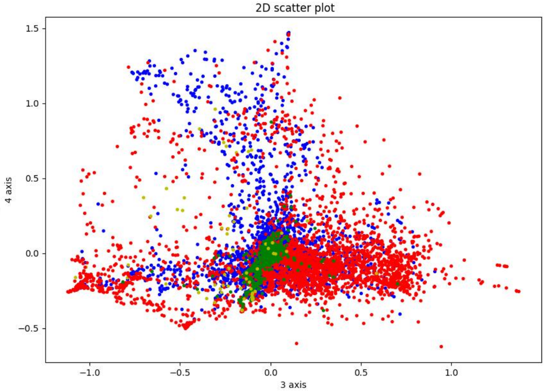


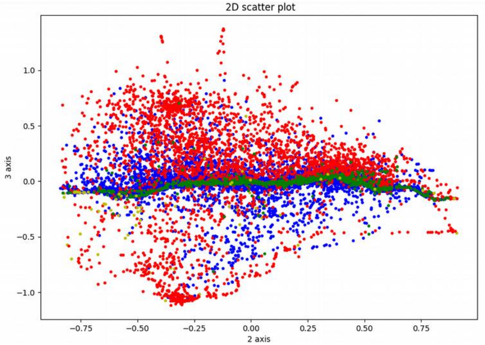


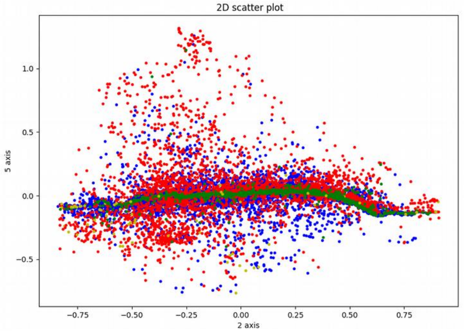


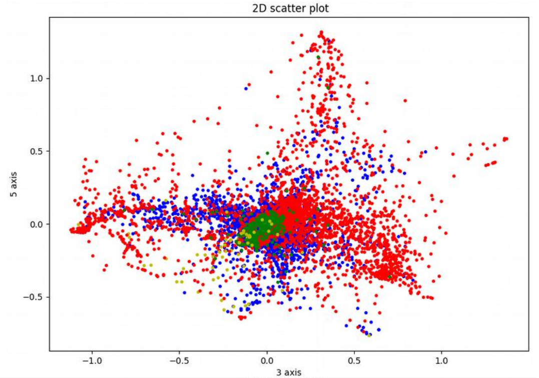


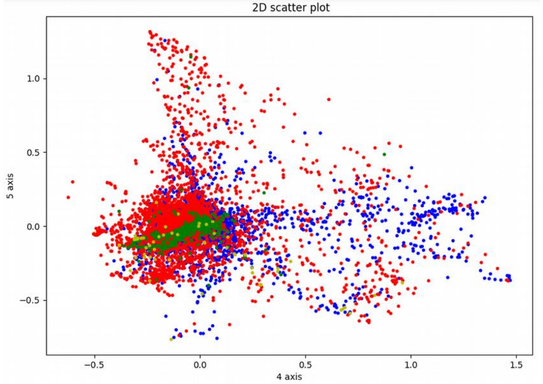






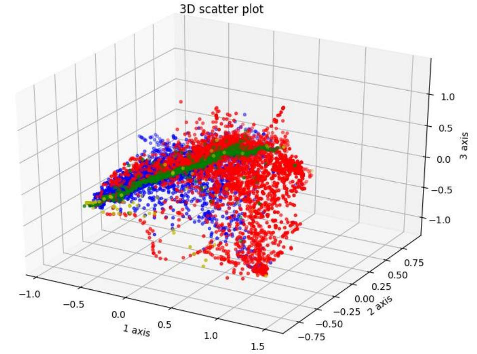


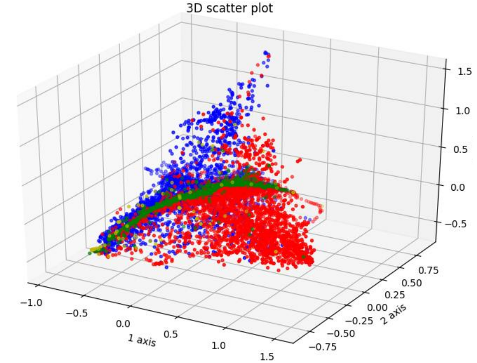


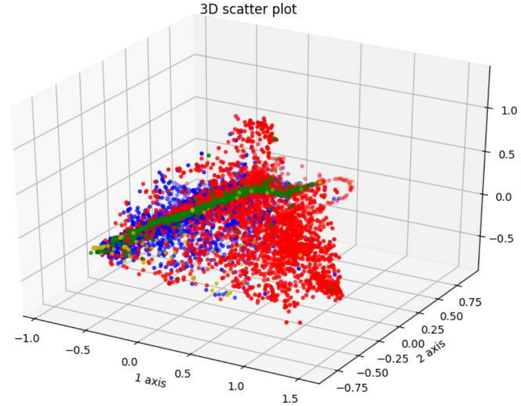


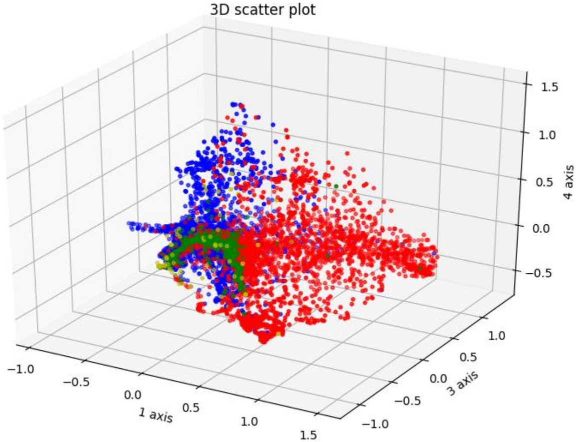
AGP (9511) – blue
PTB (3457) – red
T2D (1044) – green
IBD (86) – yellow

