# Predicting biological activity

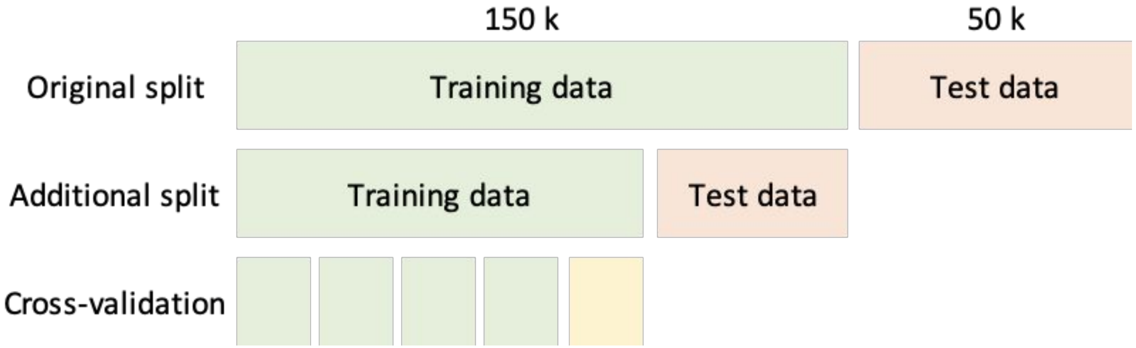
## ID2214 - Assignment 4

16 December 2022

Group 5 – Charlotte Jacquet, Isabella Rositi, Lukas Olenborg

# Methodology

1. Further divide training dataset into training-validation sets
2. Cross validation on the training set to find the optimal features subsets
3. Grid search combined with cross-validation to tune the hyperparameters
4. Validation with the best combination of hyperparameters on the validation set



150 k

50 k

Original split

Training data

Test data

Additional split

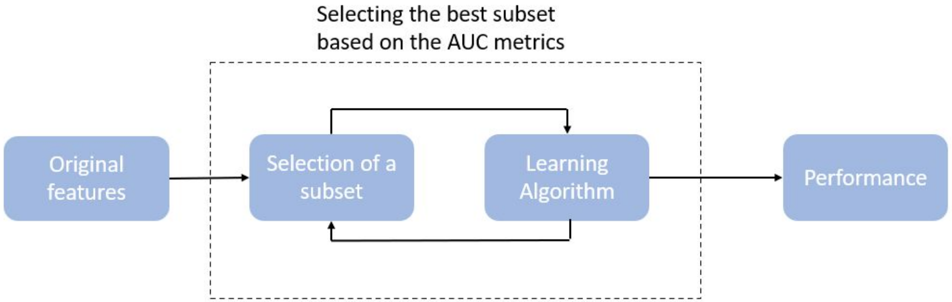
Training data

Test data

Cross-validation

# Feature Selection

1. Atomic features (Lipinski)
2. Molecular fingerprints (Morgan)
3. Combined representation

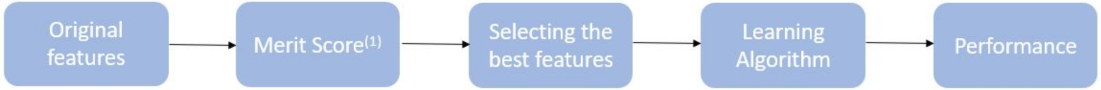


*Wrapper Method*

Feature selection methods:

- Wrapper
- Filter

→ **Use of all features**

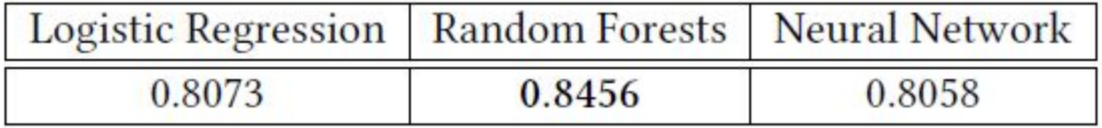


*Filter Method*

# Model selection

Fine tune hyperparameters with **grid search** and **5 folds cross-validation**:

1. Logistic Regression ( penalty = “L2”, fit_intercept = TRUE)
2. **Random Forest** ( n_estimators = 400, max_depth = 50)
3. Neural Network ( hidden_layer_sizes = (150, ), solver = “adam”, alpha = 0.0001, learning_rate = “adaptive”)



| Logistic Regression | Random Forests | Neural Network |
|---|---|---|
| 0.8073 | **0.8456** | 0.8058 |

AUCs calculated on the combined *validation set* using the best combination of hyperparameters