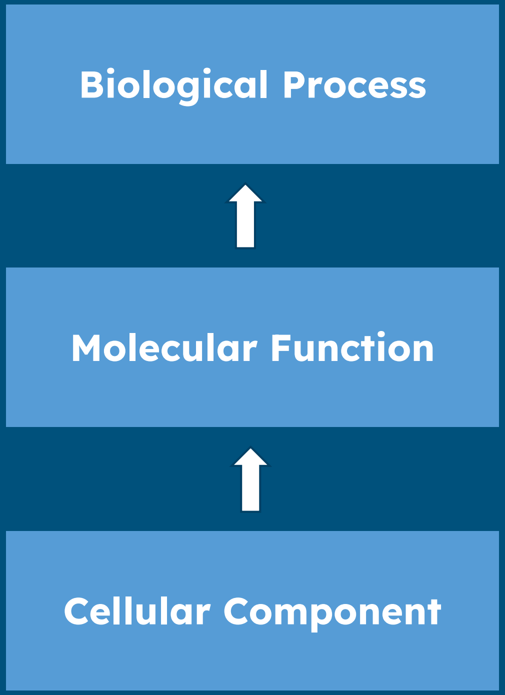
GO/HPO

Porcelli Angelica
Roveda Gianluca
Stefanelli Marta

WORK FLOW

0. The dataset
1. Feature selection
2. Jaccard function
3. TF - IDF
4. Three different views and HPO
5. Similar Network Fusion
6. HDBSCAN and UMAP clusters
7. Cluster analysis
8. Train MLP

DATASET

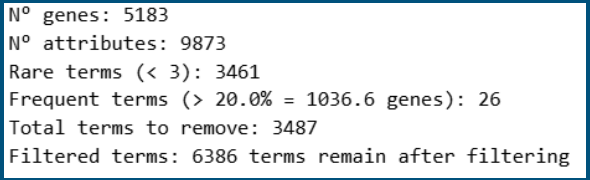


Dataset of gene, with four different binary representation:

- **CC** = Cellular Component. (Where?)
- **MF** = Molecular Function.(What?)
- **BP** = Biological Process.(In?)
- **HPO** = Phenotype.

FEATURE SELECTION

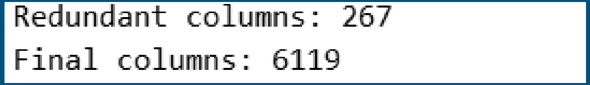
Frequency filtering



```
Nº genes: 5183
Nº attributes: 9873
Rare terms (< 3): 3461
Frequent terms (> 20.0% = 1036.6 genes): 26
Total terms to remove: 3487
Filtered terms: 6386 terms remain after filtering
```

- Removes terms that annotate **< 3 genes**
- Removes terms present in **> 20% of genes**
- Eliminates both overly frequent and overly rare terms.

Redundant column removal



```
Redundant columns: 267
Final columns: 6119
```

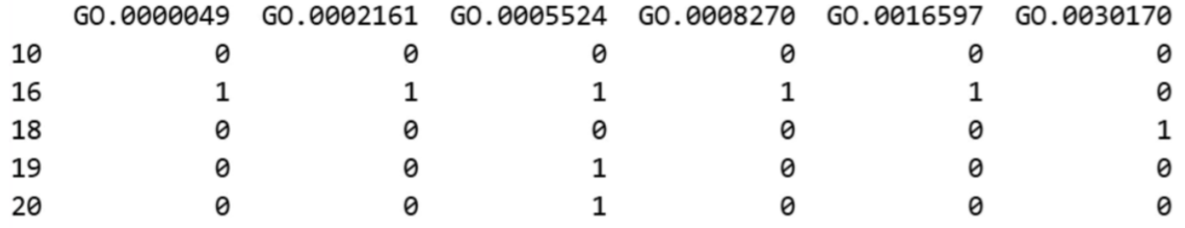
- Uses Jaccard on sparse matrices
- Removes quasi-identical columns (**Jaccard ≥ 0.9**)

SINGLE **JACCARD** FUNCTION

$$J = \frac{|intersection|}{|union|}$$

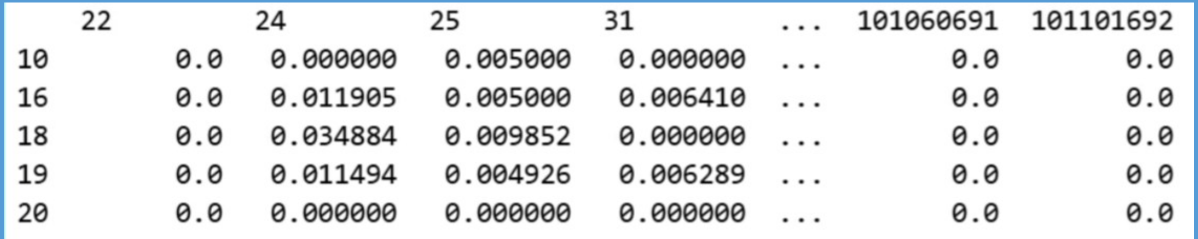
For each view, it takes a **matrix** and uses the **Jaccard index** to build the similarity matrices for each view.

Input



```
    GO.0000049  GO.0002161  GO.0005524  GO.0008270  GO.0016597  GO.0030170
10           0           0           0           0           0           0
16           1           1           1           1           1           0
18           0           0           0           0           0           1
19           0           0           1           0           0           0
20           0           0           1           0           0           0
```

Output



```
     22         24         25         31  ...  101060691  101101692
10  0.0   0.000000   0.005000   0.000000  ...        0.0        0.0
16  0.0   0.011905   0.005000   0.006410  ...        0.0        0.0
18  0.0   0.034884   0.009852   0.000000  ...        0.0        0.0
19  0.0   0.011494   0.004926   0.006289  ...        0.0        0.0
20  0.0   0.000000   0.000000   0.000000  ...        0.0        0.0
```

TF - IDF

Converts the **gene × term** matrix **into sparse (CSR) format**.
Computes the **document frequency** (df_j) **for each term**.
Computes **IDF = log(N / df_j)**

Weights each term using TF-IDF (IDF only, since TF = 1/0)

Output:

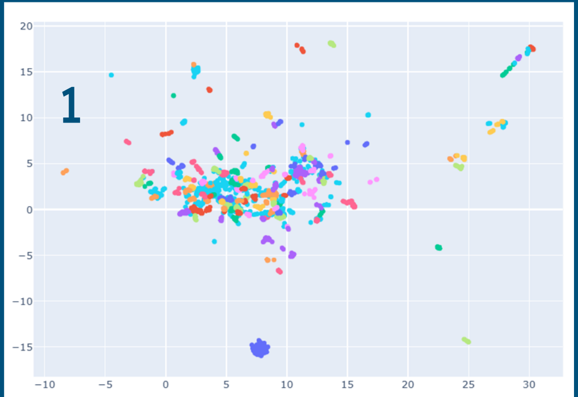
- TF-IDF matrix (gene × term)
- IDF vector for the terms

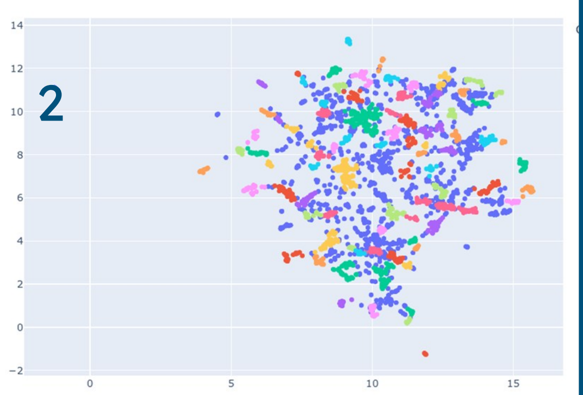
Goal: reduce the importance of very frequent terms and increase that of rare terms → more informative signals.

