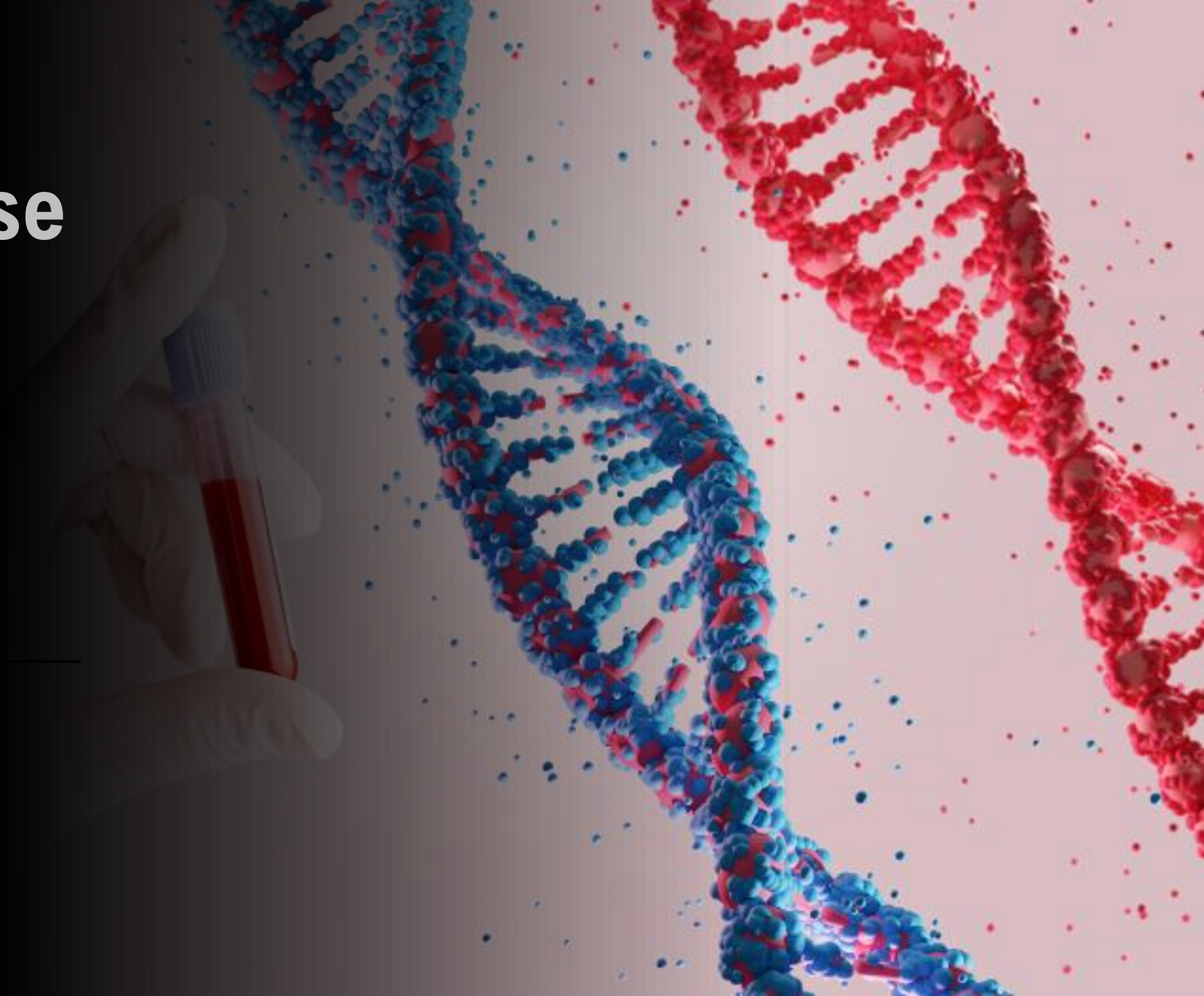
# Gene - Disease Association Prediction



Sistemas Inteligentes para a Bioinformática
Professor Miguel Rocha

Carlos Gomes (pg51681)
Diogo Esteves (pg2893)
Joana Lopes (pg53498)
Tiago Miranda (pg54437)

# Overview

PROBLEM

DATA EXPLORATION AND PRE-PROCESSING

MACHINE LEARNING TASK

DEEP LEARNING TASK

# Gene-Disease Association (GDA)

The study of **gene-disease associations is a crucial step in understanding the etiology of diseases.** Identifying the link between genes and diseases makes it possible to understand the disease's cause better and develop strategies to combat it. [1]

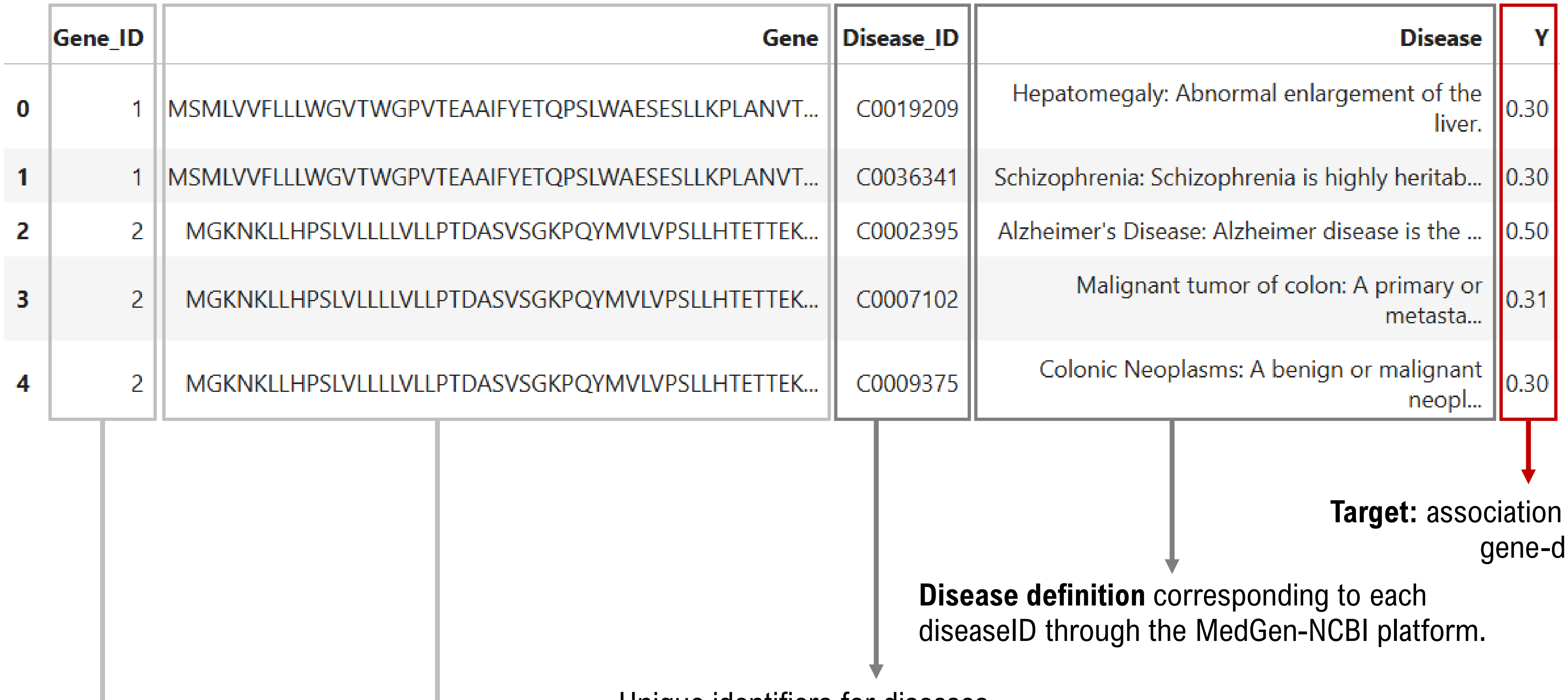
**GDA quantify the relation among a pair of gene-disease** and is one of the core concepts of **DisGeNET** platform that are integrated on TDC. DisGeNET is a discovery platform containing one of the largest publicly available collections of genes and variants associated to human diseases. [2]

Using **machine learning** accurately can contribute to **new discoveries** and bring numerous **therapeutic opportunities.**

