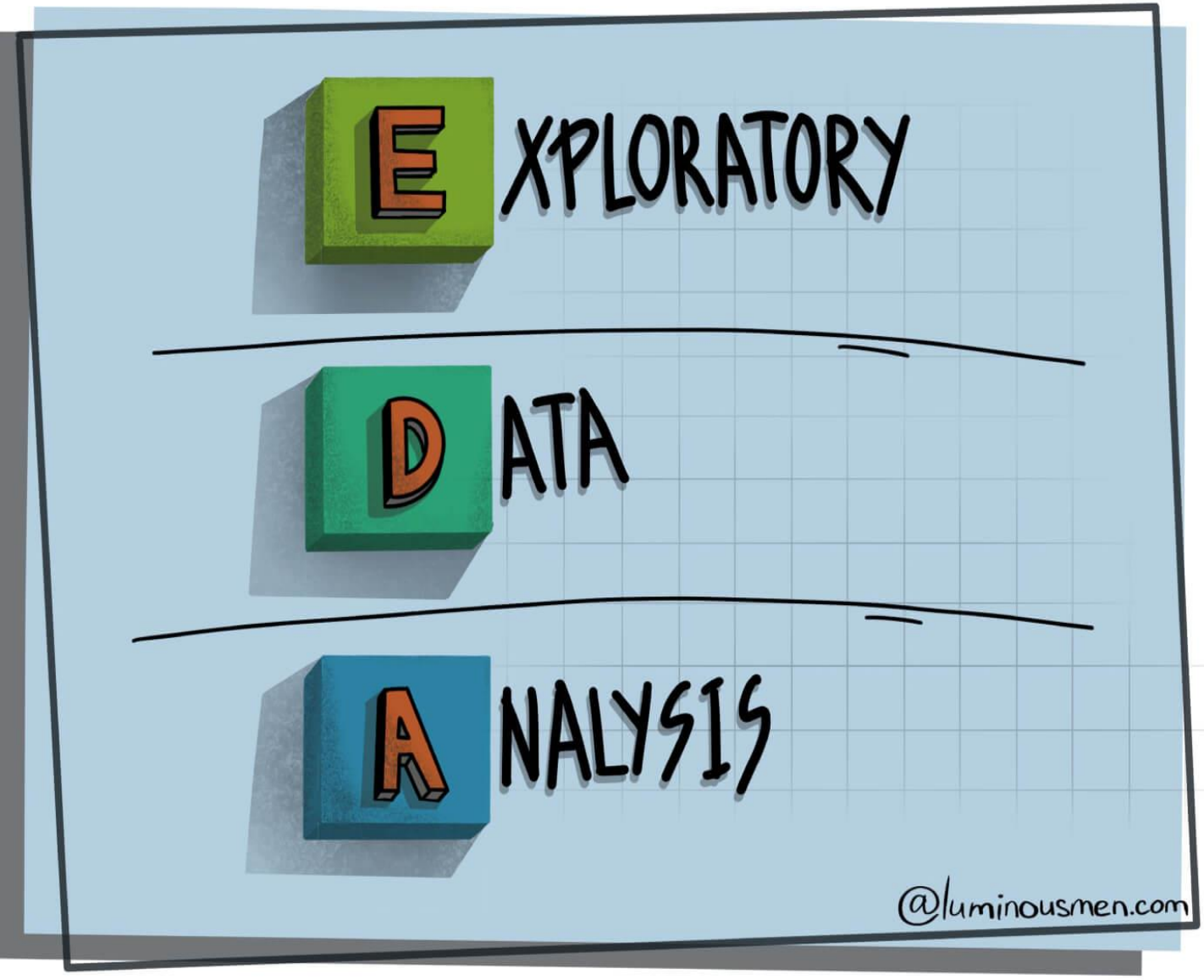


ENZYMES

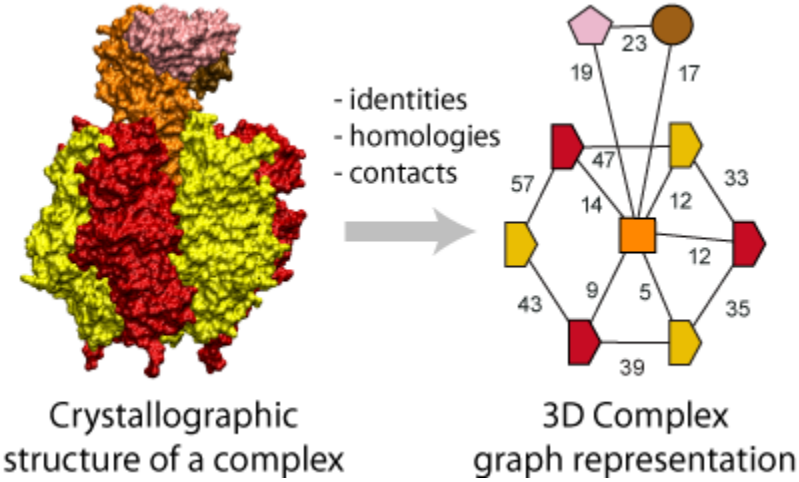
Peter Hunyadi & Nikola Vinko

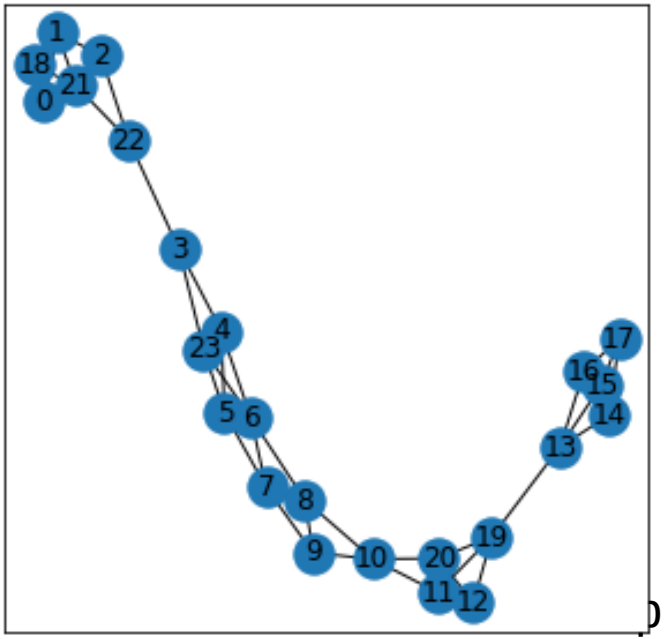
ENZYMES Overview

- 600 enzymes
- Graph representation of an enzyme
- 6 EC enzyme labels: reflect catalyzed chemical reaction

Example graph

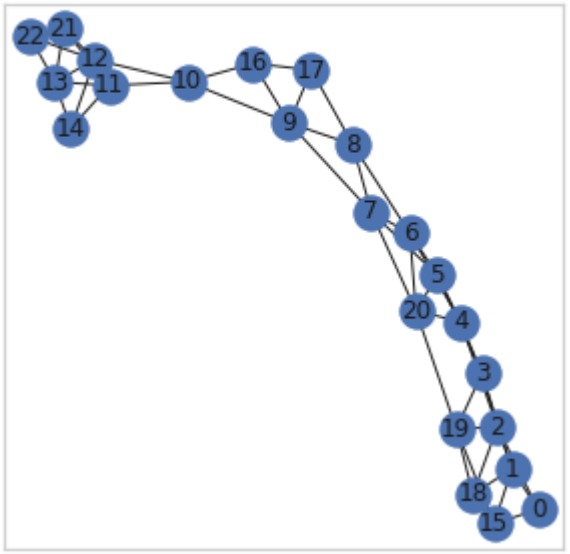
- Node represent secondary structure elements (SSE)
  - - Helix,sheet or turn
- Edge connects two SSE that are neighbors along the amino acid sequence or are one of three nearest neighbors in space