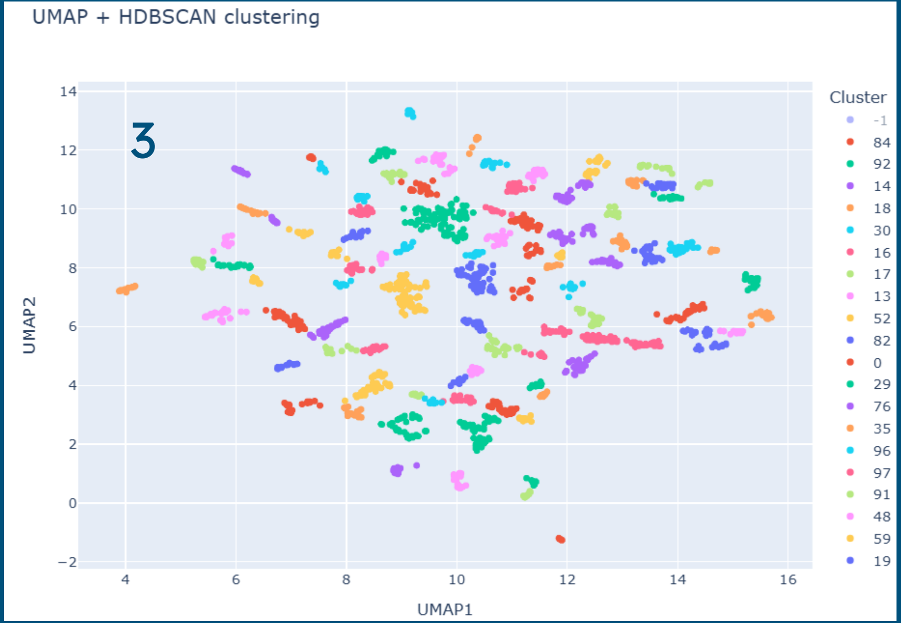
THREE DIFFERENT VIEWS

CC CLUSTERS

```
 ========== MATRIX: CC ==========

 N genes: 5183
 N terms: 1478

Input: 3248 × 882


Clusters: 98, noise points: 861
```



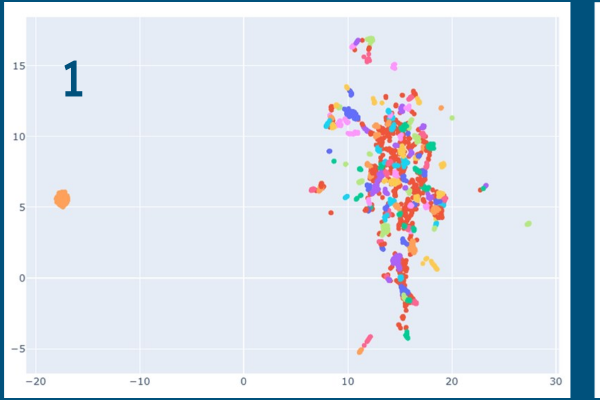


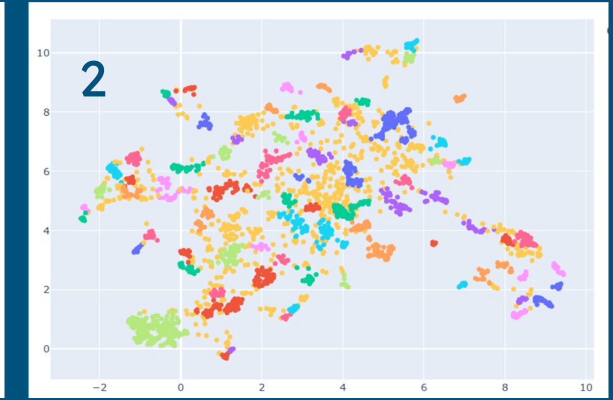

UMAP + HDBSCAN clustering

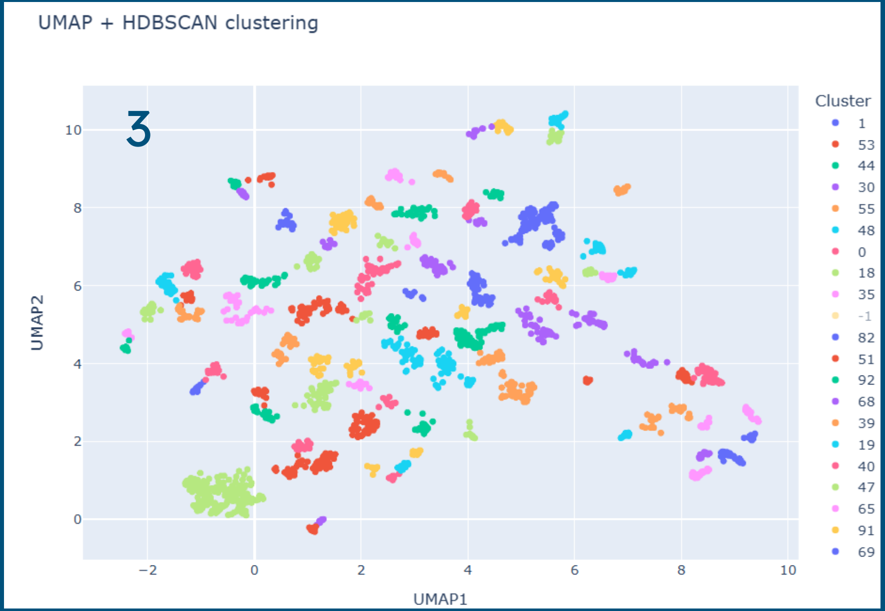
THREE DIFFERENT VIEWS

MF CLUSTERS

```
========= MATRIX: MF =========

N genes: 5183
N terms: 3258

Input: 3578 × 1337

Clusters: 94, noise points: 693
```