## Task Description

**Regression.** Given the disease description and the amino acid sequence of the gene, predict their association.

# GDA dataset : 52,476 gene-disease pairs (7 399 unique genes, 7095 unique diseases)



| | Gene_ID | Gene | Disease_ID | Disease | Y |
|---|---|---|---|---|---|
| **0** | 1 | MSMLVVFLLLWGVTWGPVTEAAIFYETQPSLWAESESLLKPLANVT... | C0019209 | Hepatomegaly: Abnormal enlargement of the liver. | 0.30 |
| **1** | 1 | MSMLVVFLLLWGVTWGPVTEAAIFYETQPSLWAESESLLKPLANVT... | C0036341 | Schizophrenia: Schizophrenia is highly heritab... | 0.30 |
| **2** | 2 | MGKNKLLHPSLVLLLLVLLPTDASVSGKPQYMVLVPSLLHTETTEK... | C0002395 | Alzheimer's Disease: Alzheimer disease is the ... | 0.50 |
| **3** | 2 | MGKNKLLHPSLVLLLLVLLPTDASVSGKPQYMVLVPSLLHTETTEK... | C0007102 | Malignant tumor of colon: A primary or metasta... | 0.31 |
| **4** | 2 | MGKNKLLHPSLVLLLLVLLPTDASVSGKPQYMVLVPSLLHTETTEK... | C0009375 | Colonic Neoplasms: A benign or malignant neopl... | 0.30 |

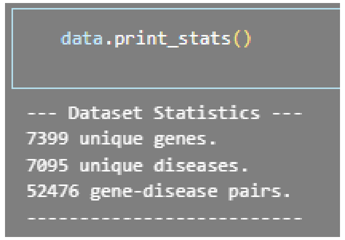
**Target:** association score for a gene-disease pair.

**Disease definition** corresponding to each diseaseID through the MedGen-NCBI platform.

Unique identifiers for diseases.

**Amino acids sequences** corresponding to each geneID through the Uniprot platform.

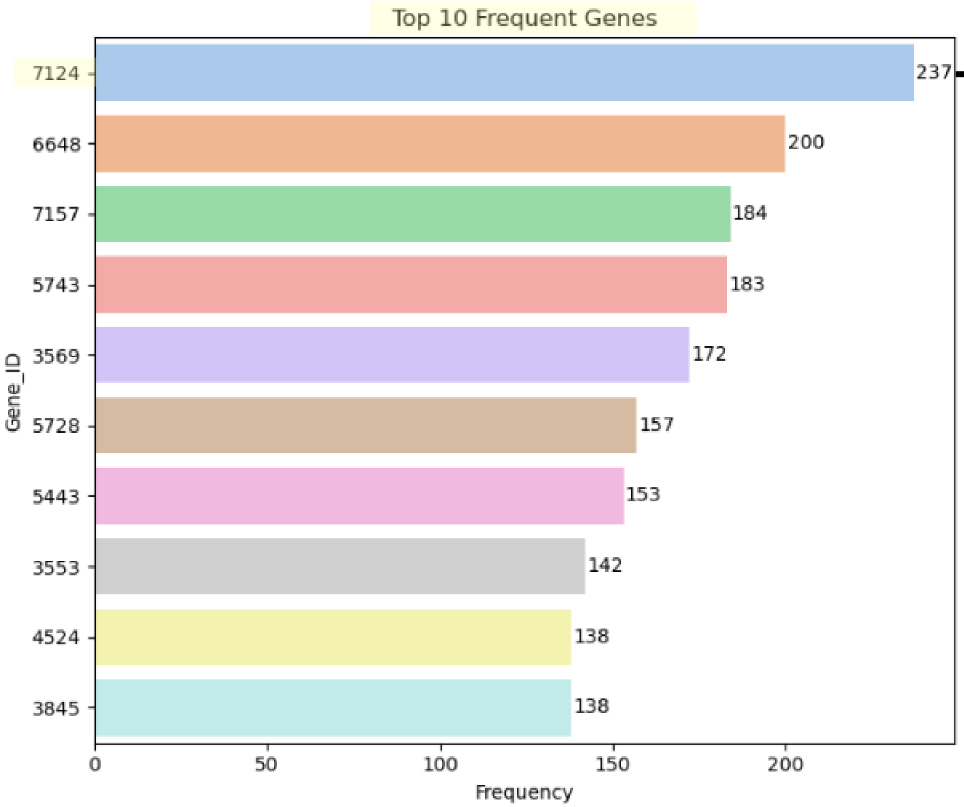
Unique identifiers for genes.

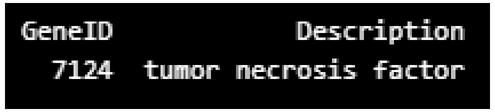


```
data.print_stats()
```

```
--- Dataset Statistics ---
7399 unique genes.
7095 unique diseases.
52476 gene-disease pairs.
--------------------------
```

# **Exploratory analysis:** gene and disease frequency



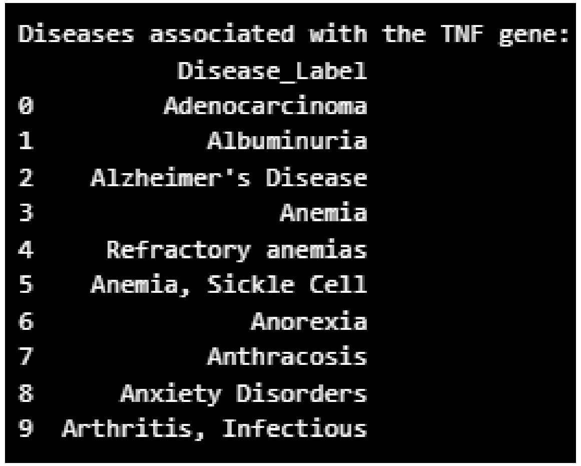


```
GeneID          Description
  7124   tumor necrosis factor
```

This gene encodes a **multifunctional proinflammatory cytokine** that belongs to the **Tumor Necrosis Factor** superfamily.

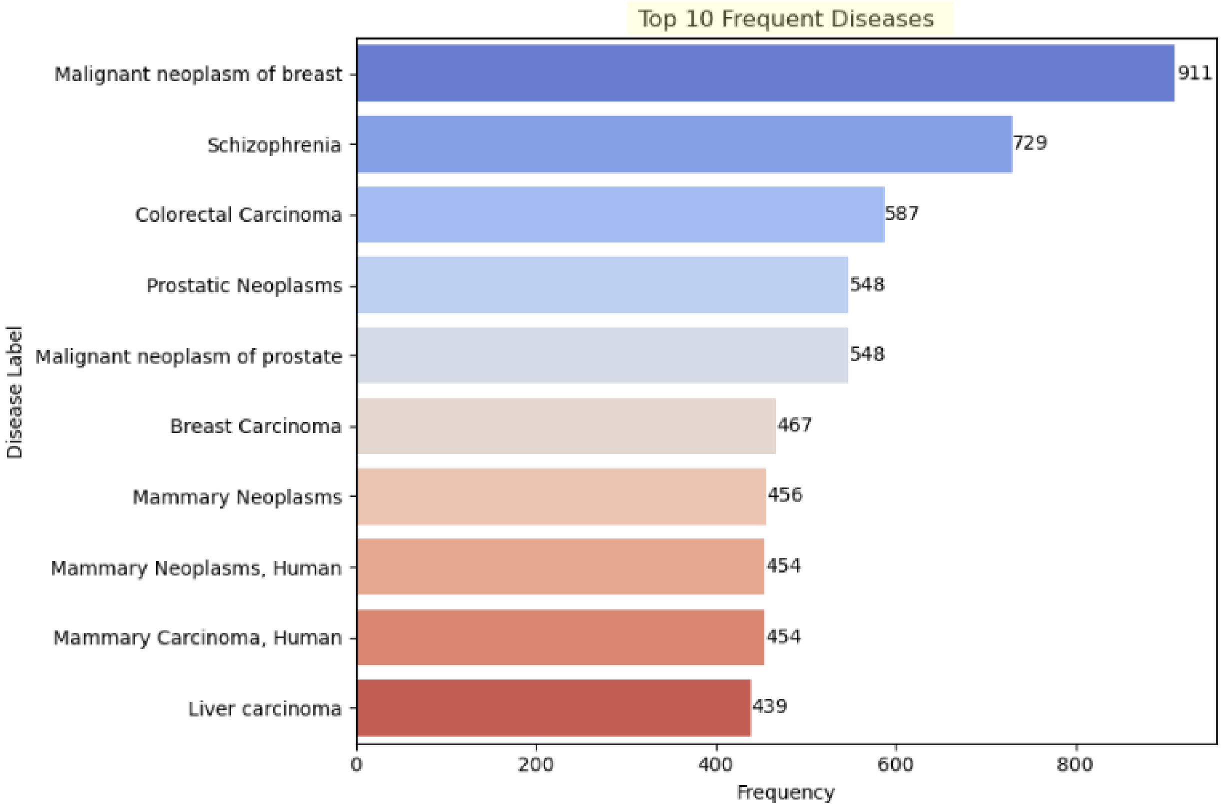
This cytokine is involved in the **regulation of a wide spectrum of biological processes** and has been **implicated in a variety of diseases**. [3]