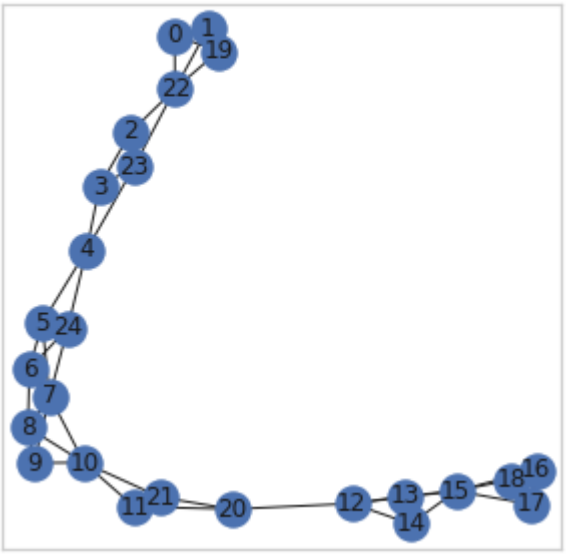


Examples of representative enzymes

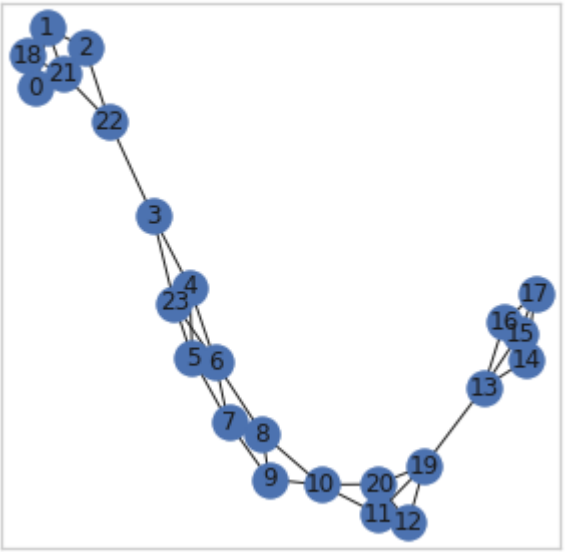
EC1



EC2



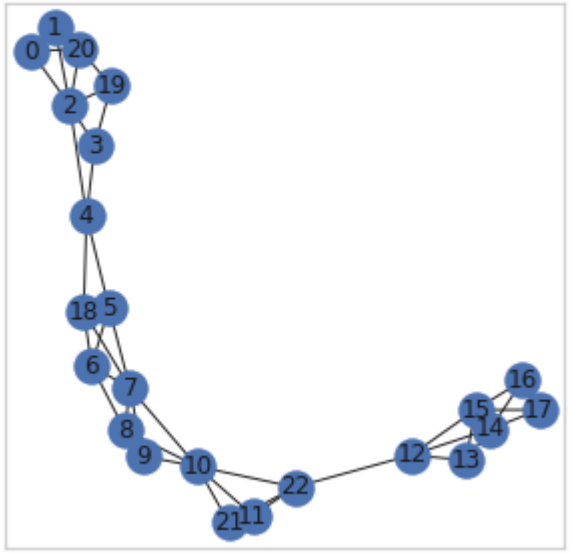
EC3



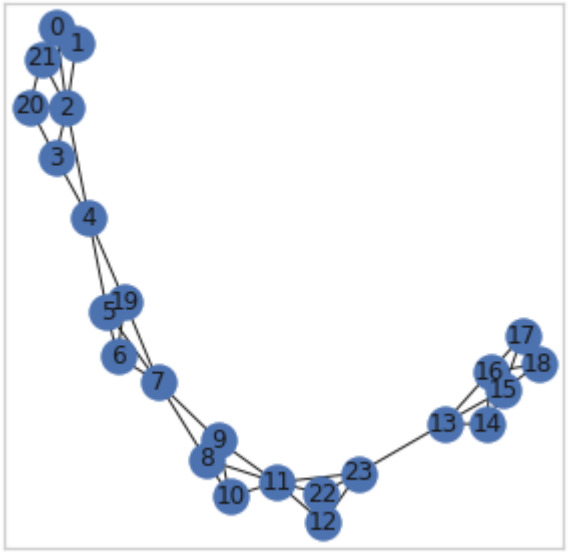
Group
0

Examples of representative enzymes

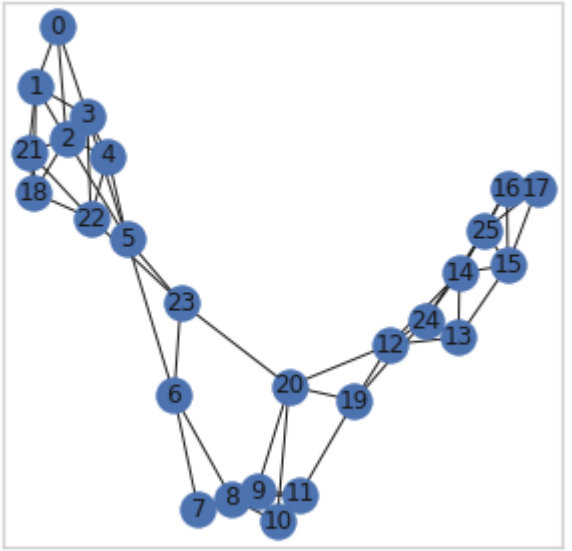
EC4



EC5



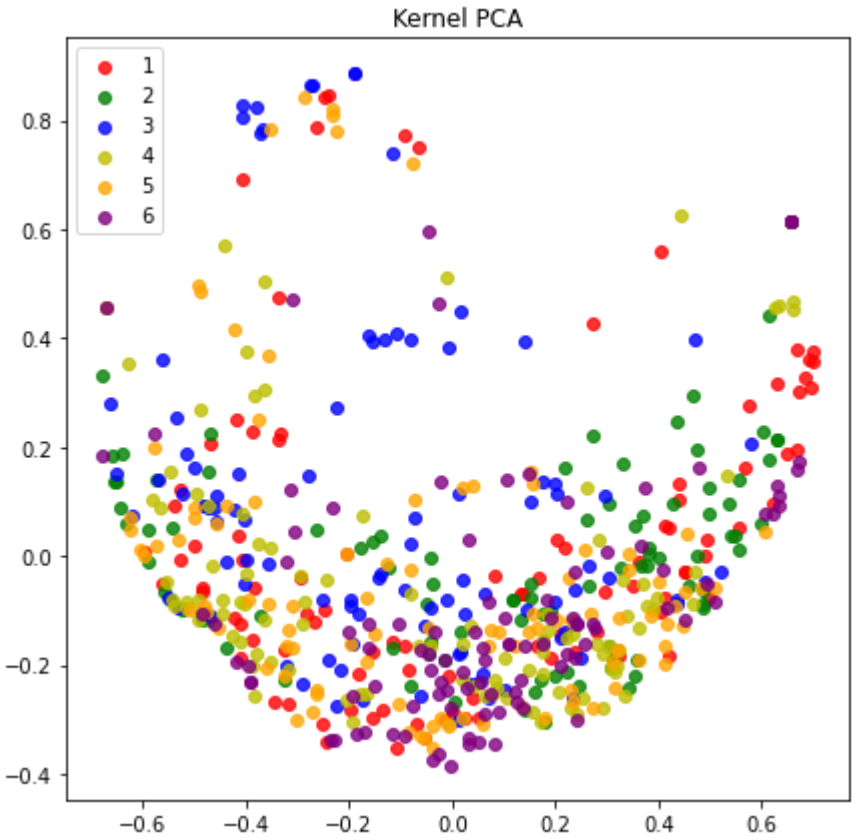
EC6



Group
0

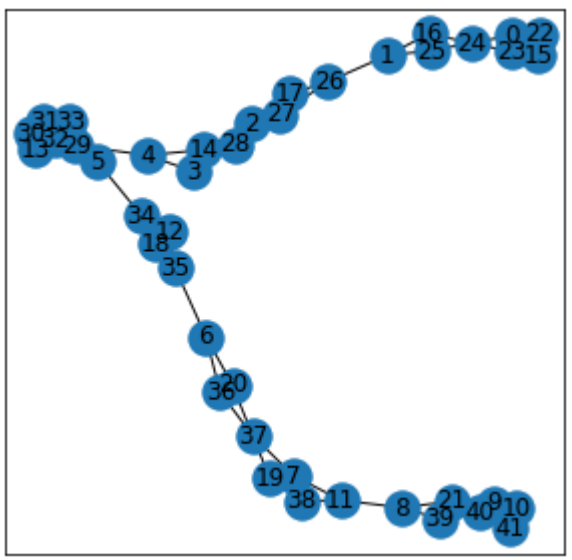
Dimensionality Reduction on unprocessed data