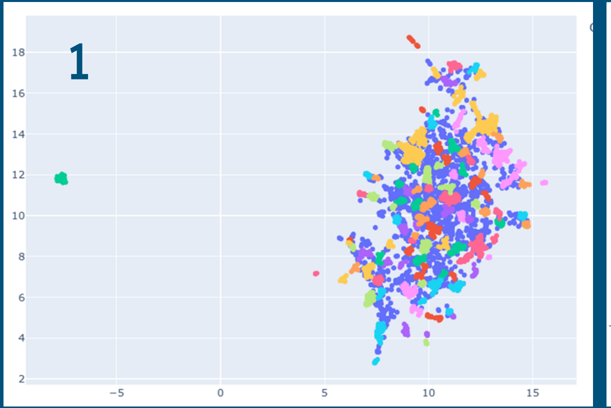


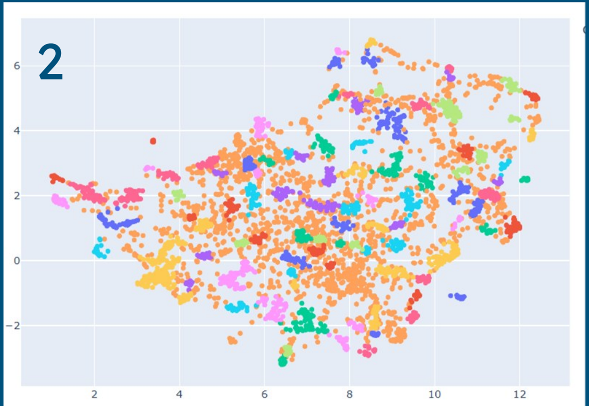




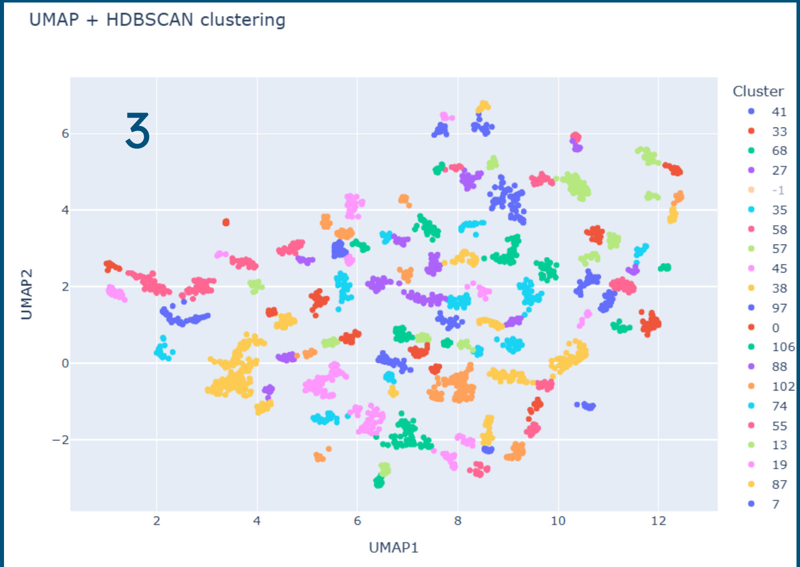
THREE DIFFERENT VIEWS

BP CLUSTERS

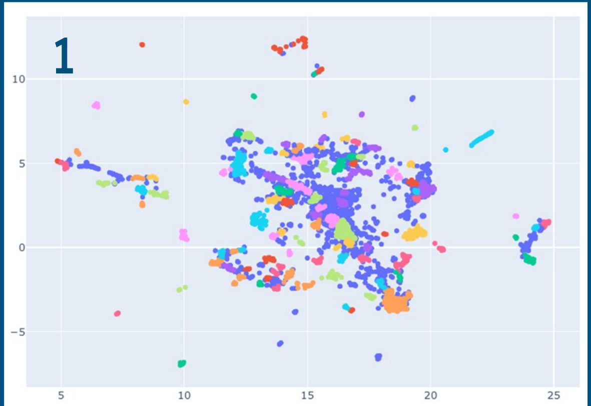


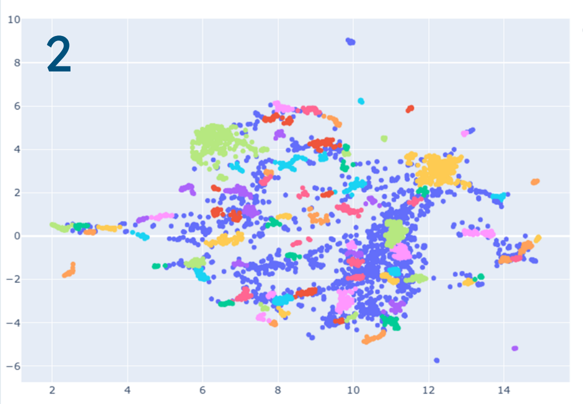


```
========== MATRIX: BP ==========

N genes: 5183
N terms: 9873

Input: 4786 × 6119

Clusters: 109, noise points: 1536
```

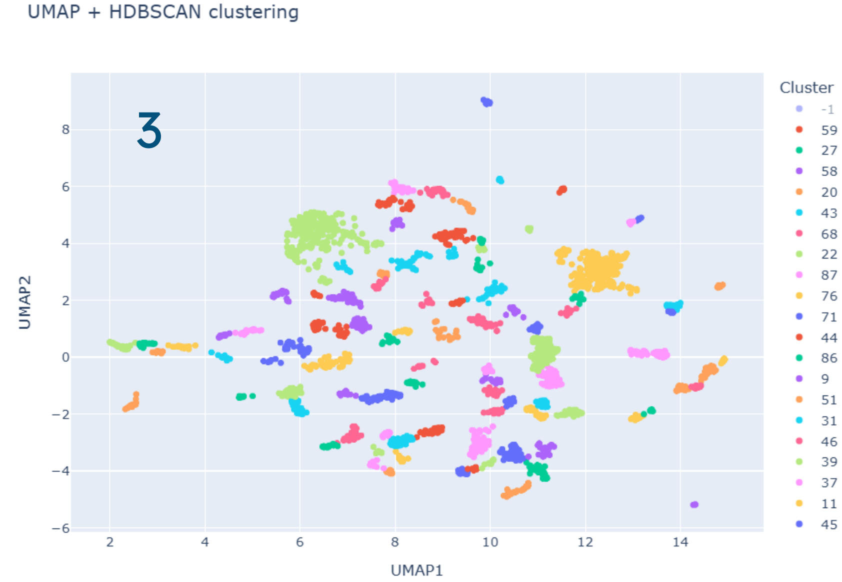

UMAP + HDBSCAN clustering

HPO CLUSTERS



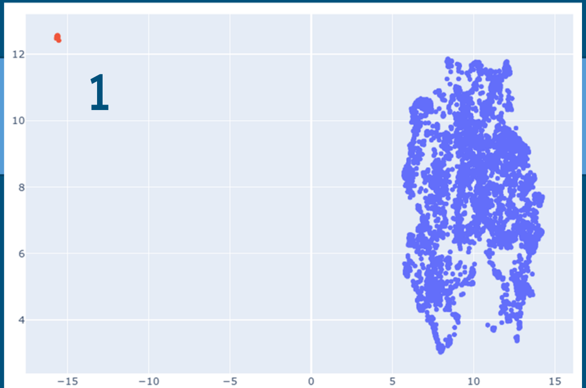


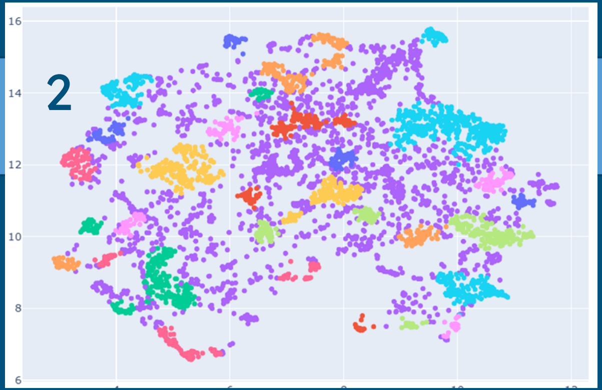
```
========= MATRIX: HPO =========

N genes: 5183
N terms: 10185

Input: 4702 × 6342

Clusters: 100, noise points: 1278
```