These results indicate that there are **genes with greater involvement in multiple pathological conditions**. The high frequency of these genes may be a result of their **participation in fundamental biological processes**, such as cell cycle regulation, apoptosis or immune response.



```
Diseases associated with the TNF gene:
          Disease_Label
0         Adenocarcinoma
1            Albuminuria
2    Alzheimer's Disease
3                 Anemia
4     Refractory anemias
5    Anemia, Sickle Cell
6               Anorexia
7            Anthracosis
8      Anxiety Disorders
9  Arthritis, Infectious
```

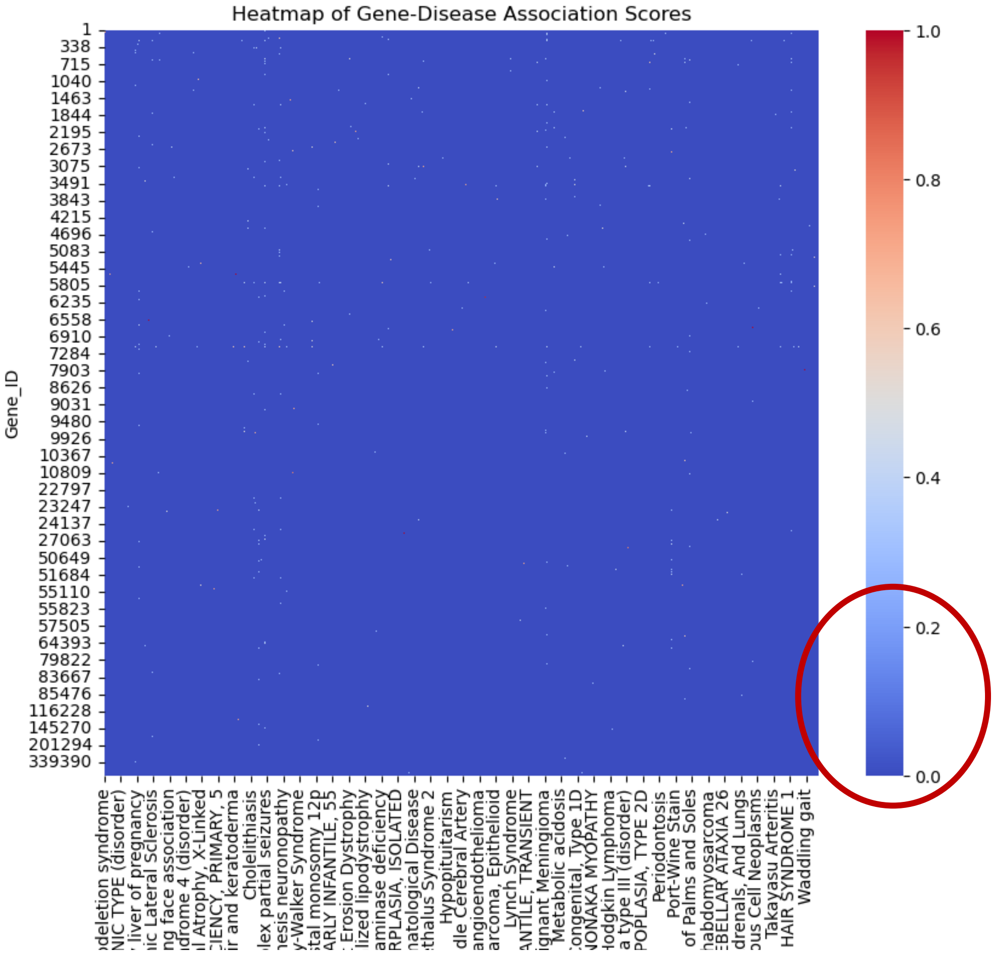
# **Exploratory analysis:** gene and disease frequency

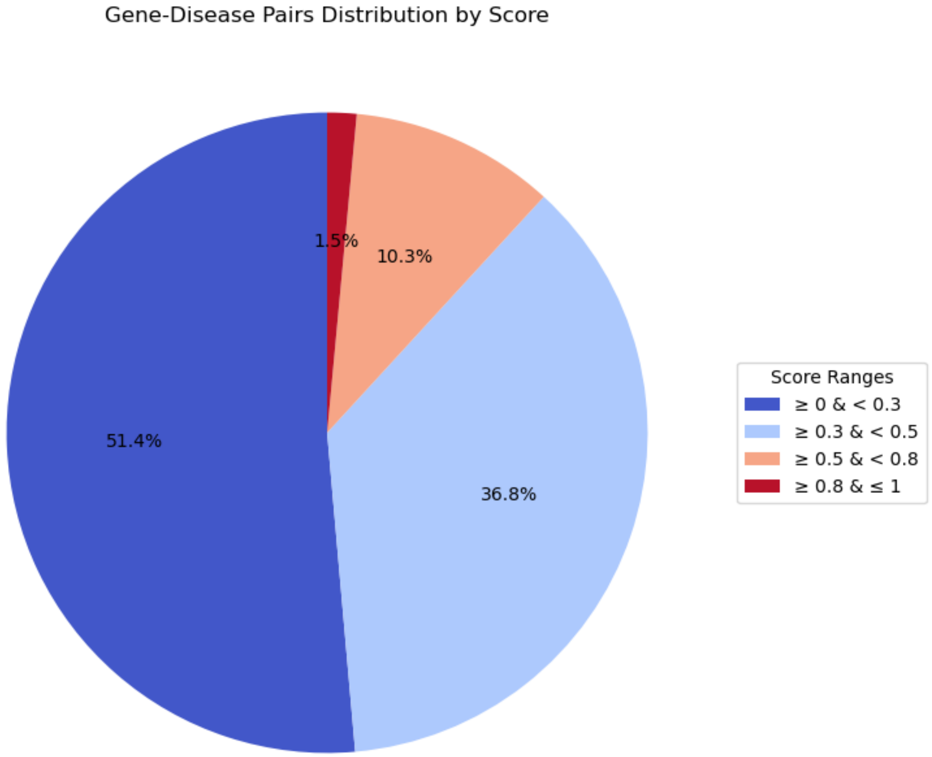


There is a clear pattern of highly frequent diseases in the dataset, led by conditions such as **malignant neoplasm of the breast** with 911 associations and **schizophrenia** with 729 associations.

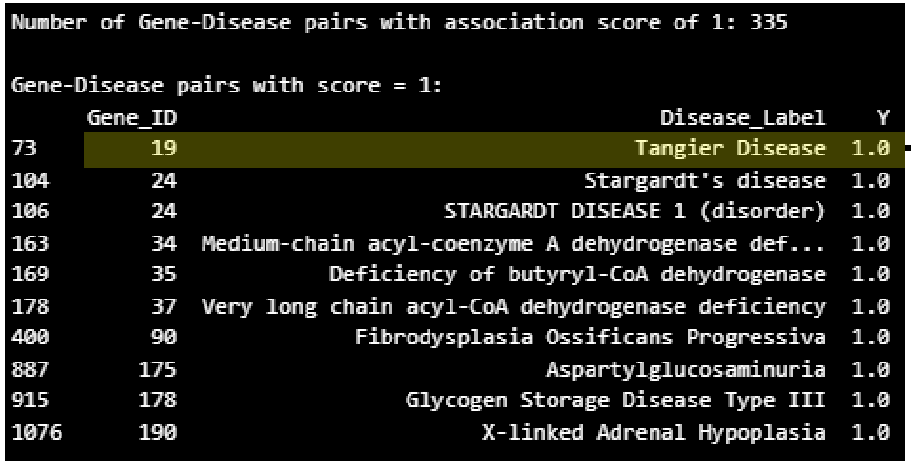
The list is dominated by **neoplasms** such as breast, prostate and liver cancer, as well as the condition schizophrenia.