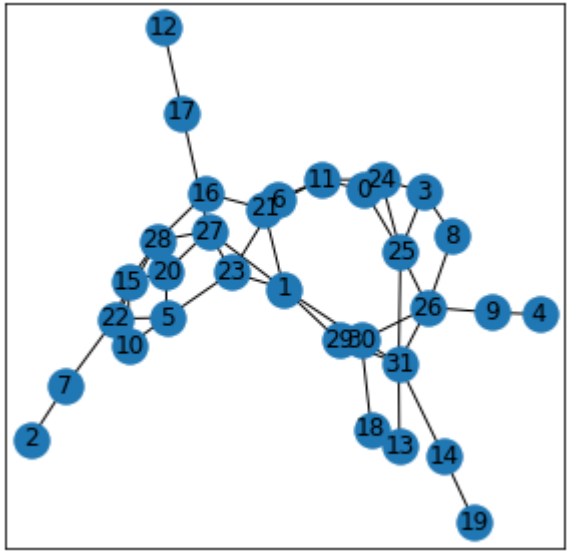
- We see some outlier points which are far separated from the main cluster

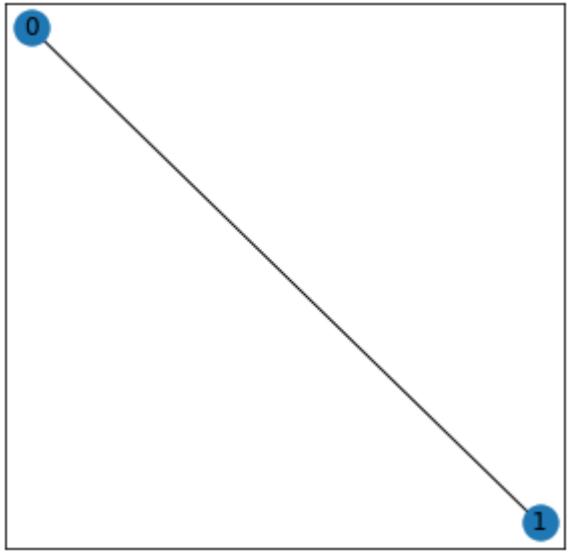


Group
9

Visualizing outliers: unique structure and high/low number of edges/nodes

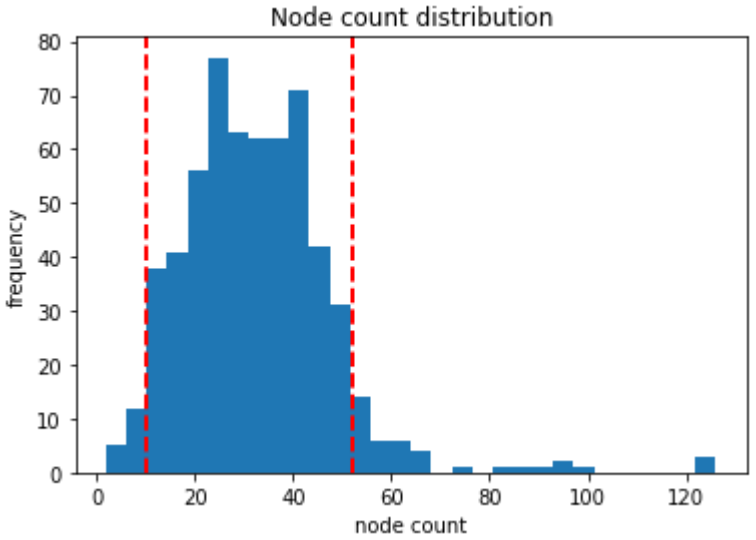


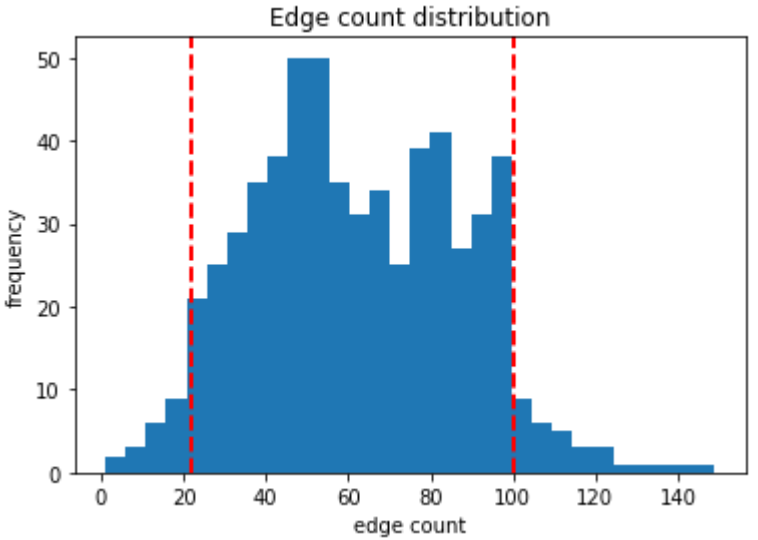




Group
0

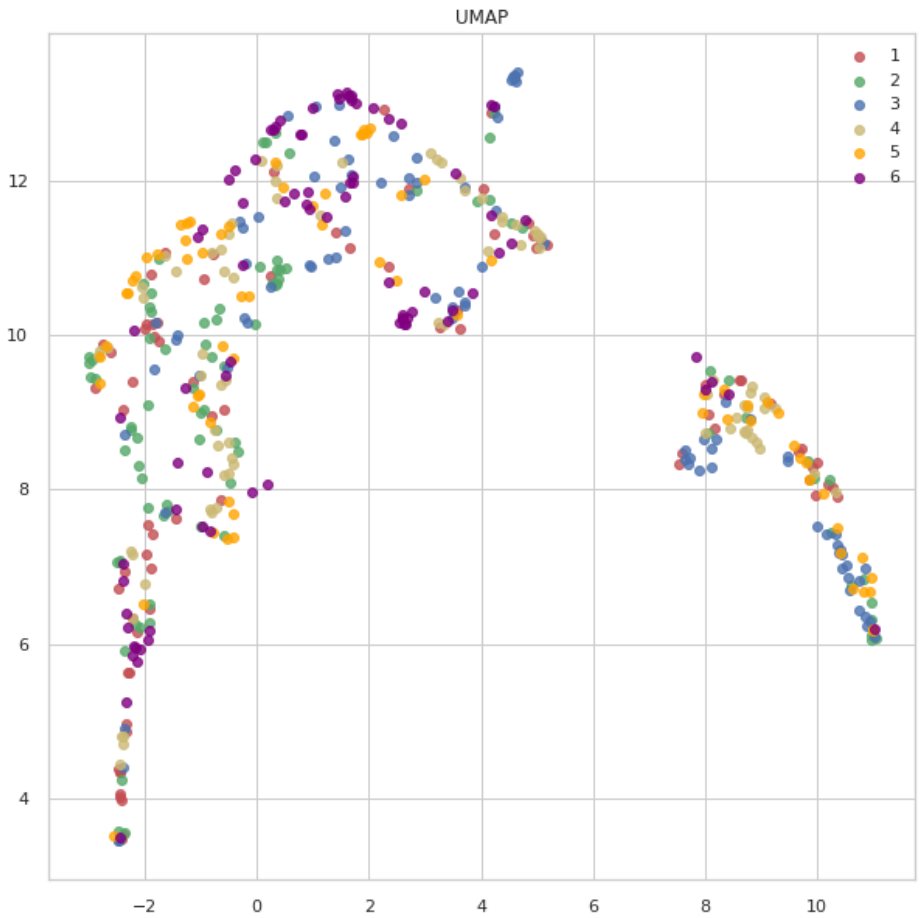
Preprocessing: removing outliers and graphs with extreme edge/node counts