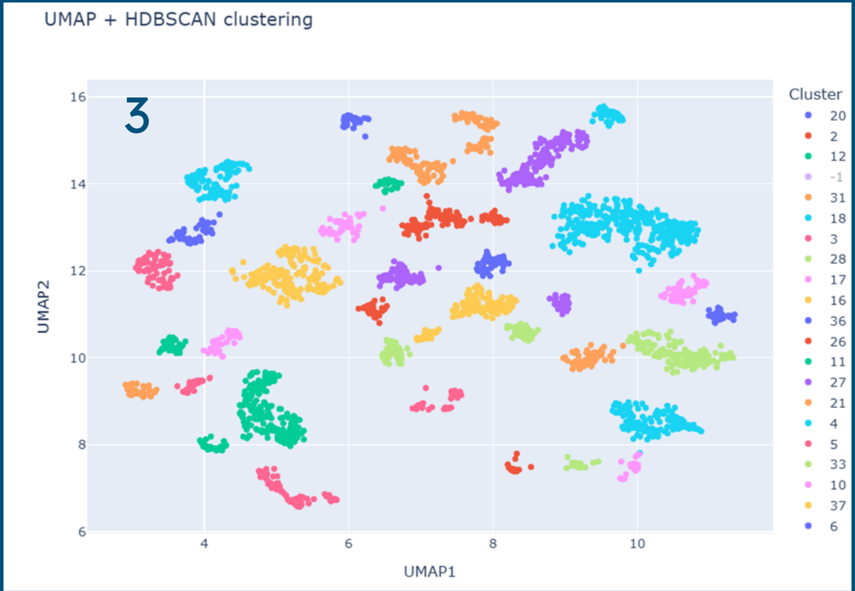
CURIOSITY FOR

Composed (CC-MF-BP)

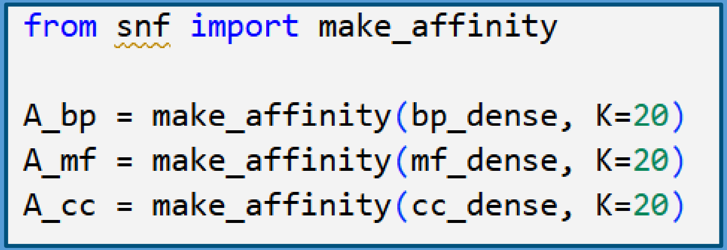




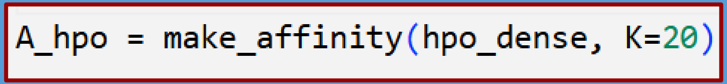
```
======== MATRIX: COMPOSED ========

N genes: 5183
N original terms: 14609

Input: 5117 × 8265

Clusters: 38, noise points: 1810
```


UMAP + HDBSCAN clustering

PREPARATION FOR SNF



```
from snf import make_affinity

A_bp = make_affinity(bp_dense, K=20)
A_mf = make_affinity(mf_dense, K=20)
A_cc = make_affinity(cc_dense, K=20)
```



```
A_hpo = make_affinity(hpo_dense, K=20)
```

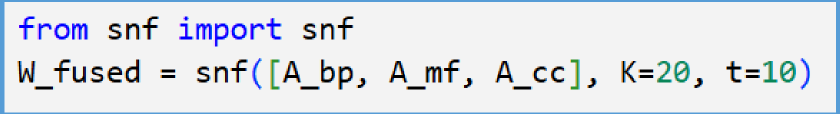
Transforms the raw similarities into a **robust affinity network** based on the K nearest neighbors.

It is a fundamental step in the SNF workflow because it ensures:

- comparability across different views
- robustness to noise
- preservation of local structures

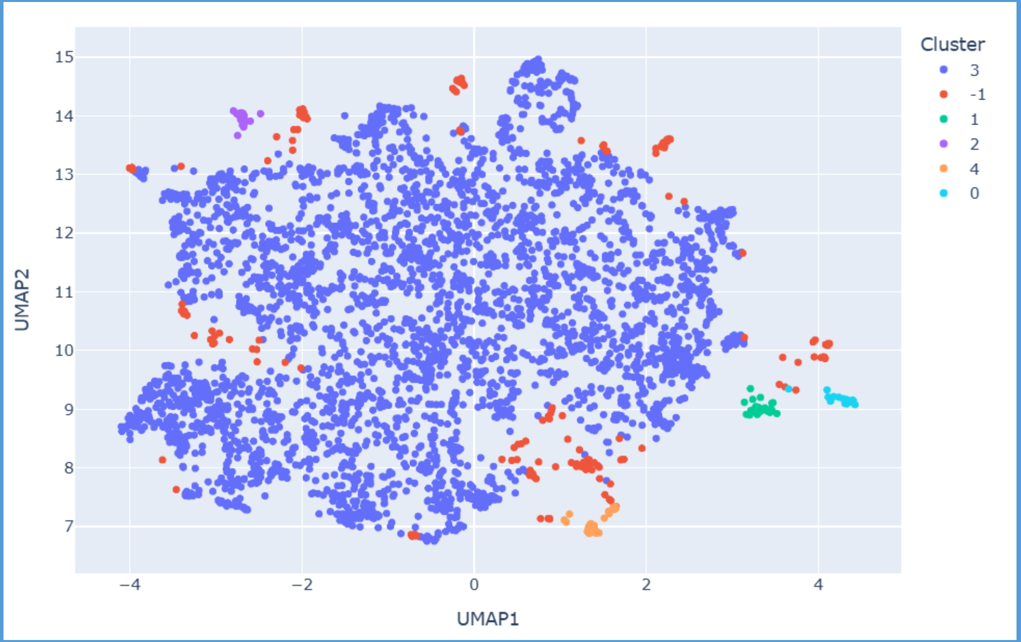
SIMILARITY NETWORK FUSION - **SNF**

SNF **merges the similarity networks of different views** (BP, MF, CC) into a **single**, more **strong network**, **amplifying shared similarities** and **reducing noise**.



```
from snf import snf
W_fused = snf([A_bp, A_mf, A_cc], K=20, t=10)
```

WHERE DID THE **HPO** GO?



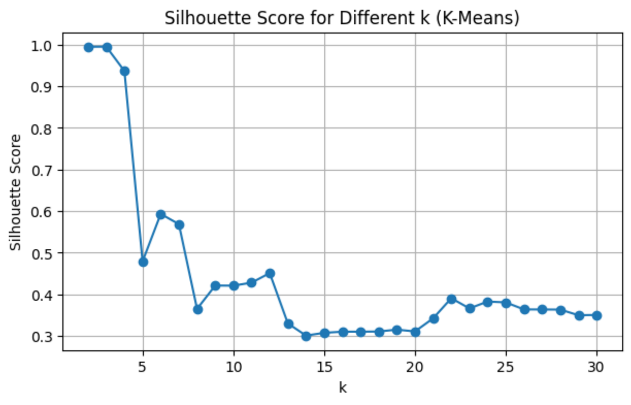
Cluster 3 is:

- poorly defined
- share few common characteristics
- elements that do not fit well into the more coherent clusters

We decided not to consider HPO but to focus on BP, MF, and CC because give us **much clearer and more coherent clusters.**