# **Exploratory analysis:** pairs gene-disease score



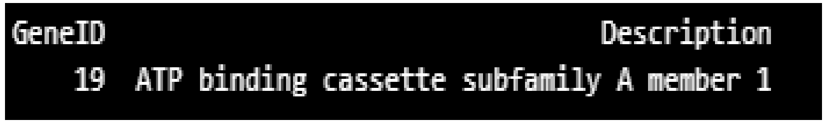


# **Exploratory analysis:** pairs gene-disease score



```
Number of Gene-Disease pairs with association score of 1: 335

Gene-Disease pairs with score = 1:
      Gene_ID                                   Disease_Label    Y
73         19                                Tangier Disease  1.0
104        24                             Stargardt's disease  1.0
106        24                     STARGARDT DISEASE 1 (disorder)  1.0
163        34  Medium-chain acyl-coenzyme A dehydrogenase def...  1.0
169        35             Deficiency of butyryl-CoA dehydrogenase  1.0
178        37  Very long chain acyl-CoA dehydrogenase deficiency  1.0
400        90                    Fibrodysplasia Ossificans Progressiva  1.0
887       175                          Aspartylglucosaminuria  1.0
915       178                     Glycogen Storage Disease Type III  1.0
1076      190                     X-linked Adrenal Hypoplasia  1.0
```



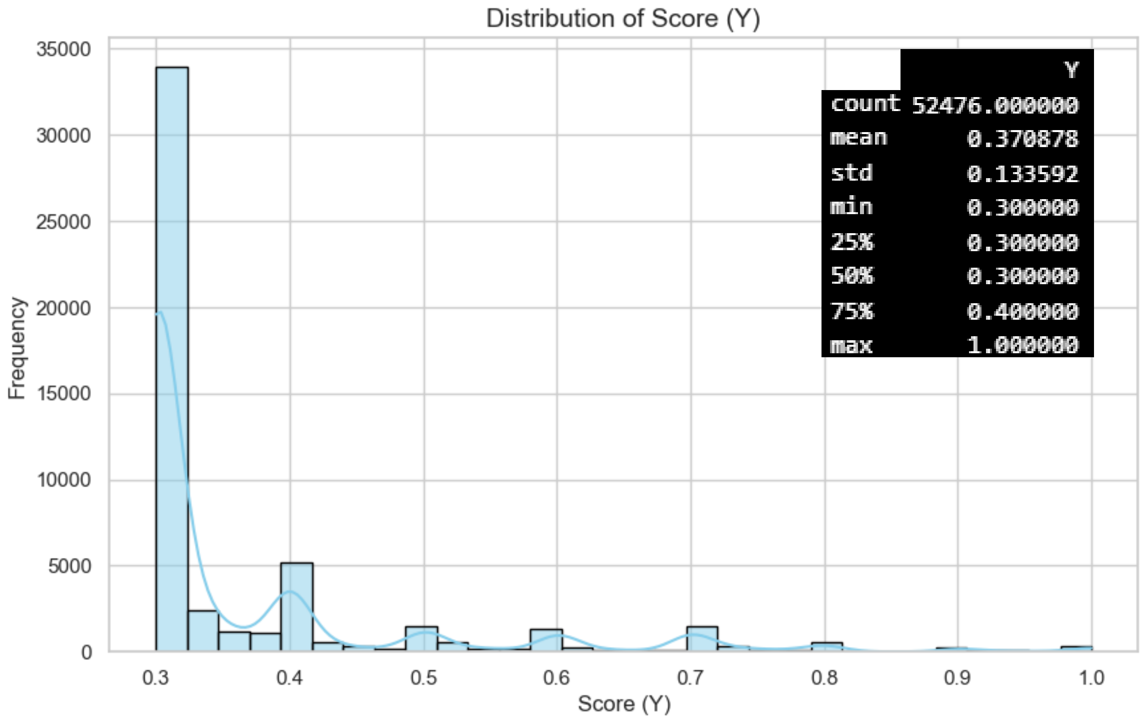
```
GeneID                              Description
    19  ATP binding cassette subfamily A member 1
```

The membrane-associated protein encoded by this gene is a member of the **superfamily of ATP-binding cassette (ABC) transporters.** [4]

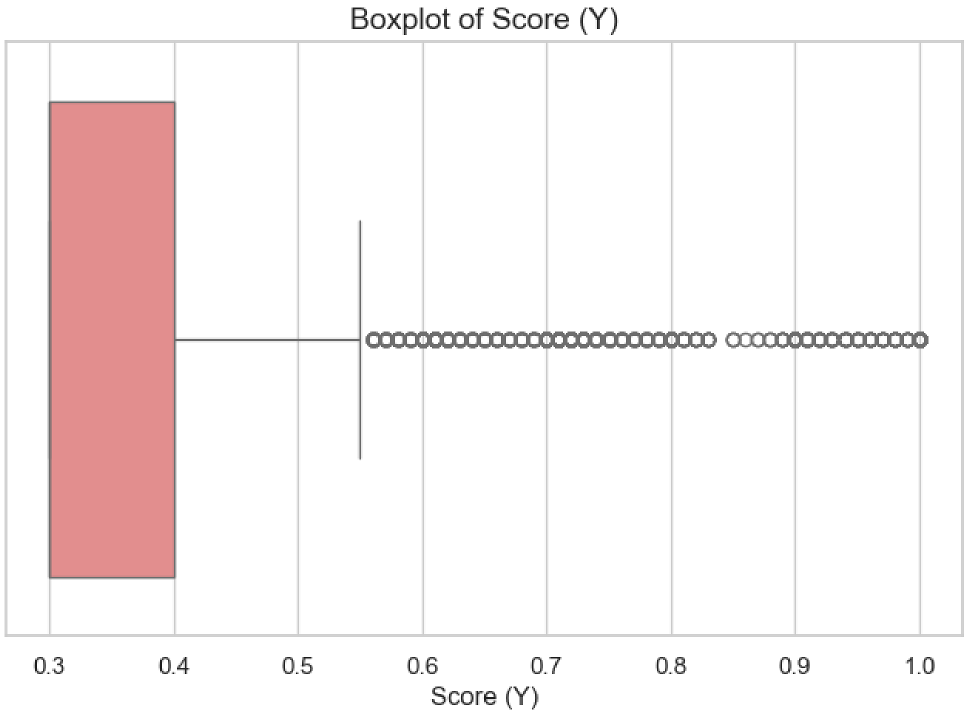
With cholesterol as its substrate, this protein functions as a **cholesterol efflux pump in the cellular lipid removal pathway**. Mutations in both alleles of this gene cause **Tangier disease** and **familial high-density lipoprotein (HDL) deficiency.** [5]

Highly associated pairs can **guide drug development**, help **identify biomarkers** for **early diagnosis** and contribute to **personalizing treatments**.

# **Exploratory analysis:** Y distribution



```
                  Y
count  52476.000000
mean       0.370878
std        0.133592
min        0.300000
25%        0.300000
50%        0.300000
75%        0.400000
max        1.000000
```



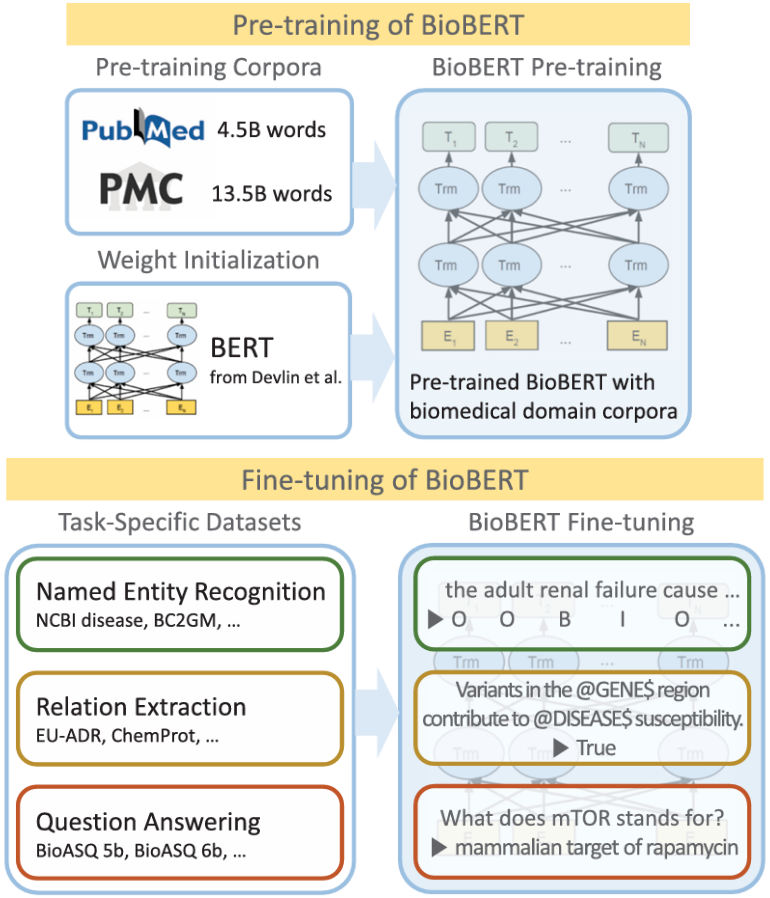
The target variable does not follow a normal distribution, with a clear **concentration of pairs in scores between 0.3 and 0.4**, indicate a low strength of association.

