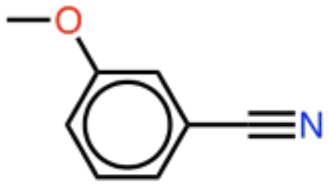
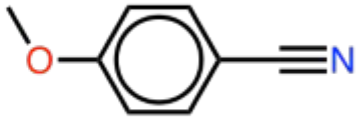


SMILES: COc(c1)cccc1C#N



SMILES: COc(cc1)ccc1C#N

MARGO X-AI HACKATHON

Juliette Anglade, Sacha Arroues-Paykin, Mario Massy, Mathias Ollu

MARGO

Qubit PHARMACEUTICALS

IBM®

Our group

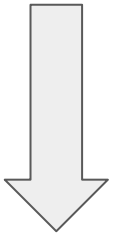
No prior experience in ML

No prior experience in bioinformatics

Little experience in chemistry

Leverage knowledge in

- statistics
- statistical learning
- graph theory

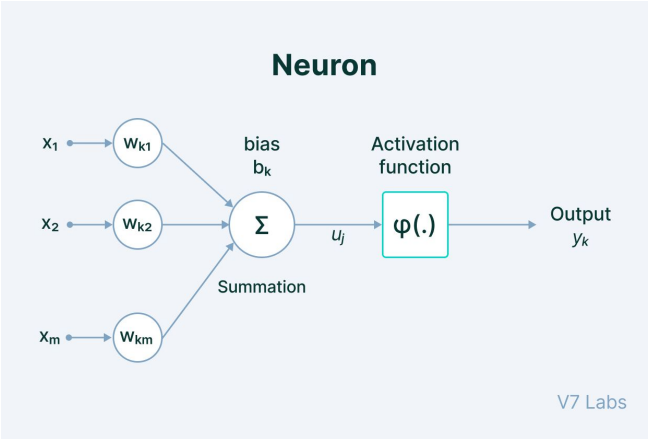


1. A first traditional approach through CNNs and GNNs, after a *Nature* paper in bioinformatics
2. A second approach based on random trees and forests

First Approach : naive CNN implementation + PCA

- base model
- PCA : principal component analysis
- Accuracy : around 70%
- **Problem** : 4296 features for 9415 rows, too many features ?

→ dimension reduction through **principal component analysis** on **footprints.** Does not improve results significantly



GNN - an attempt to a global implementation

- Approach from *Graph neural networks* **Nature** paper.

- graphs from the *Smiles* through rdkit
- other features through a MLP branch

- Accuracy : 84 % (test data from the file train.csv)
- But a lower cohen Kappa accuracy (0.44 in practice).
- Calculated a ROC (receiver operating characteristic) to find the best threshold for the binary classifier.

→ **Problem** : what threshold ? Maximize:

$$F1 = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

nature reviews methods primers

Explore content | About the journal | Publish with us | Subscribe

nature > nature reviews methods primers > primers > article

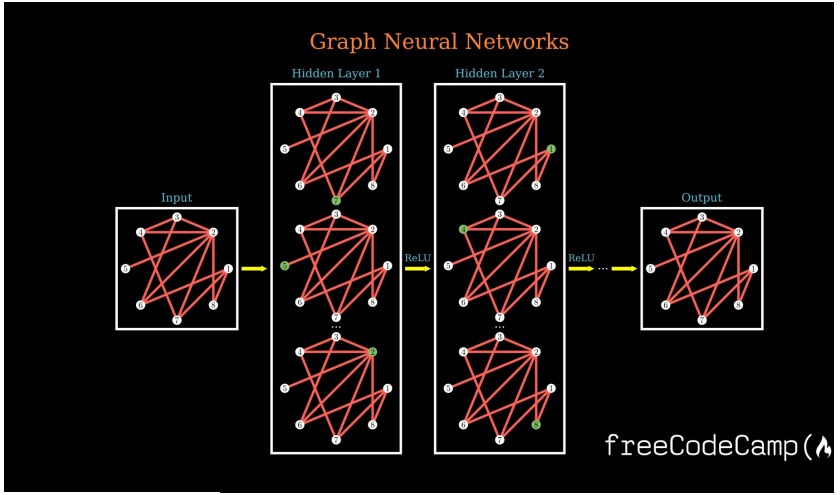
Primer | Published: 07 March 2024

Graph neural networks

Gabriele Corso, Hannes Stark, Stefanie Jegelka, Tommi Jaakkola & Regina Barzilay

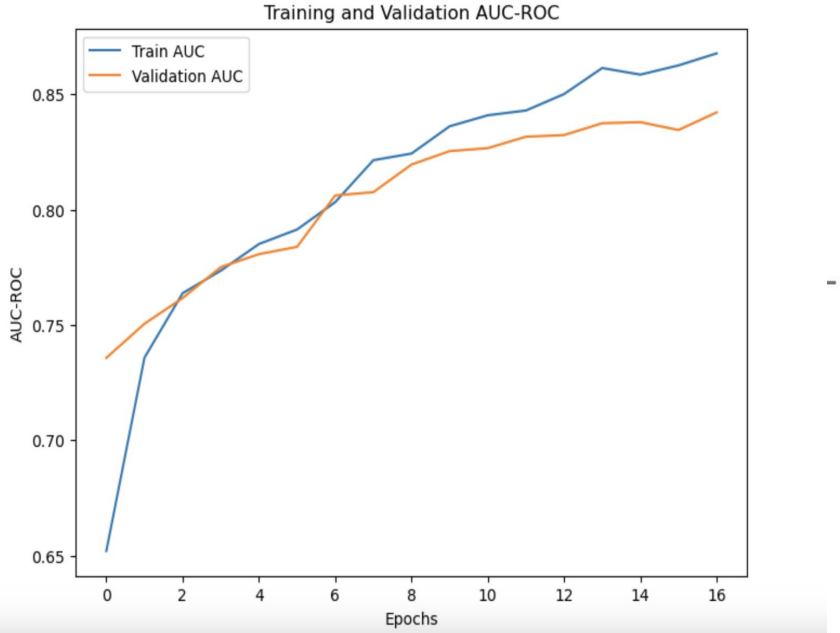
Nature Reviews Methods Primers **4**, Article number: 17 (2024) | Cite this article

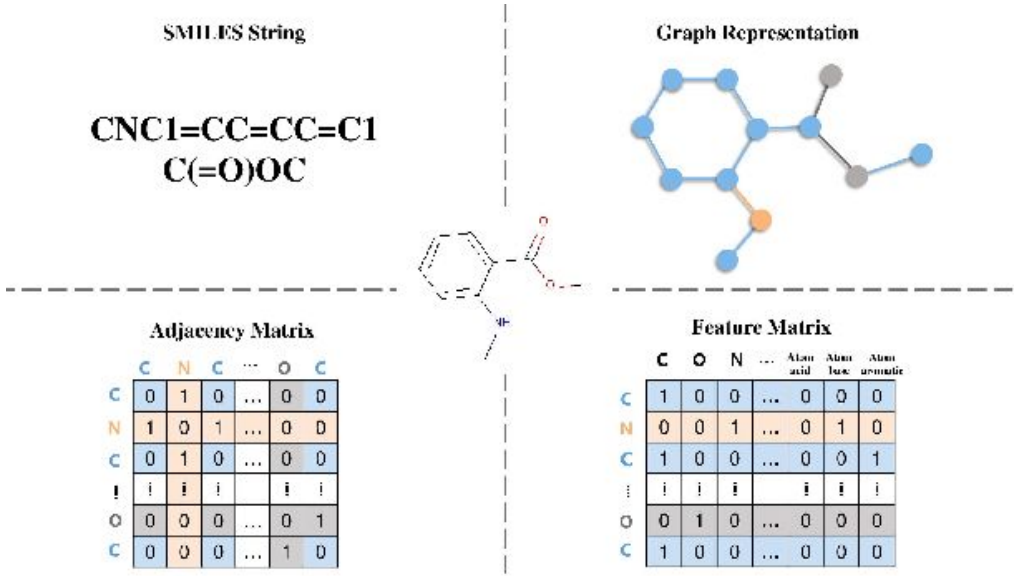
6766 Accesses | **21** Citations | **43** Altmetric | Metrics



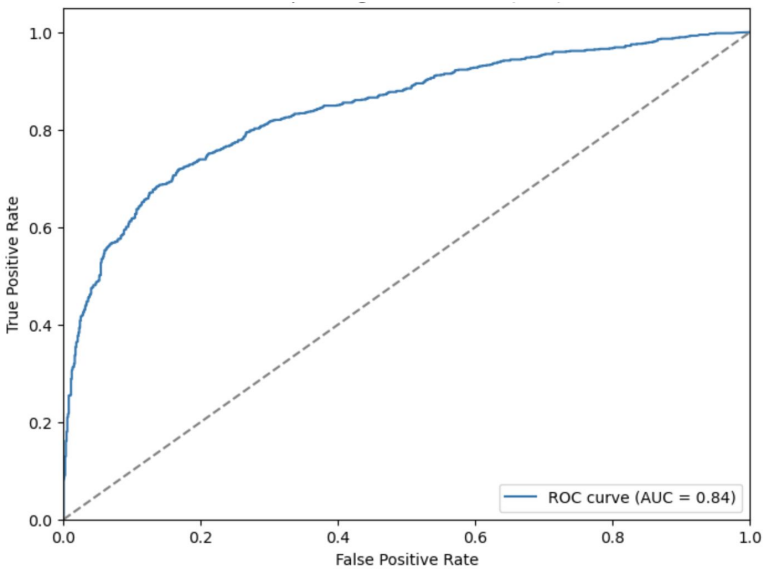
GNN - an attempt to a global implementation

```
Epoch 16/17 - Train AUC: 0.8625, Val AUC: 0.8345
Epoch 17/17 - Train AUC: 0.8677, Val AUC: 0.8421
```