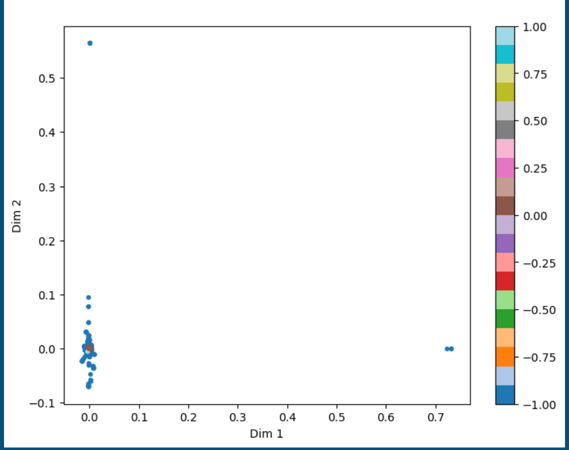
We try…

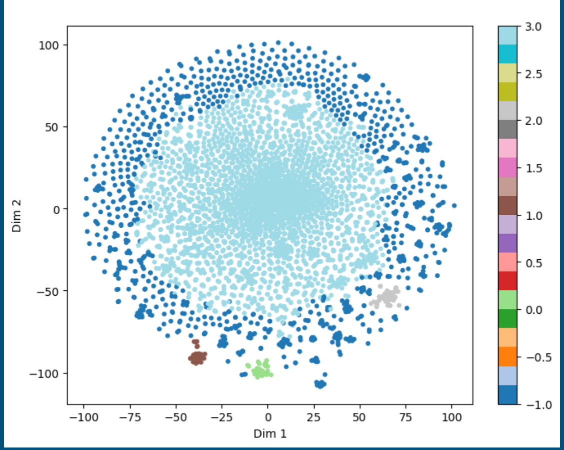
OTHER COMBINATIONS

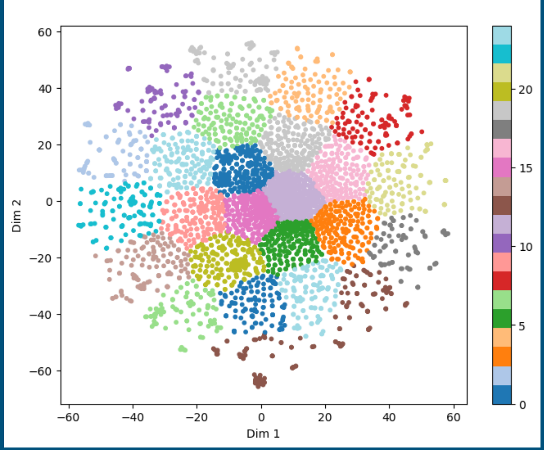


Best K = 2

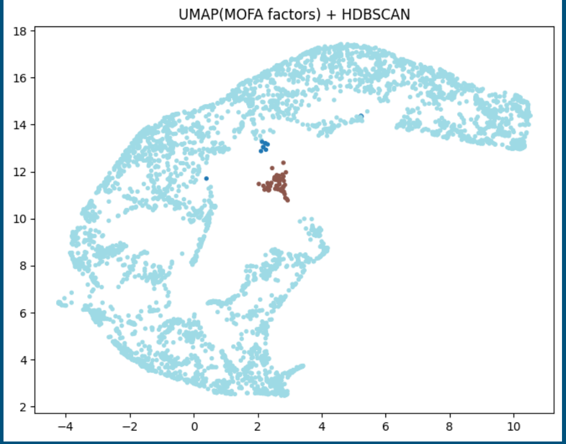
PCA + HDBSCAN	t-SNE + HDBSCAN	t-SNE + K-Means







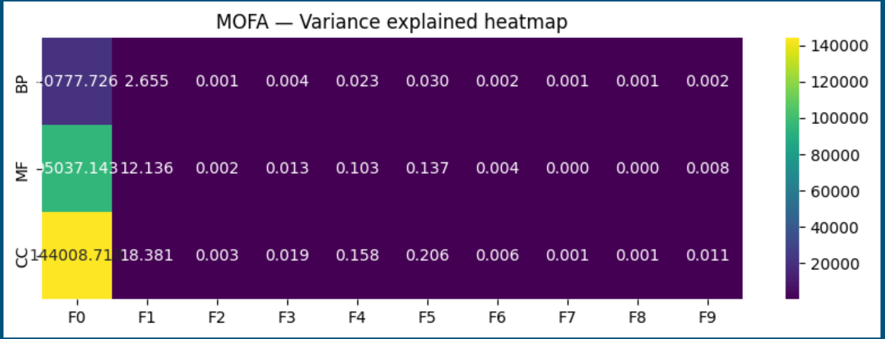
We also try… MOFA FACTORS

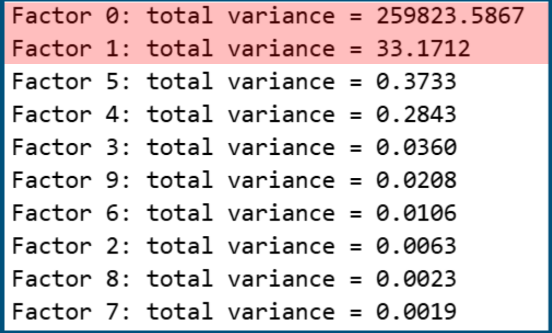


The biological processes that differentiate the samples:

- **Factor 0** → signaling/receptors/**response** to stimuli
- **Factor 1** → **differentiation** (especially neuronal)

but it **does not add** much **value to clustering**, for now…





```
Factor 0: total variance = 259823.5867
Factor 1: total variance = 33.1712
Factor 5: total variance = 0.3733
Factor 4: total variance = 0.2843
Factor 3: total variance = 0.0360
Factor 9: total variance = 0.0208
Factor 6: total variance = 0.0106
Factor 2: total variance = 0.0063
Factor 8: total variance = 0.0023
Factor 7: total variance = 0.0019
```

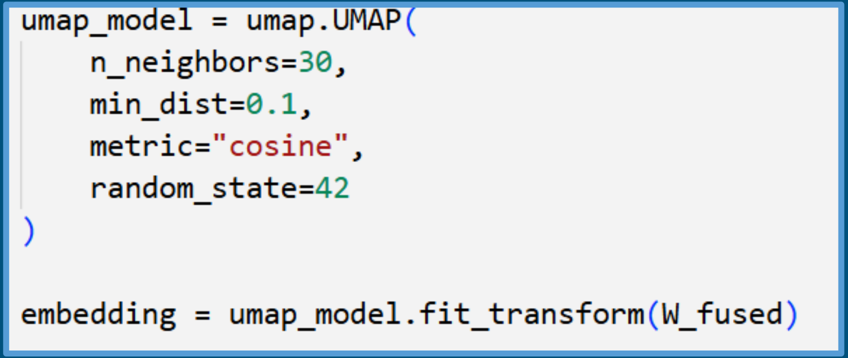
But we choose...

HDBSCAN & UMAP

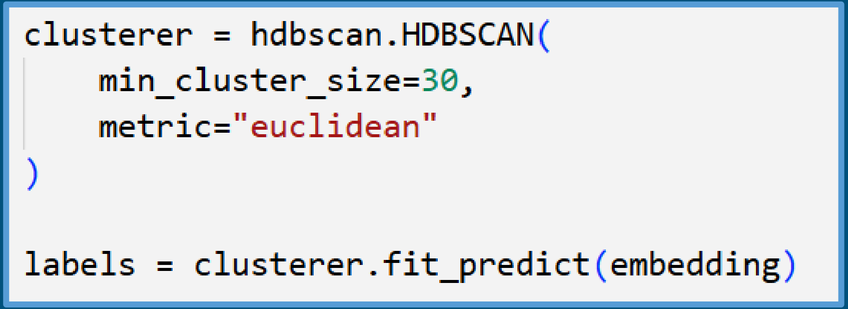
UMAP is used to project the SNF matrix into 2D while **preserving the local structure**.

HDBSCAN automatically **identifies dense clusters** and ***removes noise*** without requiring a predefined number of clusters.

The chosen parameters ensure compact, stable, and biologically consistent clusters.