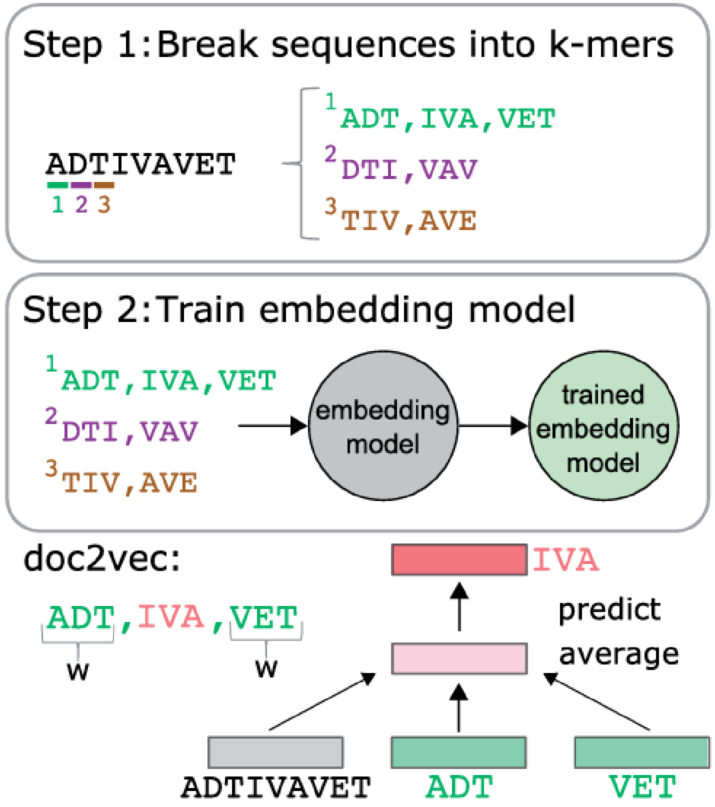
Based on the analysis of outliers using the interquartile range, **5475 outliers** were identified. In the context of our problem, these outliers represent gene-disease pairs where **a given gene contributes significantly to a given disease**.

# Pre-processing task

- ✓ Transformation of the "Disease" Column (Textual Descriptions of Diseases): generate embeddings through **BioBERT** model [6]

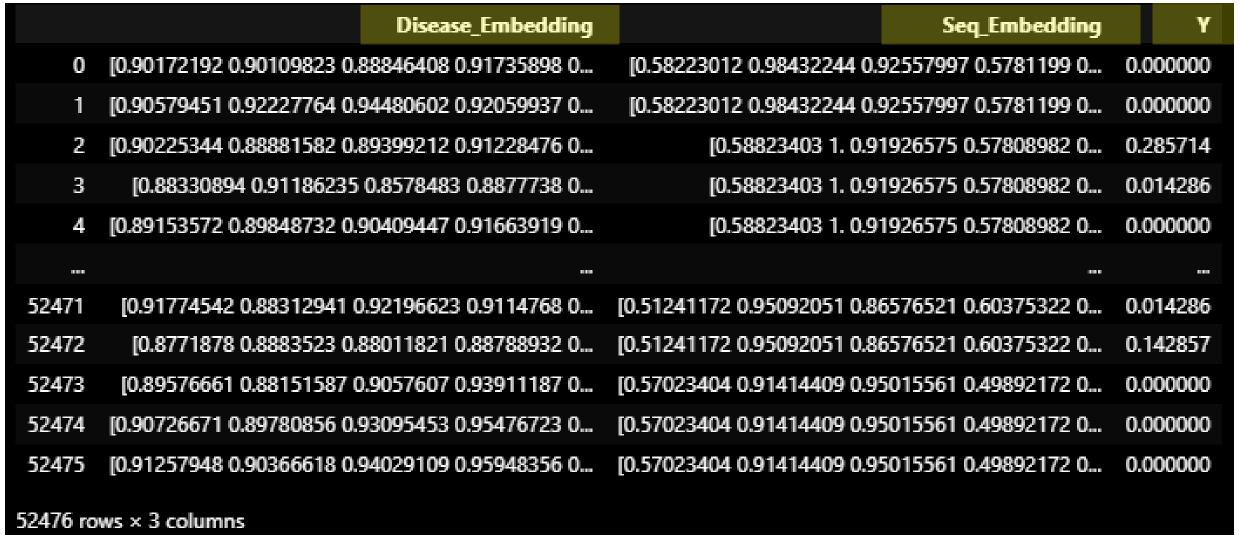
- ✓ Transformation of the "Gene" Column (Amino Acid Sequences): generate embeddings through **k-mers + Word2Vec** [7]





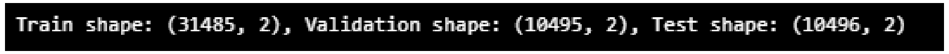
# Pre-processing task

- ✓ Conversion of the **Disease_Embedding** and **Seq_Embedding** columns from strings **to numerical arrays.**
- ✓ Feature integration: Combining generated features into a **unified dataset.**
- ✓ Data cleaning and normalization: **Handling missing values** and **scaling numerical features**.
- ✓ **Dataset splitting** dividing the data into training (60%), validation (20%), and test (20%) sets.

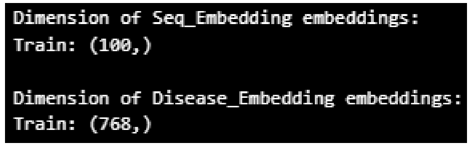


|  | Disease_Embedding | Seq_Embedding | Y |
|---|---|---|---|
| 0 | [0.90172192 0.90109823 0.88846408 0.91735898 0... | [0.58223012 0.98432244 0.92557997 0.5781199 0... | 0.000000 |
| 1 | [0.90579451 0.92227764 0.94480602 0.92059937 0... | [0.58223012 0.98432244 0.92557997 0.5781199 0... | 0.000000 |
| 2 | [0.90225344 0.88881582 0.89399212 0.91228476 0... | [0.58823403 1. 0.91926575 0.57808982 0... | 0.285714 |
| 3 | [0.88330894 0.91186235 0.8578483 0.8877738 0... | [0.58823403 1. 0.91926575 0.57808982 0... | 0.014286 |
| 4 | [0.89153572 0.89848732 0.90409447 0.91663919 0... | [0.58823403 1. 0.91926575 0.57808982 0... | 0.000000 |
| ... | ... | ... | ... |
| 52471 | [0.91774542 0.88312941 0.92196623 0.9114768 0... | [0.51241172 0.95092051 0.86576521 0.60375322 0... | 0.014286 |
| 52472 | [0.8771878 0.8883523 0.88011821 0.88788932 0... | [0.51241172 0.95092051 0.86576521 0.60375322 0... | 0.142857 |
| 52473 | [0.89576661 0.88151587 0.9057607 0.93911187 0... | [0.57023404 0.91414409 0.95015561 0.49892172 0... | 0.000000 |
| 52474 | [0.90726671 0.89780856 0.93095453 0.95476723 0... | [0.57023404 0.91414409 0.95015561 0.49892172 0... | 0.000000 |
| 52475 | [0.91257948 0.90366618 0.94029109 0.95948356 0... | [0.57023404 0.91414409 0.95015561 0.49892172 0... | 0.000000 |