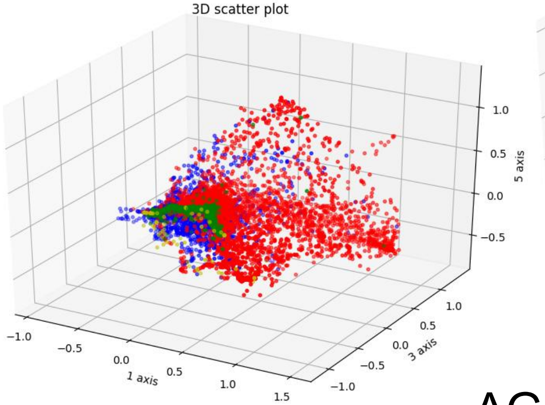
3d projections isomap

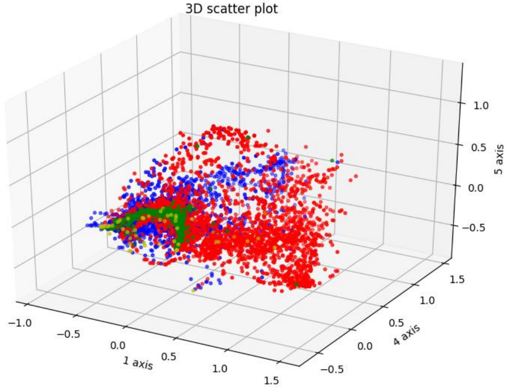


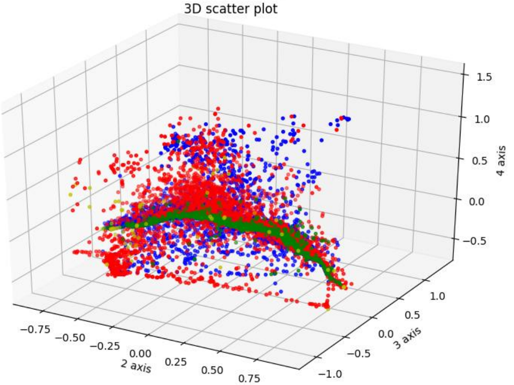


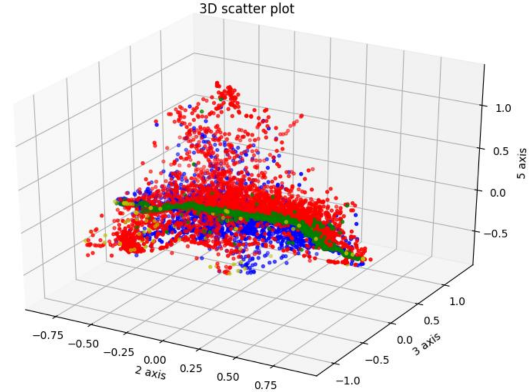


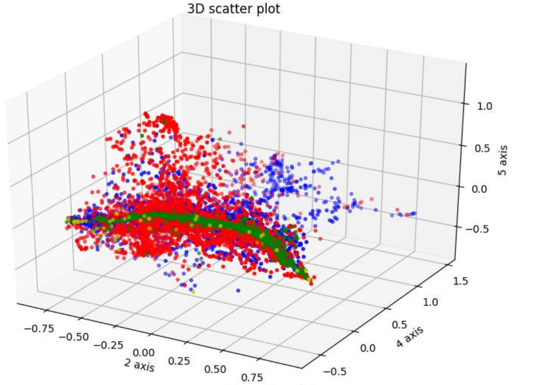


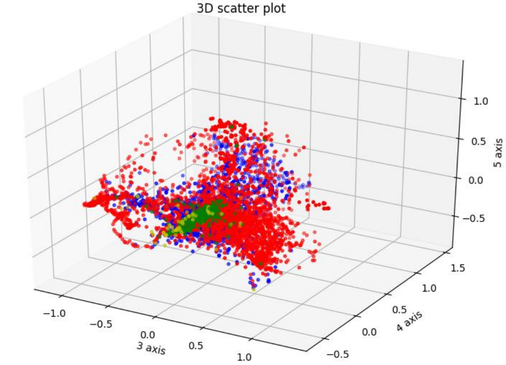