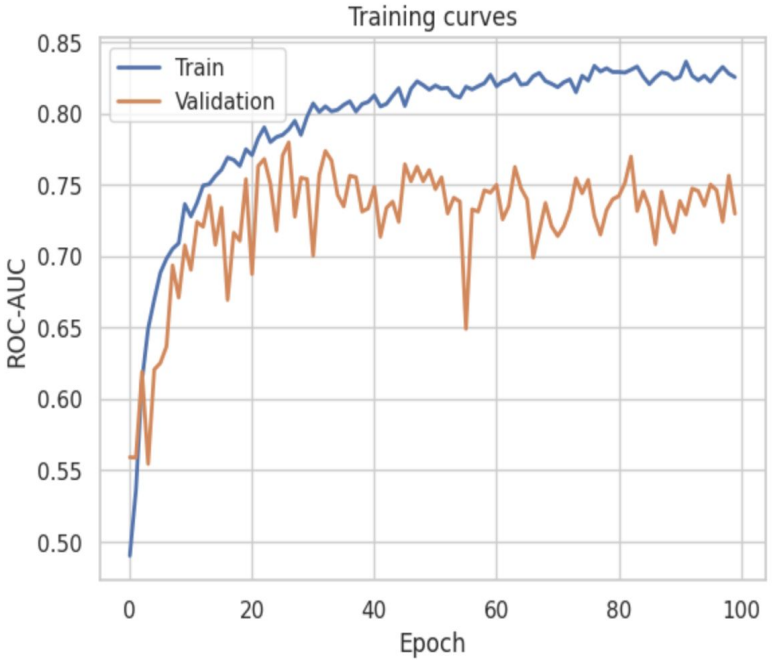




GNN - an attempt to a global implementation




```
0.5828064
Cohen's Kappa score: 0.5165
```

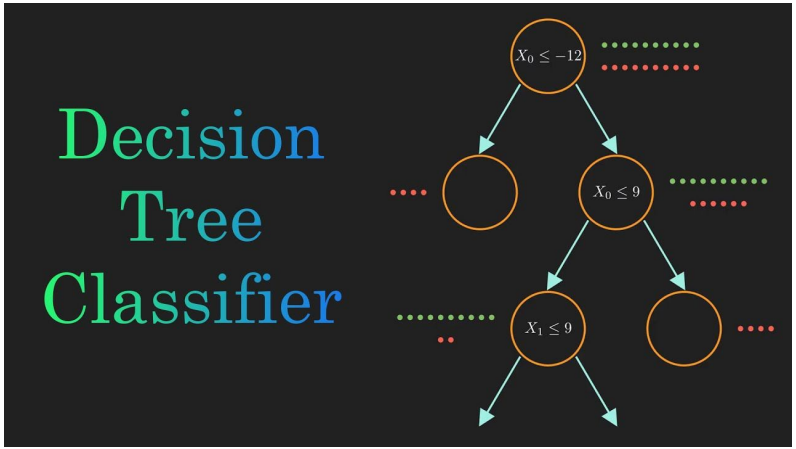


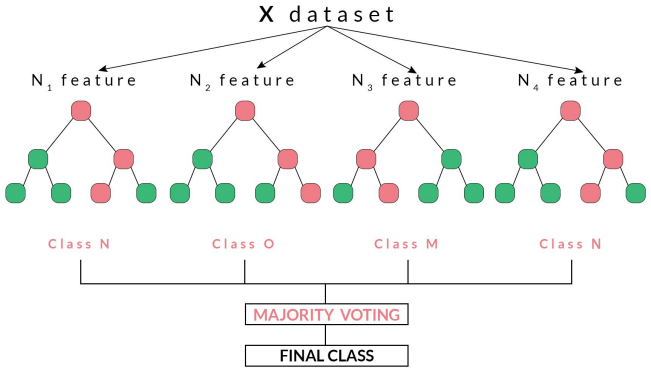
New approach : Random Forest model

Why use decision trees ? robust to irrelevant features/overfitting, tackles the “many features” problem, quick to train

Accuracy : 83% - Cohen Kappa : 0.596 - best 200: 0.955

Number of trees : 1000





Trying to select the relevant features

- random forest gives the importance of the features
- tried to select the most important
- worse accuracy…

Lessons from this hackathon

- The importance of the subject, not only applying all the ML known out there, but think of the structure
- Here it was important to not classify a toxic molecule as a non-toxic
- Team-work and communication are very important.

Conclusion: Thank you!