52476 rows × 3 columns



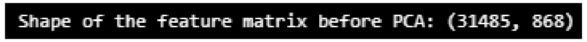
```
Train shape: (31485, 2), Validation shape: (10495, 2), Test shape: (10496, 2)
```



```
Dimension of Seq_Embedding embeddings:
Train: (100,)

Dimension of Disease_Embedding embeddings:
Train: (768,)
```

**Concatenating features** →

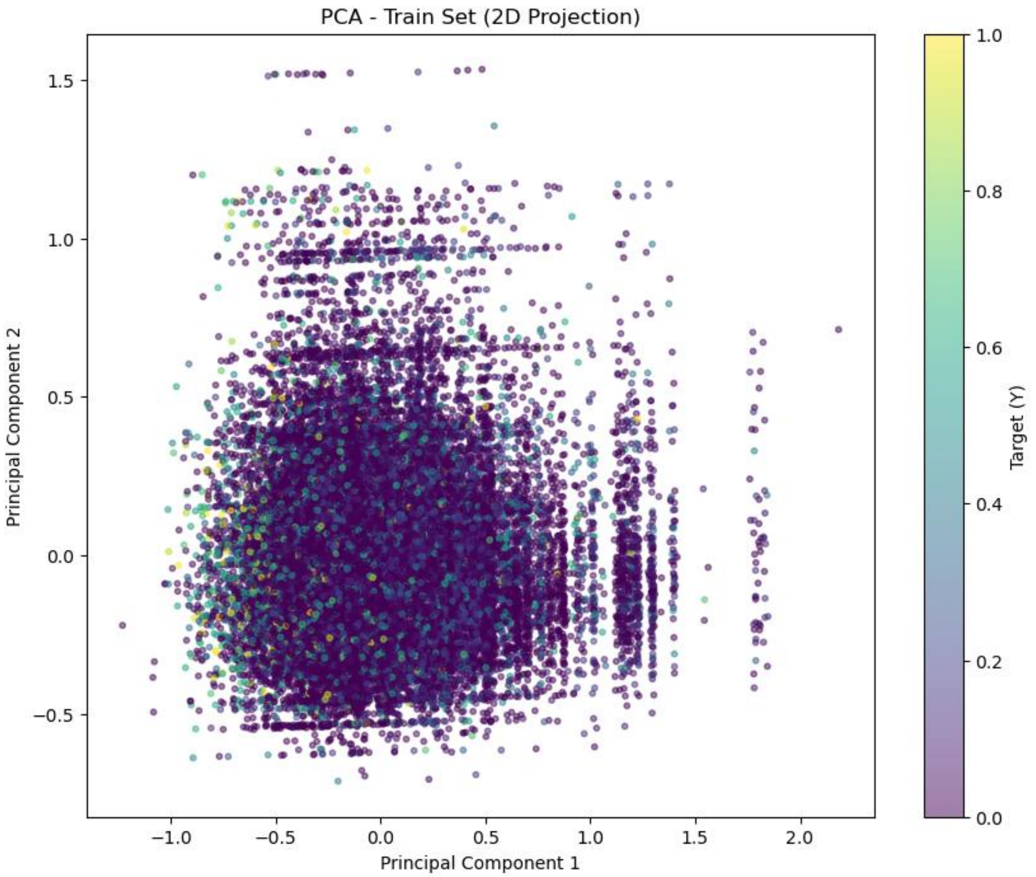


```
Shape of the feature matrix before PCA: (31485, 868)
```

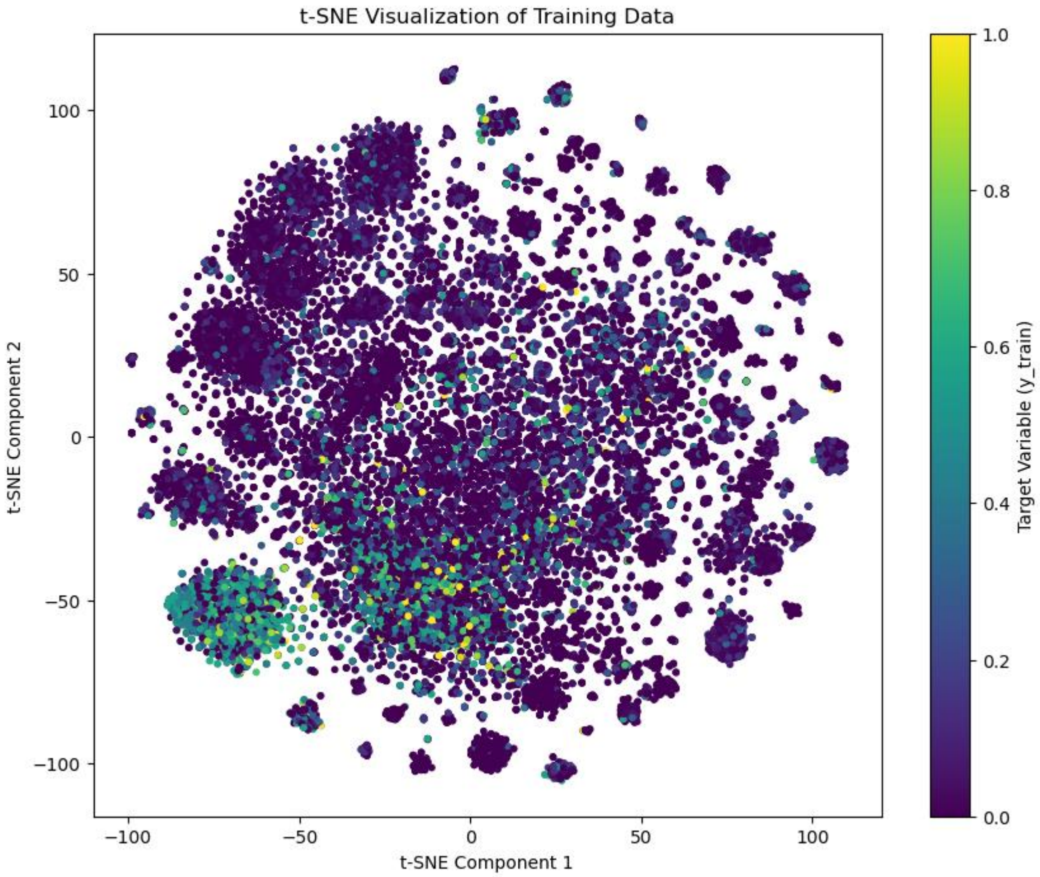
MACHINE LEARNING

# Unsupervised Learning

- After applying **PCA**, the original embeddings with 868 dimensions were reduced to **181 principal components**, preserving **95% of the explained variance.**



- **T-SNE** proved to be a **better choice for visualizing** the data.



# **Machine Learning task:** main results

| | Linear Regression | | Ridge Regression | | | DecisionTreeRegressor | | | RandomForestRegressor | | | Suport Vector Regressor | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Hyperparameters** | Def. | Def. | Def. | Def. | Randomized SearchCV | Def. | Def. | Randomized SearchCV | Def. | Def. | **Randomized SearchCV** | Def. | Def. | Randomized SearchCV |
| **Reduced data by PCA** | Yes | No | Yes | No | No | Yes | No | No | Yes | No | **No** | Yes | No | No |
| **MSE** | 0,028 | 0,025 | 0,028 | 0,025 | 0,025 | 0,056 | 0,044 | 0,281 | 0,026 | 0,023 | **0,023** | 0,026 | 0,028 | 0,025 |
| **R-squared** | 0,229 | 0,306 | 0,229 | 0,306 | 0,311 | -0,561 | - 0,239 | 0,214 | 0,274 | 0,353 | **0,356** | 0,284 | 0,213 | 0,294 |

- ✓ Better performance of the models when using the **original data** (without applying PCA);

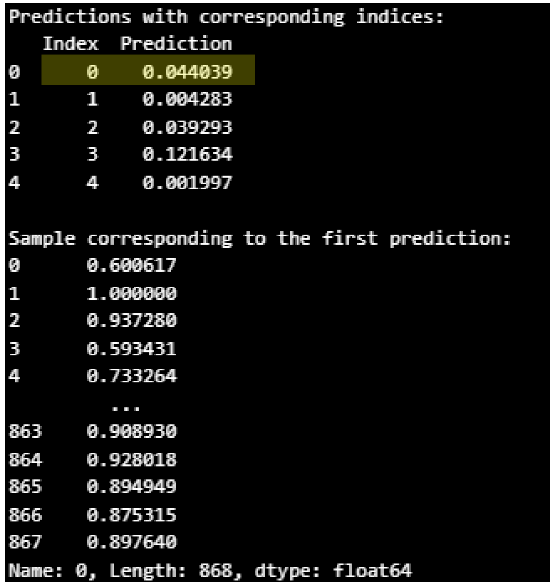
- ✓ Better performance of the models when using **RandomizedSearchCV for hiperparameter optimization**.