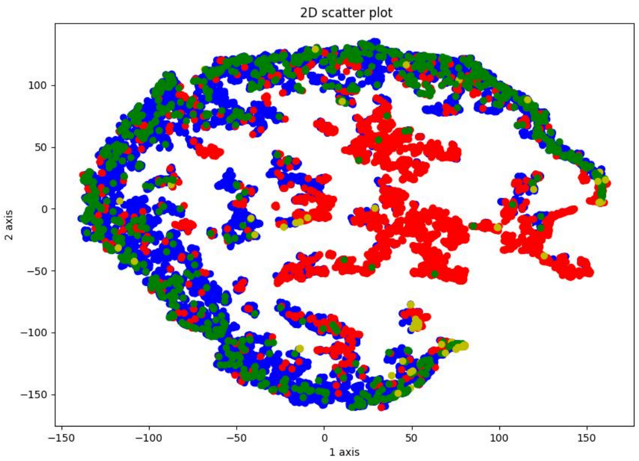




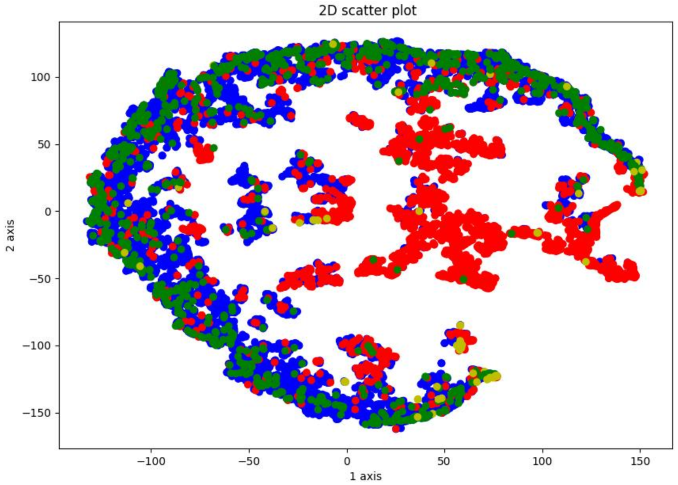


AGP (9511) – blue
PTB (3457) – red
T2D (1044) – green
IBD (86) – yellow

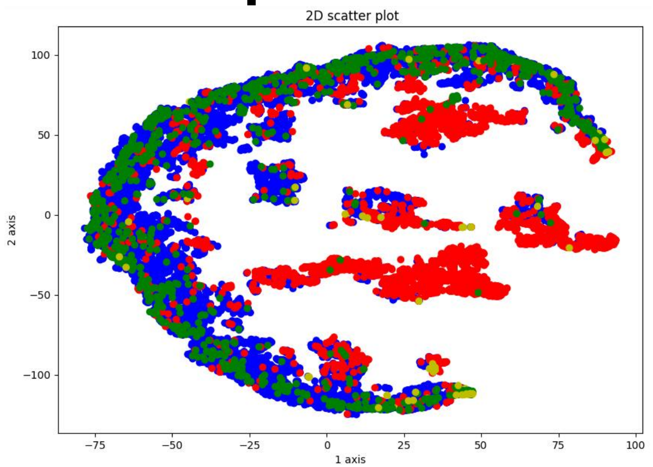
t-SNE visualization 2D isomap