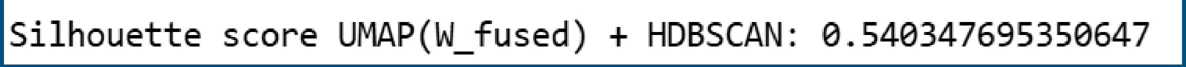


```
umap_model = umap.UMAP(
    n_neighbors=30,
    min_dist=0.1,
    metric="cosine",
    random_state=42
)

embedding = umap_model.fit_transform(W_fused)
```



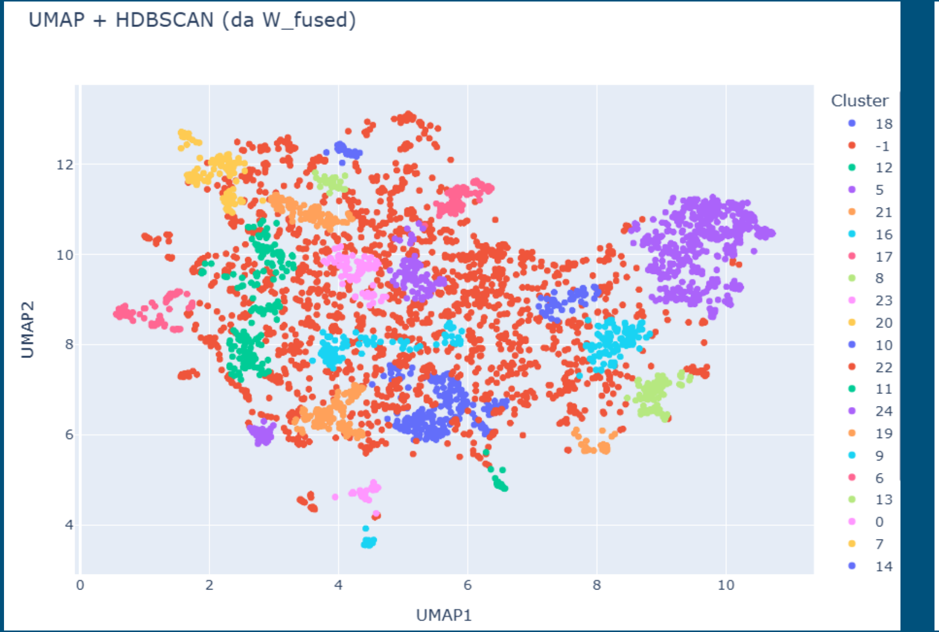
```
clusterer = hdbscan.HDBSCAN(
    min_cluster_size=30,
    metric="euclidean"
)

labels = clusterer.fit_predict(embedding)
```

FINAL CLUSTERS

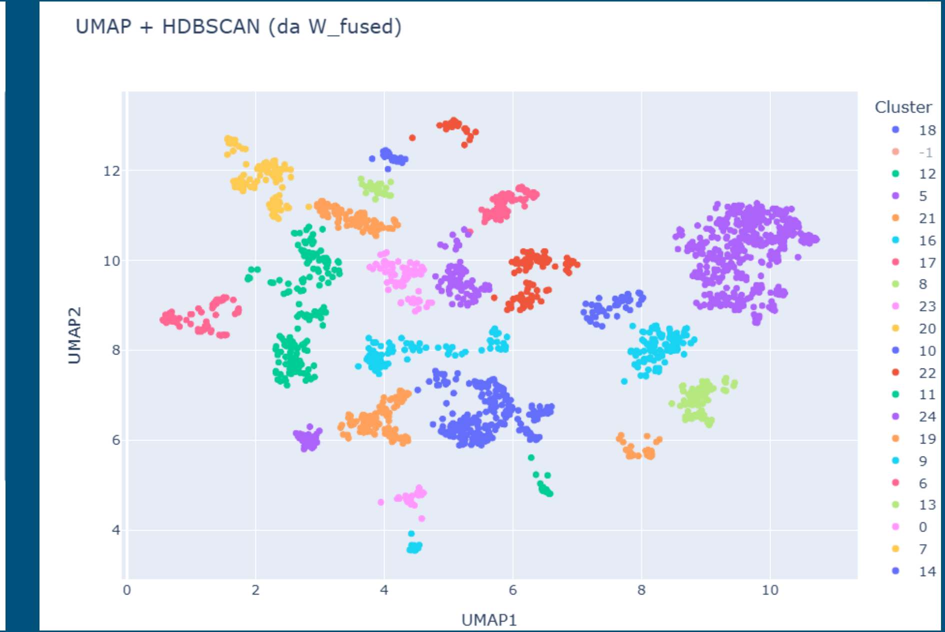


```
Silhouette score UMAP(W_fused) + HDBSCAN: 0.540347695350647
```