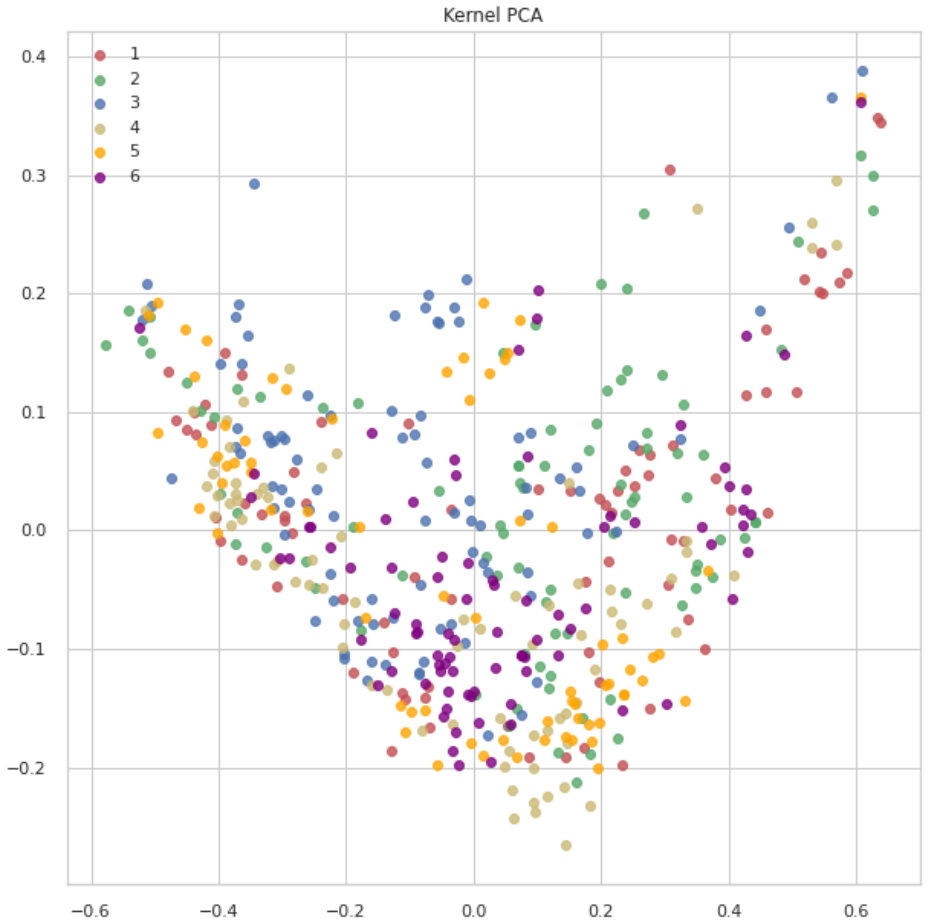




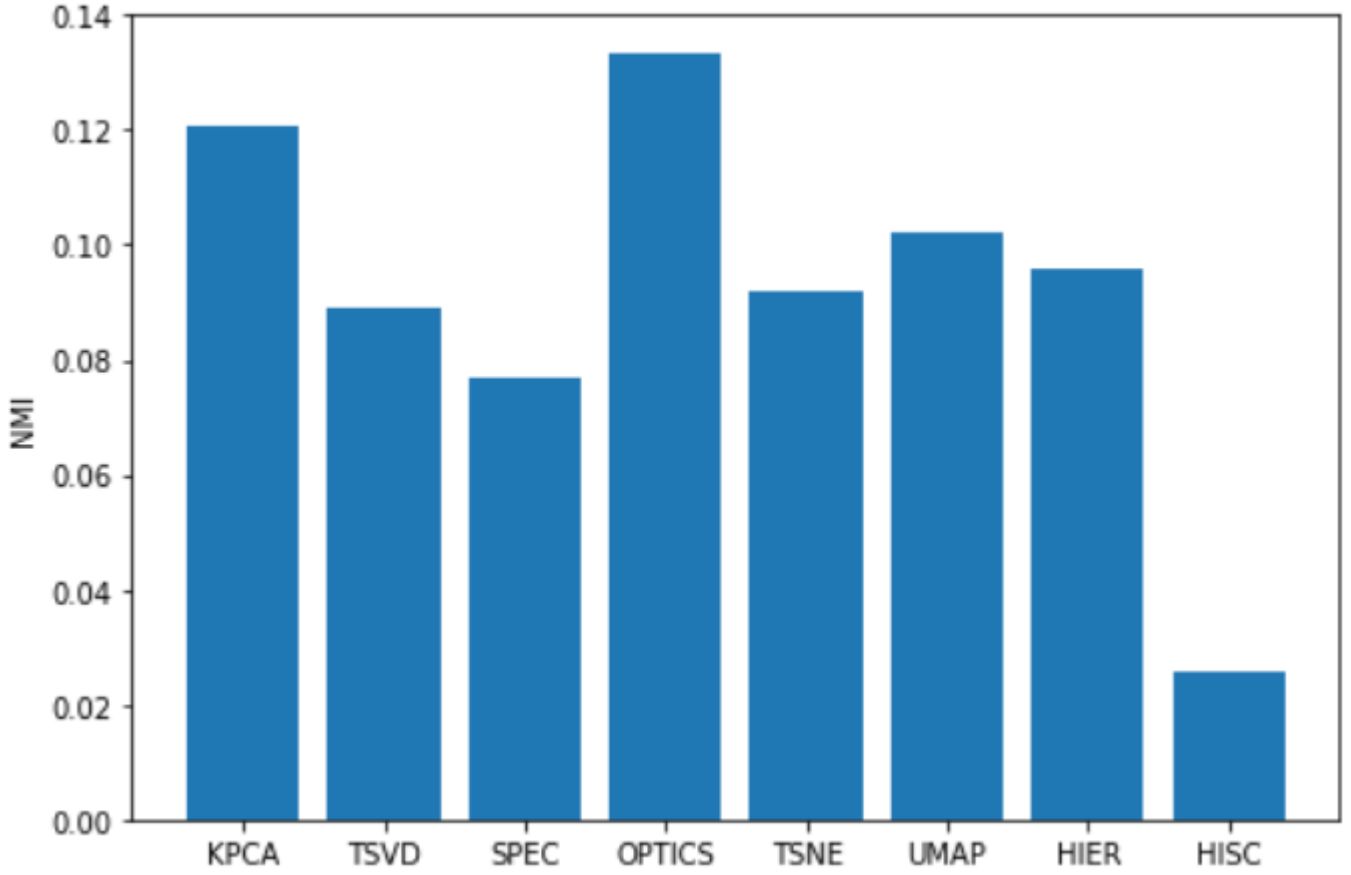
Group 0

Dimensionality Reduction on labeled dataset: WL4





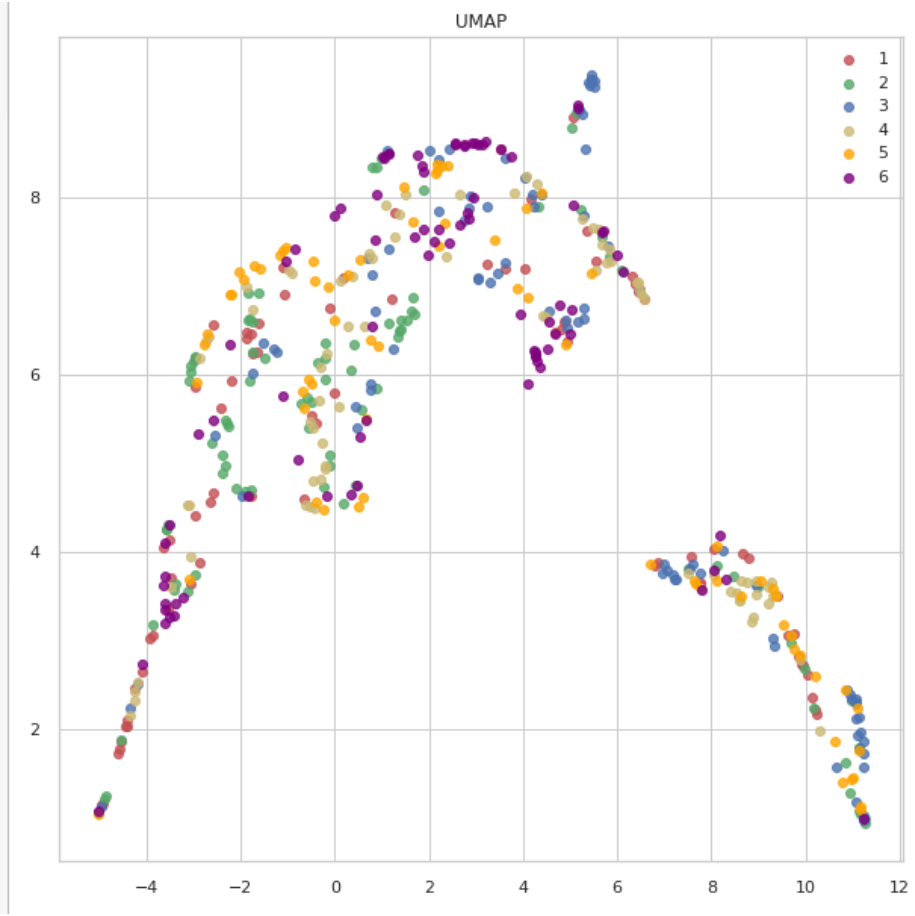
Cluster evaluation based on NMI



Group
0

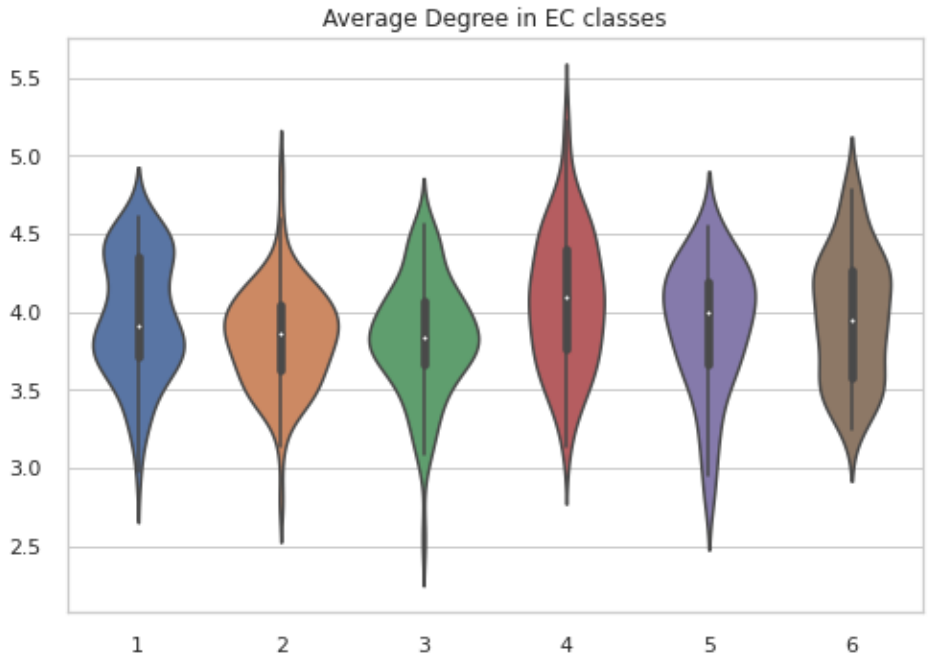
Clusters do not represent the true labels

True labels

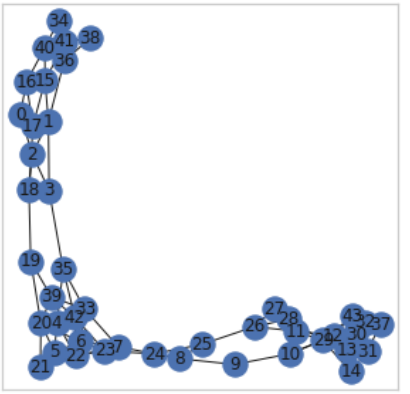


Kmeans Clusters

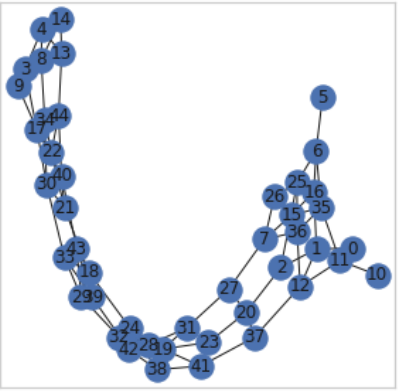
- Clusters have different degree distributions



Cluster 1, Indices [233 240]