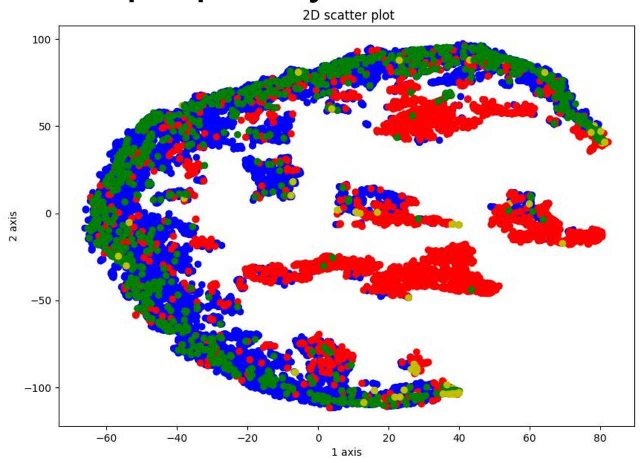
perplexity = 40



perplexity = 50

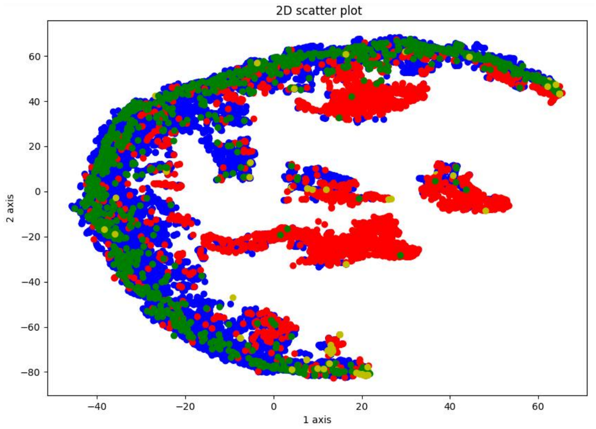


perplexity = 100



perplexity = 120

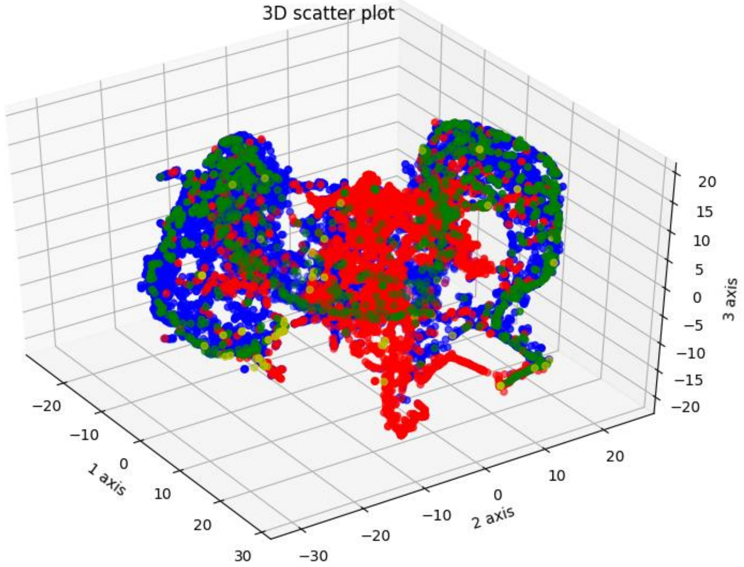
AGP (9511) – blue
PTB (3457) – red
T2D (1044) – green
IBD (86) – yellow