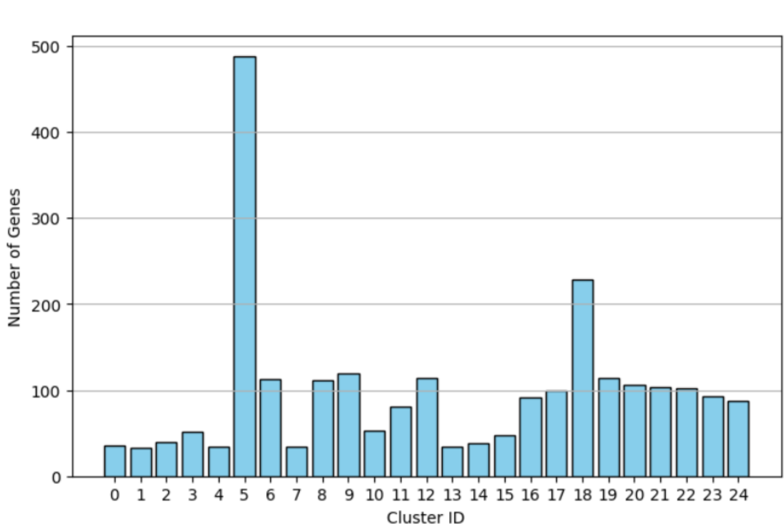
With noise



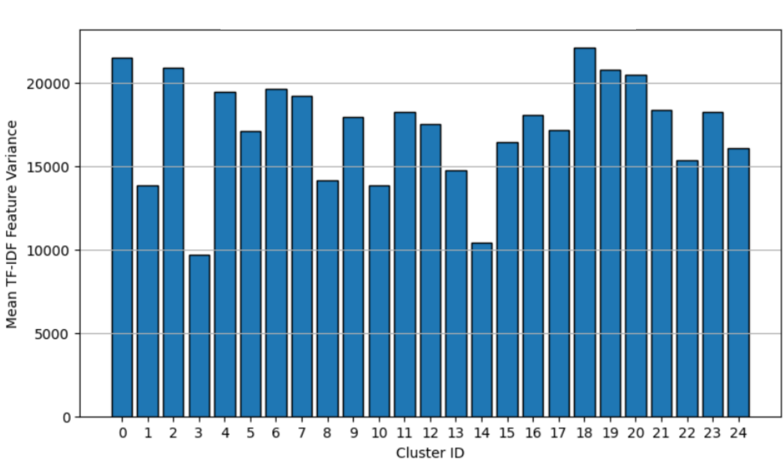
Without noise



HISTOGRAMS OF **CLUSTERS**

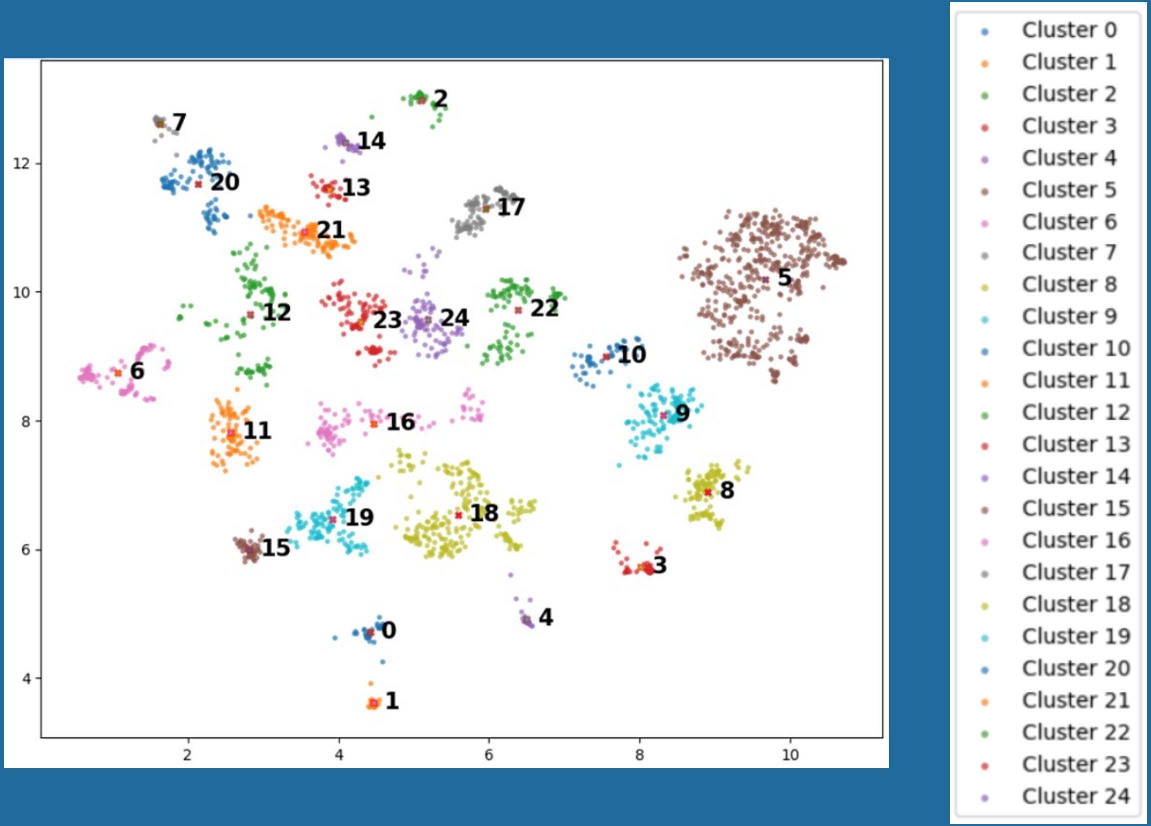


Cluster size distribution, without noise.



Variance of the clusters, without noise.

CENTROIDS OF THE CLUSTERS



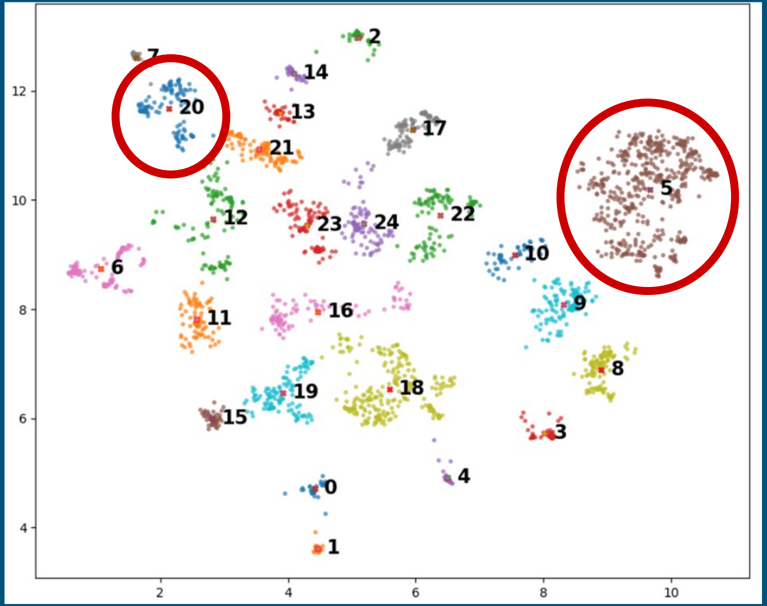
CENTROIDS

Cluster 5

Category	GO Term	Description
MF	GO:0140110	Transcription regulator activity
CC	GO:0000785	Centromeric region
BP	GO:0006357	Regulation of transcription by RNA Pol II

Cluster 20

Category	GO Term	Description
MF	GO:0005201	Extracellular matrix structural constituent
CC	GO:0031012	Extracellular matrix
BP	GO:0030198	Extracellular matrix organization