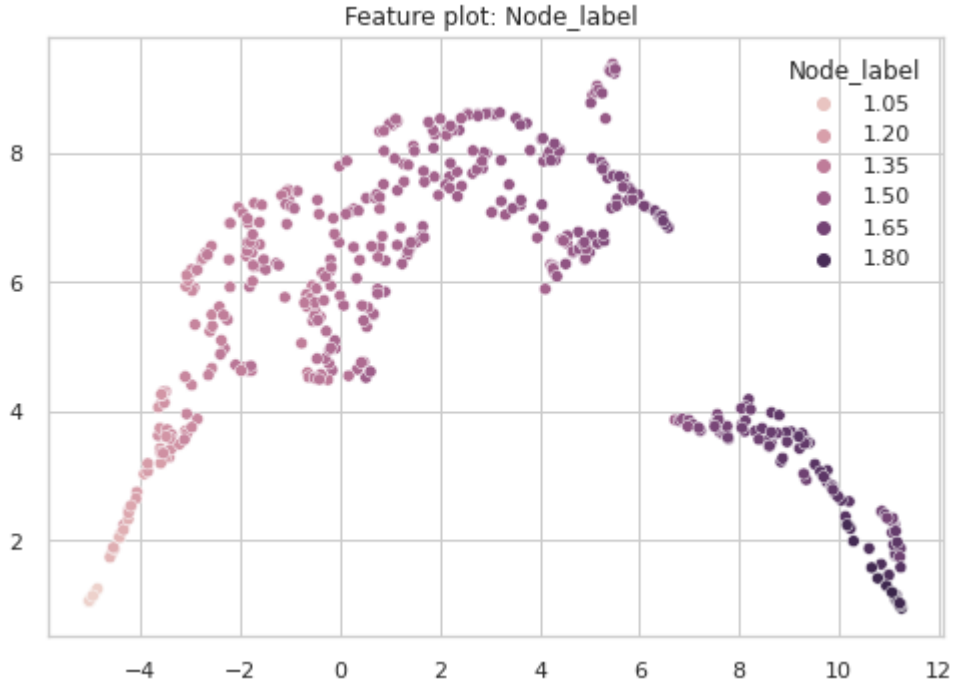


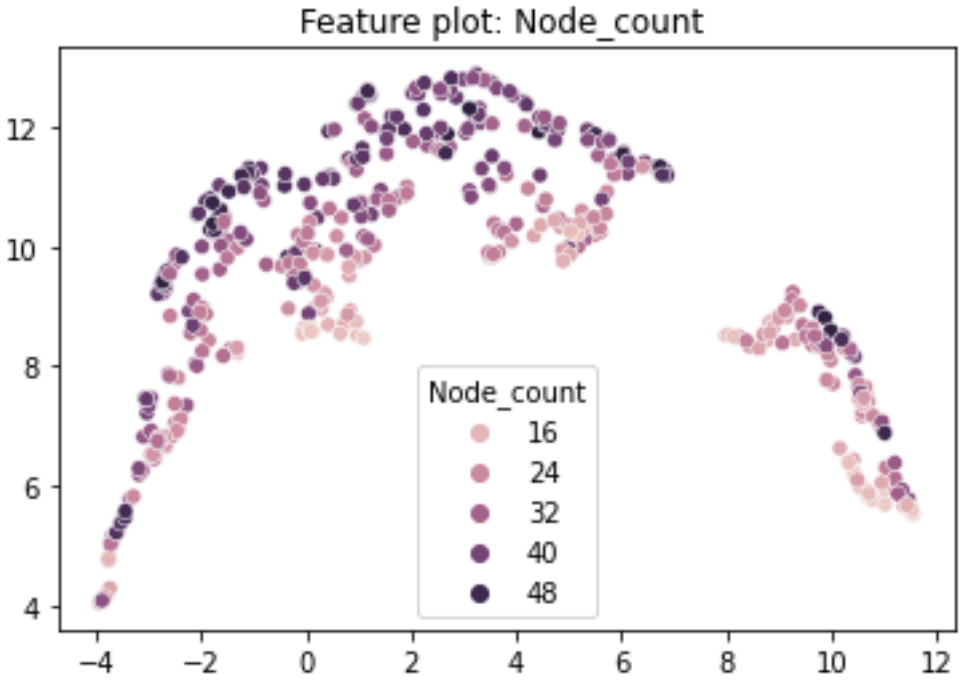
Cluster 2, Indices [485 232]



Group
0

DR describes node labels and node count

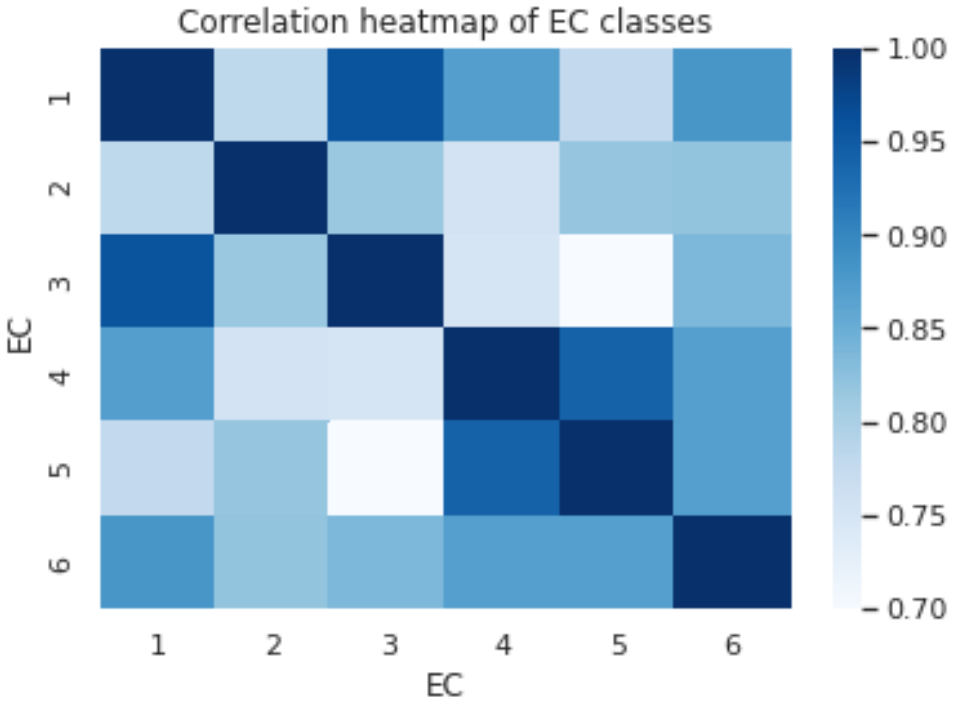




Group
0

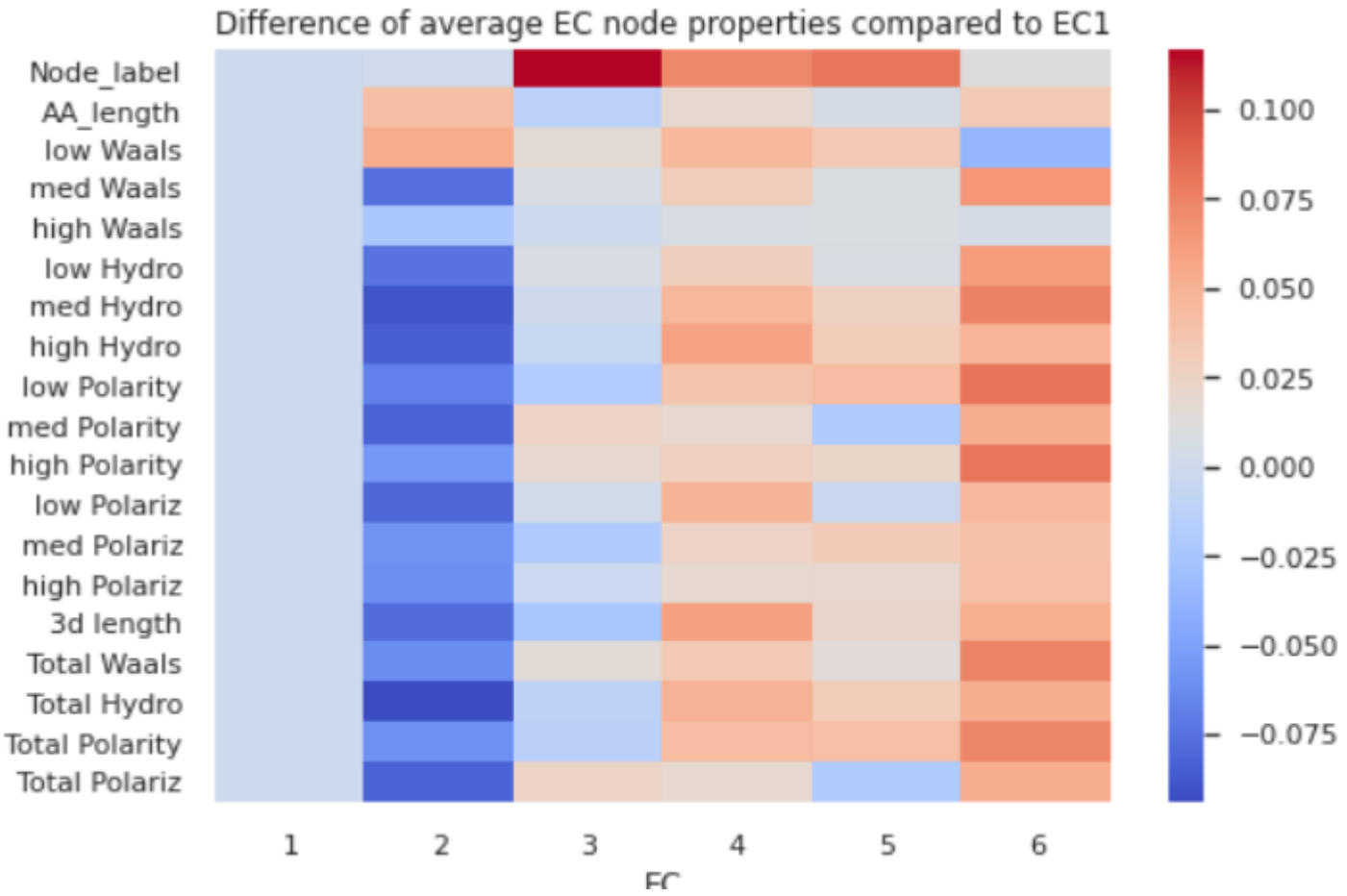
True label analysis based on node properties

- We notice higher correlation between EC 4,5
- Also EC 1,3 are highly correlated