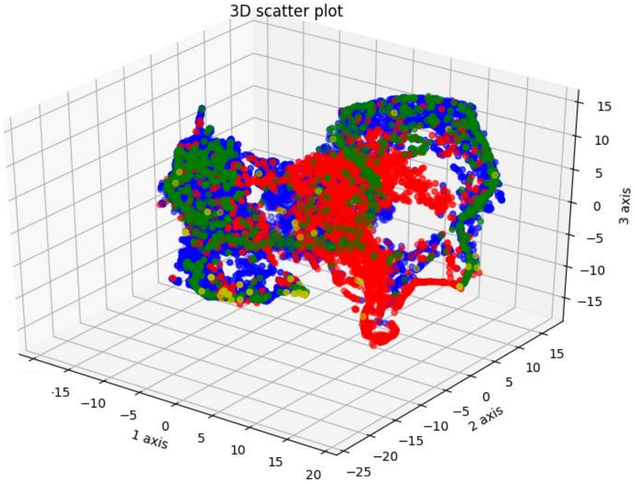
perplexity = 200

t-SNE visualization 3D isomap

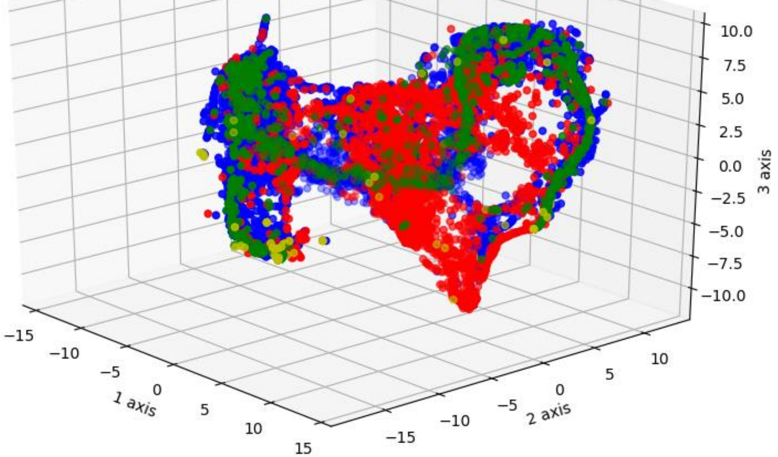
AGP (9511) – blue
PTB (3457) – red
T2D (1044) – green
IBD (86) – yellow