**Best hyperparameters founded**
n_estimators: 150
min_samples_splitt: 5
min_samples_leaf: 2
max_features: sqrt
max_depth: 30
bootstrap: True

**Most promising model**

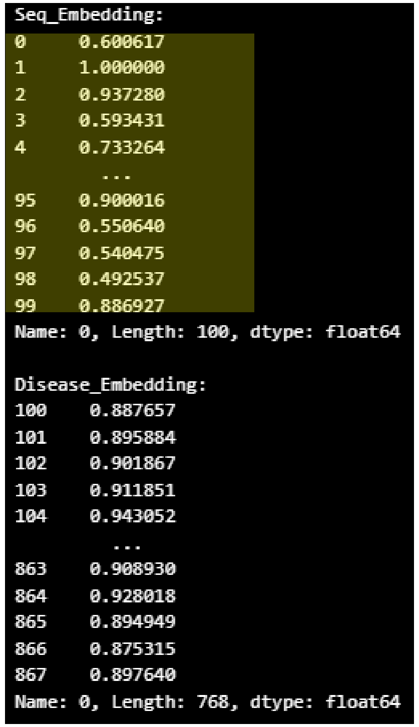
# **Machine Learning task:** main results

| | RFR |
|---|---|
| Hiperparameters | Randomized SearchCV |
| Reduced data by PCA | No |
| MSE | 0,023 |
| R-squared | 0,357 |



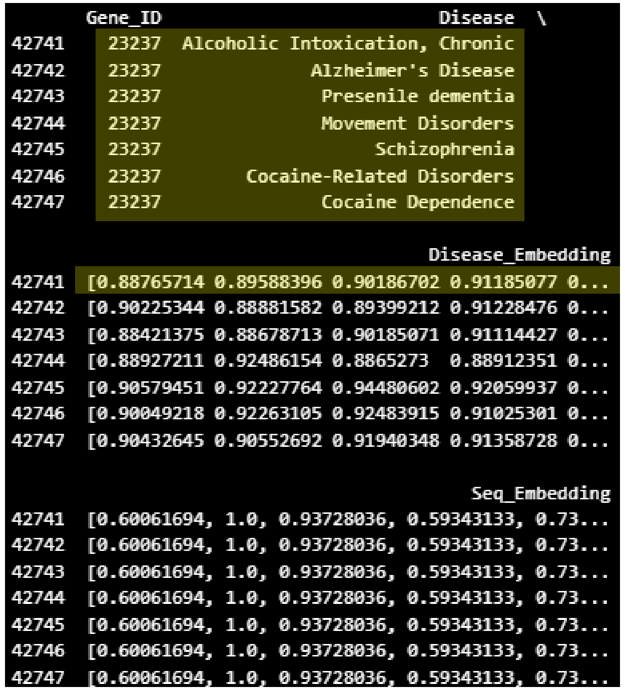
```
Predictions with corresponding indices:
   Index  Prediction
0      0    0.044039
1      1    0.004283
2      2    0.039293
3      3    0.121634
4      4    0.001997

Sample corresponding to the first prediction:
0      0.600617
1      1.000000
2      0.937280
3      0.593431
4      0.733264
         ...
863    0.908930
864    0.928018
865    0.894949
866    0.875315
867    0.897640
Name: 0, Length: 868, dtype: float64
```



```
Seq_Embedding:
0      0.600617
1      1.000000
2      0.937280
3      0.593431
4      0.733264
         ...
95     0.900016
96     0.550640
97     0.540475
98     0.492537
99     0.886927
Name: 0, Length: 100, dtype: float64

Disease_Embedding:
100     0.887657
101     0.895884
102     0.901867
103     0.911851
104     0.943052
          ...
863     0.908930
864     0.928018
865     0.894949
866     0.875315
867     0.897640
Name: 0, Length: 768, dtype: float64
```



```
       Gene_ID                          Disease  \
42741    23237  Alcoholic Intoxication, Chronic
42742    23237              Alzheimer's Disease
42743    23237               Presenile dementia
42744    23237               Movement Disorders
42745    23237                    Schizophrenia
42746    23237         Cocaine-Related Disorders
42747    23237               Cocaine Dependence

                                        Disease_Embedding
42741  [0.88765714 0.89588396 0.90186702 0.91185077 0...
42742  [0.90225344 0.88881582 0.89399212 0.91228476 0...
42743  [0.88421375 0.88678713 0.90185071 0.91114427 0...
42744  [0.88927211 0.92486154 0.8865273  0.88912351 0...
42745  [0.90579451 0.92227764 0.94480602 0.92059937 0...
42746  [0.90049218 0.92263105 0.92483915 0.91025301 0...
42747  [0.90432645 0.90552692 0.91940348 0.91358728 0...

                                            Seq_Embedding
42741  [0.60061694, 1.0, 0.93728036, 0.59343133, 0.73...
42742  [0.60061694, 1.0, 0.93728036, 0.59343133, 0.73...
42743  [0.60061694, 1.0, 0.93728036, 0.59343133, 0.73...
42744  [0.60061694, 1.0, 0.93728036, 0.59343133, 0.73...
42745  [0.60061694, 1.0, 0.93728036, 0.59343133, 0.73...
42746  [0.60061694, 1.0, 0.93728036, 0.59343133, 0.73...
42747  [0.60061694, 1.0, 0.93728036, 0.59343133, 0.73...
```

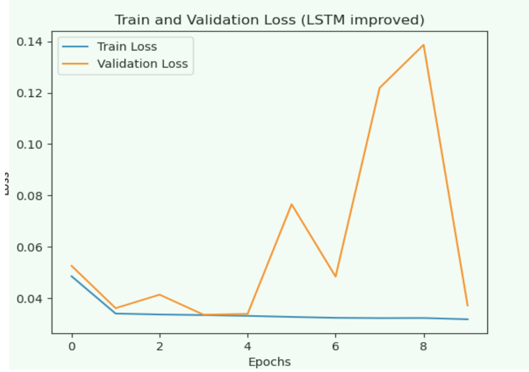
✓ **Gene_ID:** 23237

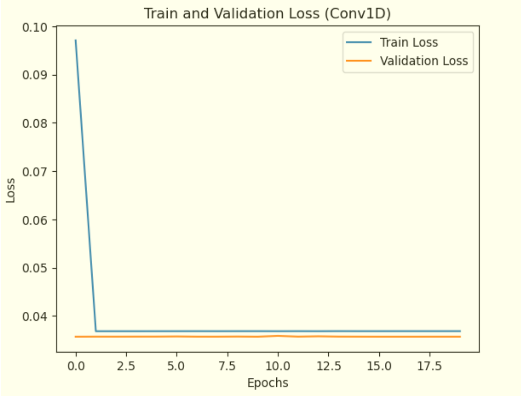
Activity regulated cytoskeleton associated protein [8]

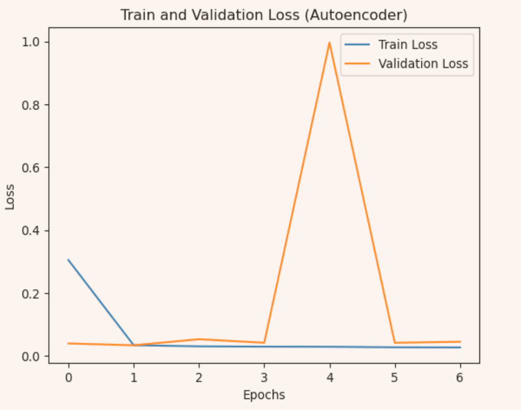
Potential role in **vulnerability to addiction** [9]