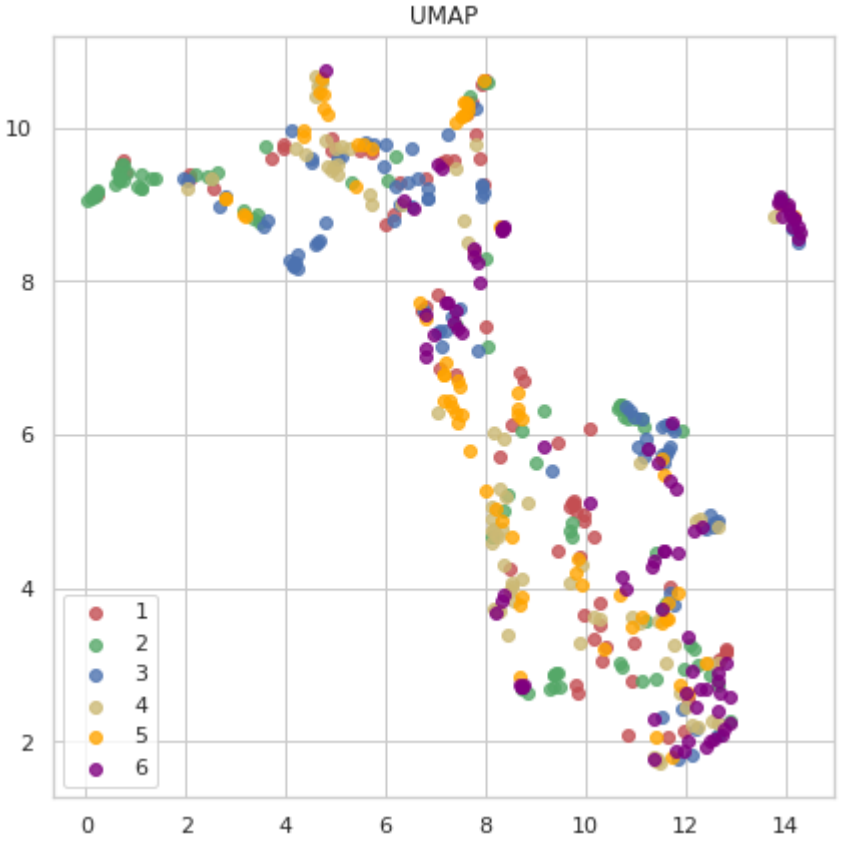
Group

Difference between EC1 and other EC classes



Group
0

UMAP of only node attribute data (wtihout labels)



Group
0

- SVM accuracy comparison of kernels

WL iterations 1-5 With preprocessing	0,44898
	0,56122
	0,57143
	0,62245
	0,61225
WL iterations 1-5 WithOUT preprocessing	0,45834
	0,575
	0,625
	0,625
	0,60834
Graphlet	0,32653
Shortest Paths	0,48979

Conclusion

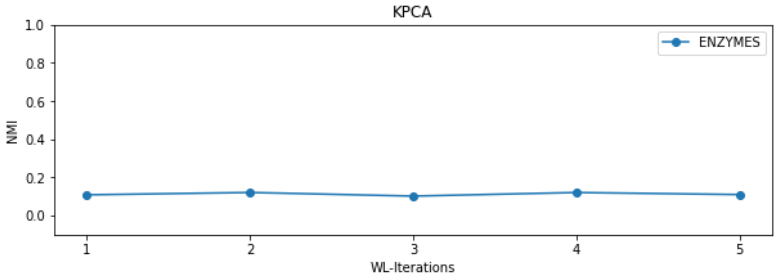
- UMAP vector representation is strongly driven by node labels (SSE elements), edge/node count
- Clusters do not represent the ground truth labels
- Clusters are formed based on different SSE composition, and structural differences e.g. degree distribution
- Node attributes play also an important role in EC classification which could be used for building better kernels

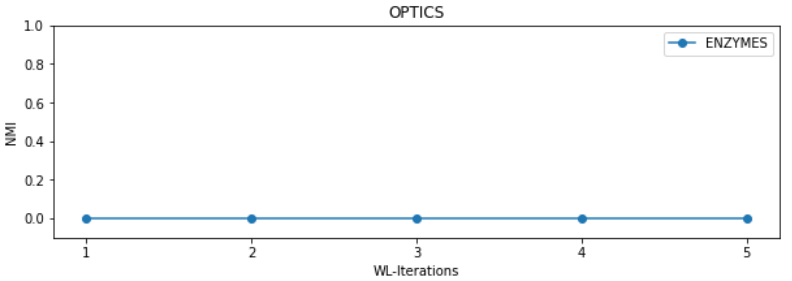
Thank you for your attention!

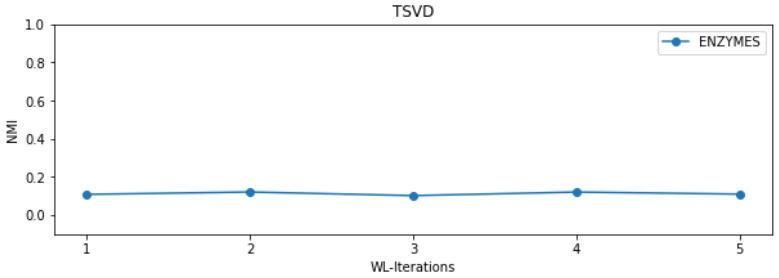
- NMI values for the 3 kernels (with removing the extreme values first)

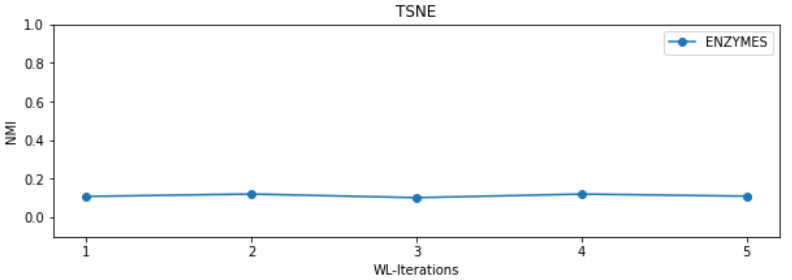
	WL iter 3	**Graphlet**	**Shortest Paths**
Kmeans	0,1018	0,0828	0,1091
Hierarchical Ward	0,2079	0,1710	0,1866
Hierarchical Complete	0,1774	0,1576	0,1601
Hierarchical Average	0,0446	0,0887	0,0602
Hierarchical Single	0,0338	0,0551	0,0508
Subkmeans	0,0833	0,0717	0,0754

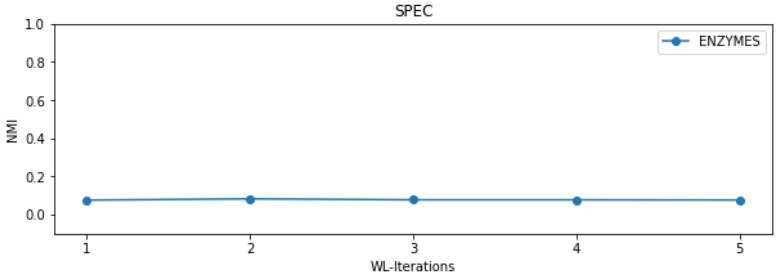
Overview of NMI performance for different representations