perplexity = 50

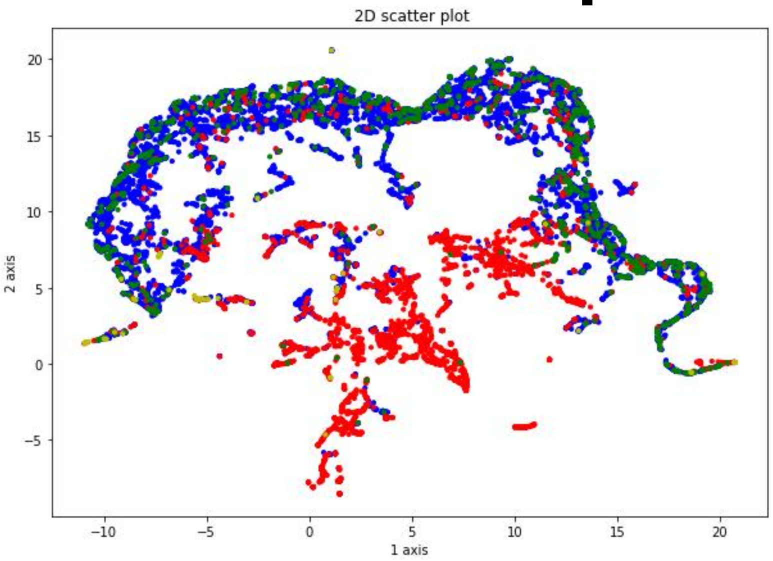


perplexity = 120



perplexity = 300

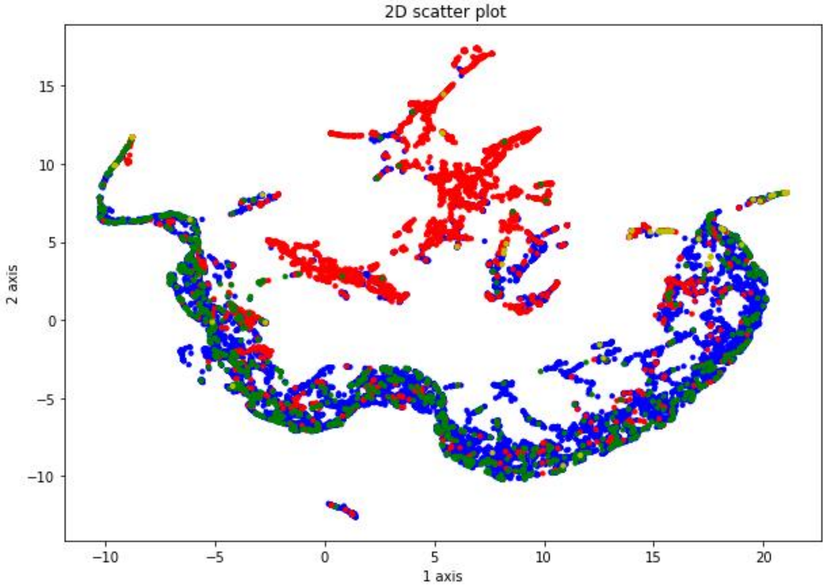
Umap visualization 2D Isomap

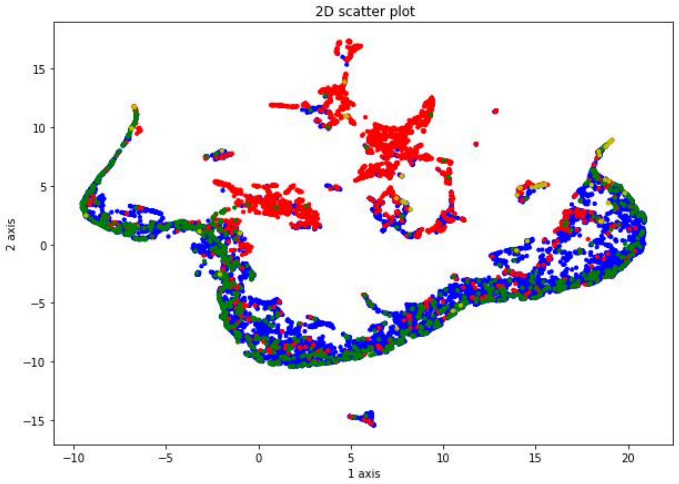


n_neighbors=10

AGP (9511) – blue
PTB (3457) – red
T2D (1044) – green
IBD (86) – yellow

n_neighbors=15

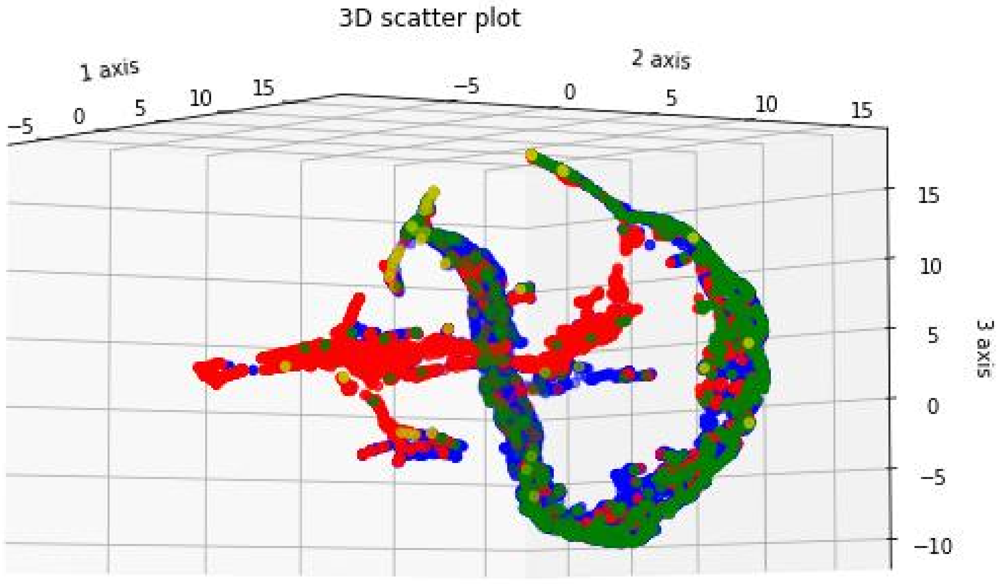