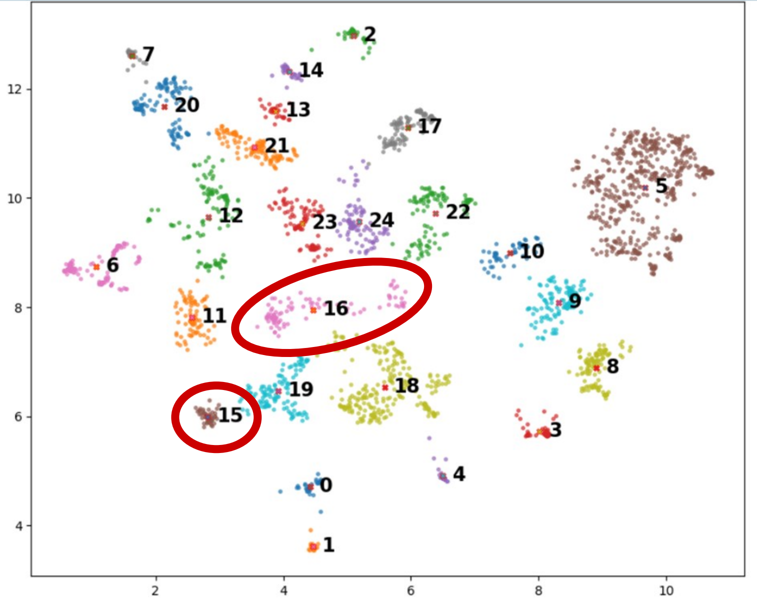
CENTROIDS

Cluster 15

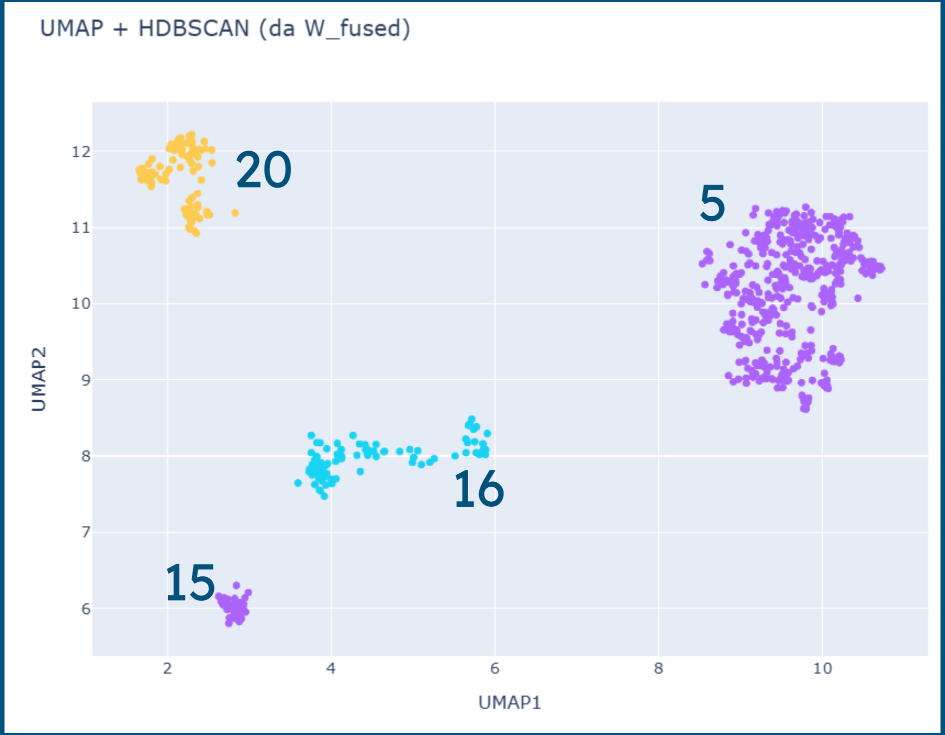
Category	GO Term	Description
MF	GO:0016757	Glycosyltransferase activity
CC	GO:0000139	Golgi membrane
BP	GO:0009100	Glycolipid biosynthetic process

Cluster 16

Category	GO Term	Description
MF	GO:0004553	Hydrolase activity (O-glycosyl compounds)
CC	GO:0005775	Vacuolar lumen
BP	GO:0005975	Carbohydrate metabolic process



CLUSTER ANALYSIS



5

These genes localize mainly to the **nucleus**, **chromosomes**, and **transcription-related complexes**, consistent with their **regulatory role in transcription**.

20

The localizations align with the BP and MF terms: **extracellular matrix, collagen, basement membrane, and cortical cytoskeleton**, all structures involved in **tissue support and adhesion**.

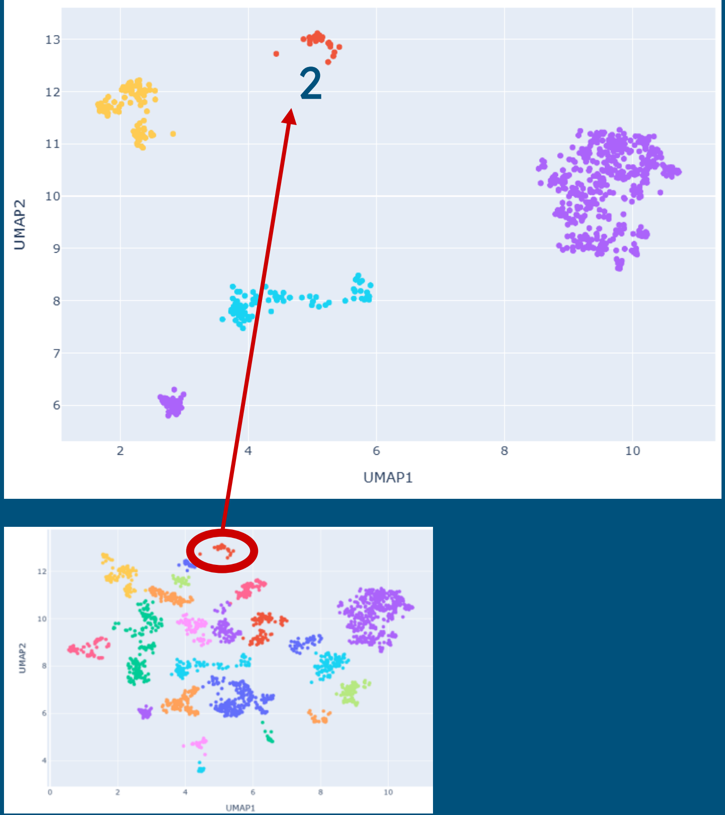
15

These genes are mainly located in the **Golgi apparatus** and related compartments, consistent with roles in **carbohydrate processing** and **protein glycosylation/transport**.

16

The genes localize to **vacuoles, lysosomes, and vesicles**, consistent with **metabolic and enzymatic functions** that process sugars within intracellular compartments.

CLUSTER 2



BP (Biological Process)

- **GO.0045109** → *regulation of receptor signaling pathway via JAK-STAT*
- **GO.0045104** → *positive regulation of somatic stem cell proliferation*
- **GO.0031424** → *keratinocyte differentiation*
- **GO.0030216** → *keratinocyte proliferation*
- **GO.0043588** → *skin development*

The cluster includes genes involved in **skin development** and **keratinocyte proliferation and differentiation**, along with **JAK–STAT signaling regulation** related to cell growth and differentiation.

MF (Molecular Function)

- **GO.0030280** → *potassium ion transmembrane transporter activity*
- **GO.0005198** → *structural molecule activity*
- **GO.0005200** → *structural constituent of cytoskeleton*
- **GO.0019215** → *transmembrane receptor protein tyrosine kinase activity*
- **GO.1990254** → *voltage-gated ion channel activity*

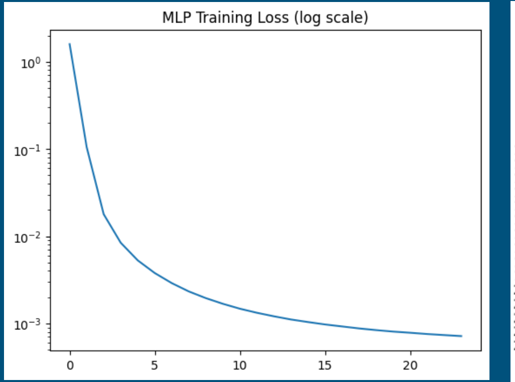
The molecular functions include **ion transport**, **cytoskeletal structural components**, and **tyrosine-kinase receptors**, consistent with the regulation of **keratinocyte proliferation and differentiation**.

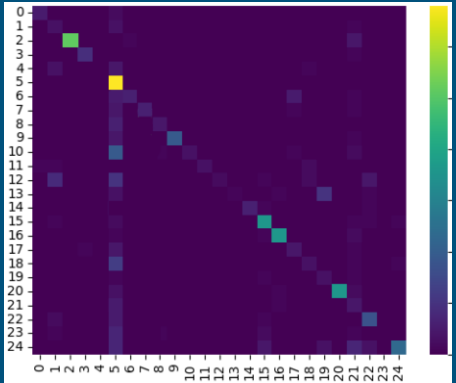
CC (Cellular Component)

- **GO.0005882** → *intermediate filament*
- **GO.0045111** → *intercellular junction*
- **GO.0045095** → *cell junction*
- **GO.0099512** → *transmembrane transporter complex*
- **GO.0001533** → *cornified envelope*

The genes localize to **intermediate filaments**, **cell junctions**, and **skin-related structures** (the **cornified envelope**), consistent with the BP and MF terms.

Training MLP





The clusters that the **MLP predicts well** are **clusters that are logically separated:**

ex. 5 15 16 20

```
accuracy 0.57
```

We can see that the training loss curve shows a rapid decline in the early epochs and stable convergence towards very low values.

low accuracy → presence of **very small clusters**: these clusters do not contain enough information to be learned by the model,

So **only predicts** the **most robust** clusters

THANKS

Porcelli Angelica - 78083A

Roveda Gianluca - 73814A

Stefanelli Marta - 84393A

AI
EXPERIENCE

codes errors

GO terms explanation