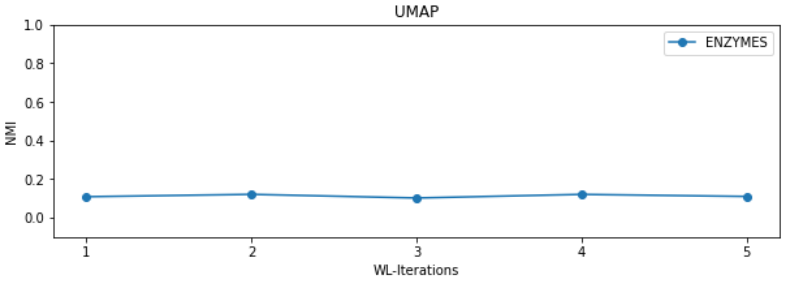




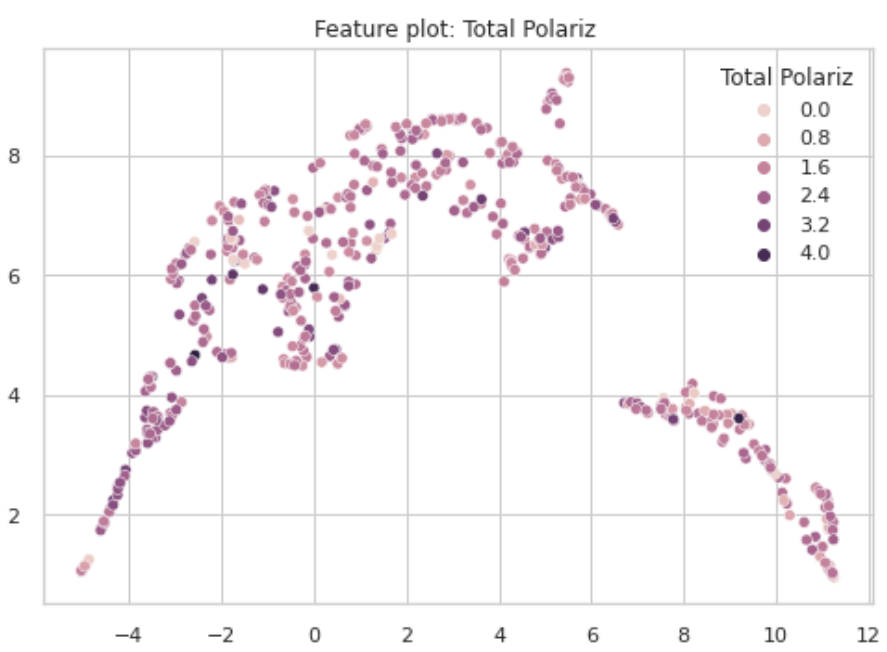


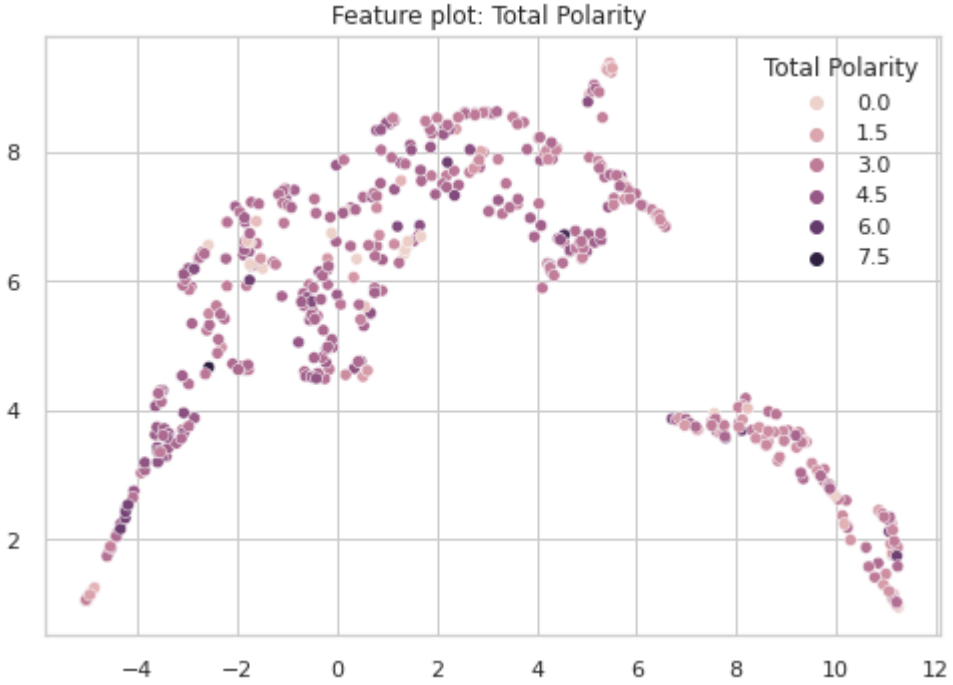






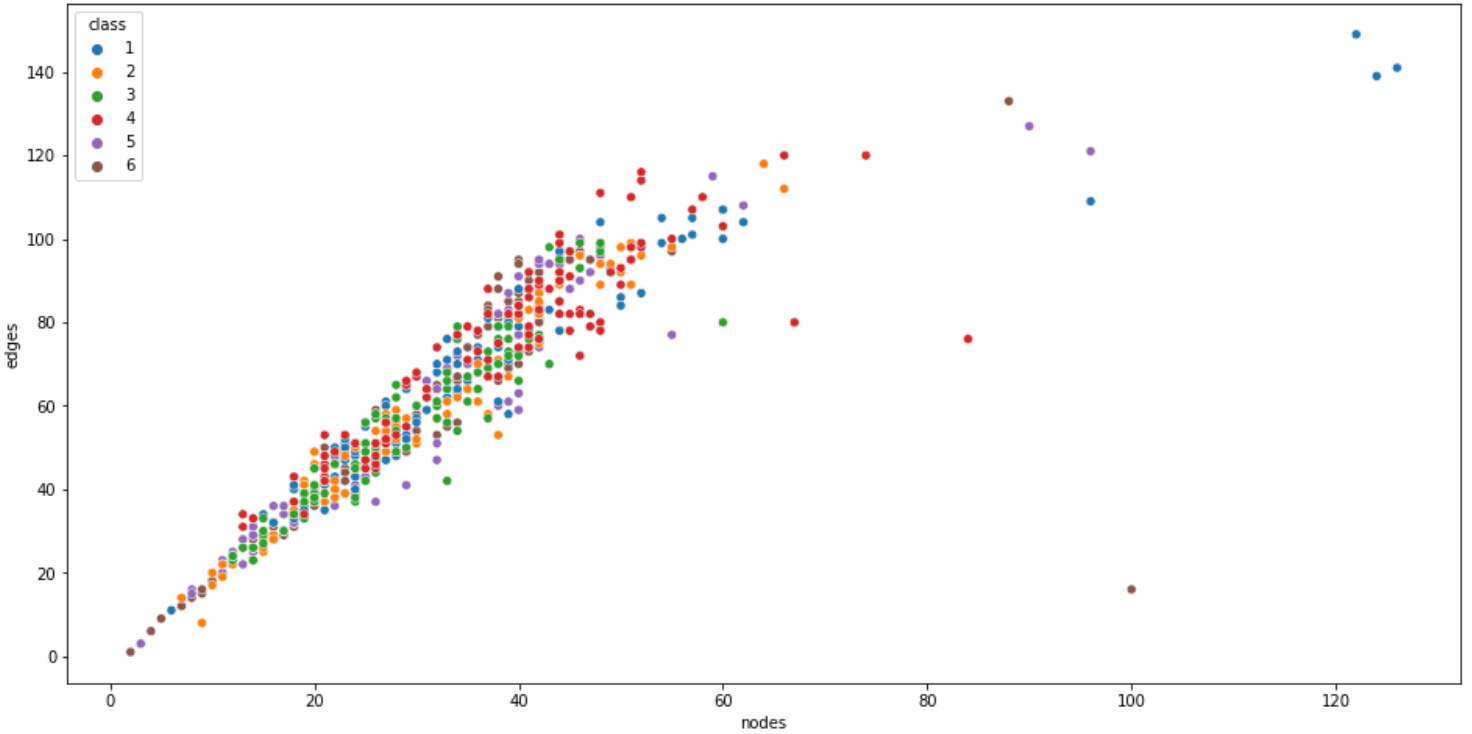
Group
9


Feature plot: Total Polariz
Total Polariz
0.0
0.8
1.6
2.4
3.2
4.0
8
6
4
2
−4
−2
0
2
4
6
8
10
12


Feature plot: Total Polarity
Total Polarity
0.0
1.5
3.0
4.5
6.0
7.5
8
6
4
2
−4
−2
0
2
4
6
8
10
12

Group
0

Scatter plot of avg. nodes and avg. edges



Group
9