n_neighbors=20

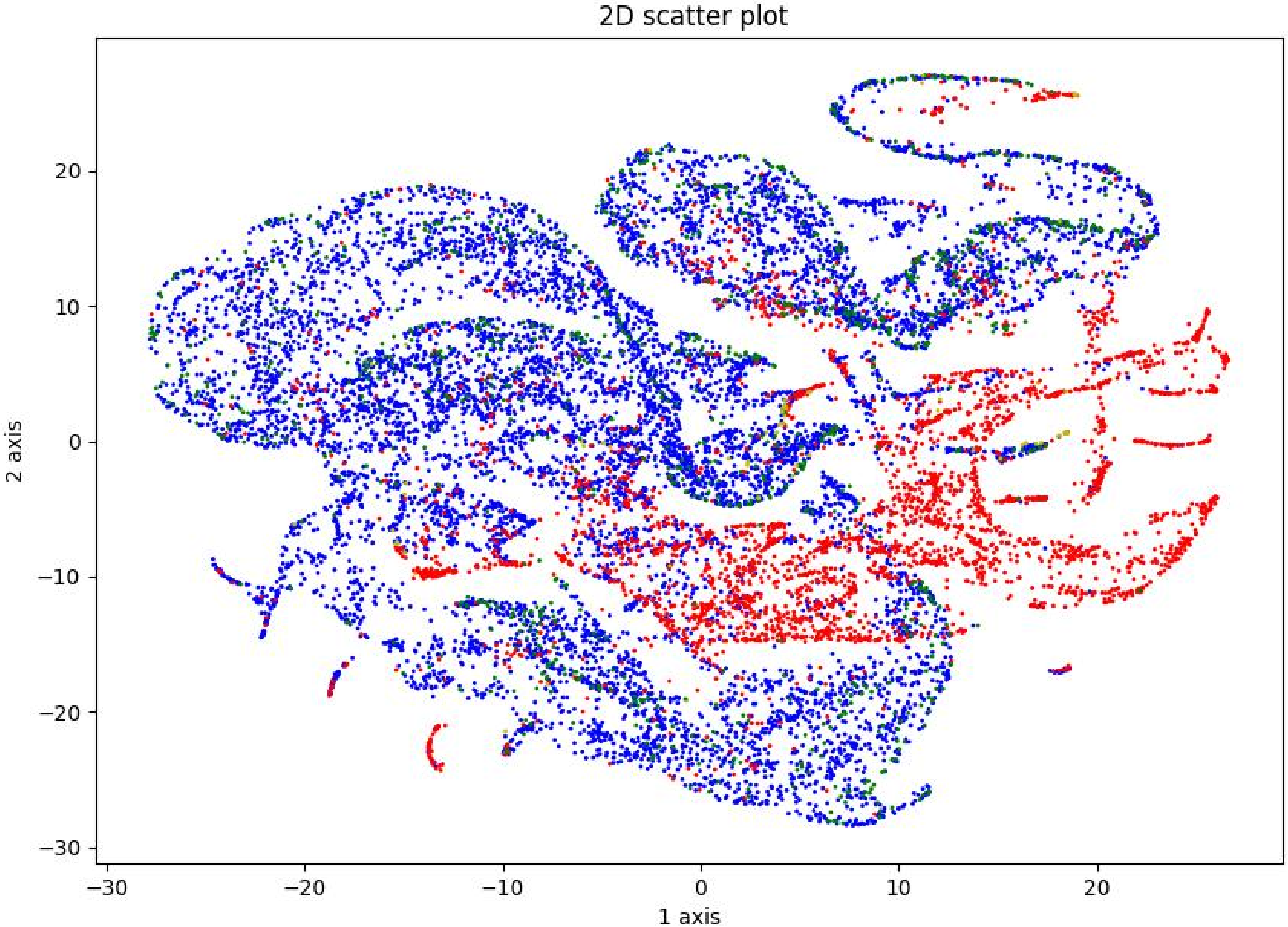
Umap visualization 3D Isomap

AGP (9511) – blue
PTB (3457) – red
T2D (1044) – green
IBD (86) – yellow



n_neighbors=30
The same picture from different points of view

NCVis visualization 2D Isomap



AGP (9511) – blue
PTB (3457) – red
T2D (1044) – green
IBD (86) – yellow