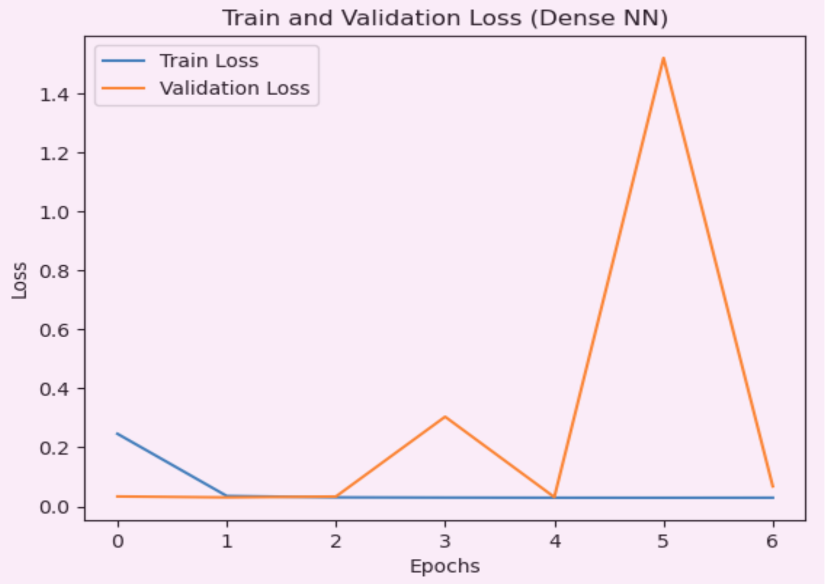
Targeting Acr may represent a new mechanism for **preventing relapses related to chronic alcoholism** [10]

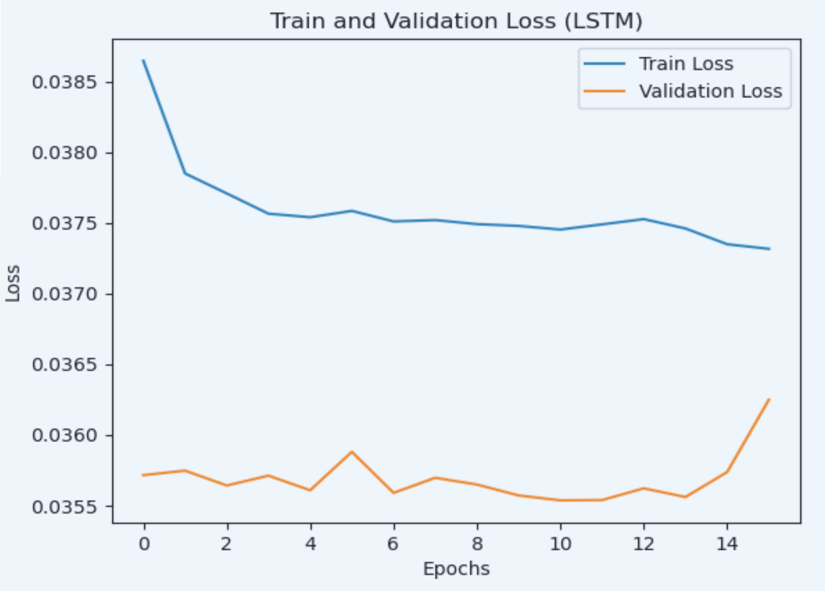
# **Deep Learning task :** model metrics


Train and Validation Loss (LSTM improved)


Train and Validation Loss (Conv1D)


Train and Validation Loss (Autoencoder)


Train and Validation Loss (Dense NN)


Train and Validation Loss (LSTM)

| | **Dense NN** | **LSTM (v3)** | **LSTM Improved** | **Autoencoder** | **Conv1D** |
|---|---|---|---|---|---|
| **Validation Loss (MSE)** | **0.02** | 0.035 | 0.04 | 0.05 | 0.03 |
| **Train Loss (MSE)** | **0.015** | 0.033 | 0.037 | 0.045 | 0.02 |
| **Overfitting (val_loss - train_loss)** | **0.005** | 0.002 | 0.003 | 0.005 | 0.01 |
| **Generalization** | **High** | Moderate | Moderate | Low | High |
| **Stability** | **Stable** | Slight Oscillations | Unstable | Unstable | Stable |
| **Validation Loss (MSE)** | **0.02** | 0.035 | 0.04 | 0.05 | 0.03 |

- ✓ Better performances of the trained models;

# **Deep Learning task :** model evaluation

| | Dense NN | Conv1D | LSTM (v3) | LSTM (improved) | Autoencoder |
|---|---|---|---|---|---|
| **Main Characteristics** | General-purpose feedforward network, suited for non-sequential data. | Specialized for sequential data, effective for local patterns in time series. | Sequential network, capable of capturing long-term dependencies. | Incorporates Batch Normalization for more stable learning. | Learns compressed representations of data for reconstruction tasks. |
| **Architecture** | 3 Dense layers (256 → 128 → 64) with Batch Normalization and Dropout. | 2 Conv1D layers (64, 128 filters), MaxPooling1D, Flatten layer. | 2 LSTM layers (32 → 16 units), Dense output layer. | 2 LSTM layers with BatchNorm, followed by Dense (8 units). | Encoder (Dense 128 → 64), Decoder (Dense 128 → original dim). |
| **Regularization** | Dropout (30%/20%) | Dropout (30%) | Dropout (20%/10%) | Dropout (20%/10%) | BatchNorm, Sigmoid output |
| **Validation Loss** | 0.02 | 0.03 | 0.035 | 0.04 | 0.05 |
| **Best Features** | Effective for non-sequential data like embeddings | Ideal for sequential features in embeddings | Best of the 4 models for capturing gene-disease temporal relationships | Improves temporal dependencies with added stability | Best for dimensionality reduction in large datasets |
| **Notes** | Performs well with tabular data from gene-disease associations, good generalization | Stable loss; captures local patterns in sequential embeddings | Handles long-term dependencies in sequential embeddings | Overfitting observed; Used BatchNorm to help stabilize learning | Good for compressing embeddings but struggles with generalization |

- ✓ Better performance – Dense NN, Conv1D
- ✓ High Potential – LSTM, improved LTSM

# Future Work

- Final notebook adjustments.
- Optimize DL models, exploring a multimodal approach.

Thank you for your attention.

# References

- [1] Opap, K., & Mulder, N. (2017). Recent advances in predicting gene–disease associations. F1000Research, 6, 578. https://doi.org/10.12688/f1000research.10788.1
- [2] Therapeutics Data Commons. (2024). TDC. https://tdcommons.ai/
- [3] TNF tumor necrosis factor [Homo sapiens (human)] - Gene - NCBI. (n.d.). Www.ncbi.nlm.nih.gov. https://www.ncbi.nlm.nih.gov/gene/7124
- [4] ABCA1 ATP binding cassette subfamily A member 1 [Homo sapiens (human)] - Gene - NCBI. (2020). Nih.gov. https://www.ncbi.nlm.nih.gov/gene/19
- [5] Koseki, M., Yamashita, S., Ogura, M., Ishigaki, Y., Ono, K., Tsukamoto, K., Hori, M., Matsuki, K., Yokoyama, S., & Harada-Shiba, M. (2021). Current Diagnosis and Management of Tangier Disease. *Journal of Atherosclerosis and Thrombosis*, *28*(8), 802–810. https://doi.org/10.5551/jat.rv17053
- [6] Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H., & Kang, J. (2019). BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, *36*(4). https://doi.org/10.1093/bioinformatics/btz682
- [7] Yang, K. K., Wu, Z., Bedbrook, C. N., & Arnold, F. H. (2018). Learned protein embeddings for machine learning. *Bioinformatics*, *34*(15), 2642–2648. https://doi.org/10.1093/bioinformatics/bty178
- [8] *ARC activity regulated cytoskeleton associated protein [Homo sapiens (human)] - Gene - NCBI*. (2025). Nih.gov. https://www.ncbi.nlm.nih.gov/gene/23237
- [9] Penrod, R. D., Thomsen, M., Taniguchi, M., Guo, Y., Cowan, C. W., & Smith, L. N. (2020). The activity-regulated cytoskeleton-associated protein, Arc/Arg3.1, influences mouse cocaine self-administration. Pharmacology, Biochemistry, and Behavior, 188, 172818. https://doi.org/10.1016/j.pbb.2019.172818
- [10] Pagano, R., Salamian, A., Zielinski, J., Beroun, A., Nalberczak-Skóra, M., Skonieczna, E., Cały, A., Tay, N., Banaschewski, T., Desrivières, S., Grigis, A., Garavan, H., Heinz, A., Brühl, R., Martinot, J.-L., Martinot, M.-L. P., Artiges, E., Nees, F., Orfanos, D. P., & Poustka, L. (2022). Arc controls alcohol cue relapse by a central amygdala mechanism. Molecular Psychiatry. https://doi.org/10.1038/s41380-022-01849-4