ROBERT v 1.0.5 2024/09/06 19:23:24

**How to cite:** Dalmau, D.; Alegre Requena, J. V. ChemRxiv, 2023, DOI: 10.26434/chemrxiv-2023-k994h

## ROBERT SCORE

*This score is designed to analyze the predictive ability of the models using different metrics.*

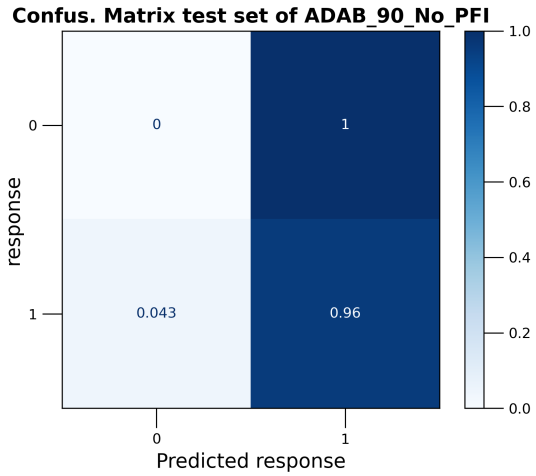
**No PFI (all descriptors):**

ML model: ADAB
Proportion Train:Validation:Test = 81:9:10

VERY WEAK

**The model has a score of 2/10**

- ● The test set shows an accuracy of 0.79
- - The test set uses 259:115 points:descriptors
- ● The test set passes 1 VERIFY tests
- - The test set passes 0 y-mean/y-shuffle




Train : Accuracy = 0.83, F1 score = 0.9, MCC = 0.46
Valid. : Accuracy = 0.96, F1 score = 0.98, MCC = 0.8
Test : Accuracy = 0.79, F1 score = 0.88, MCC = -0.09

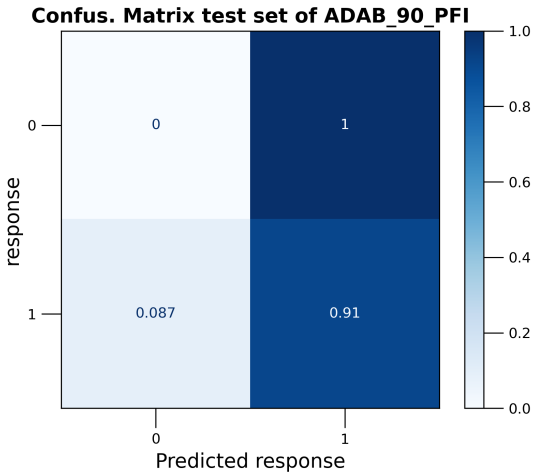
**PFI (only important descriptors):**

ML model: ADAB
Proportion Train:Validation:Test = 81:9:10

WEAK

**The model has a score of 4/10**

- ● The test set shows an accuracy of 0.75
- ●● The test set uses 259:3 points:descriptors
- ● The test set passes 1 VERIFY tests
- - The test set passes 0 y-mean/y-shuffle




Train : Accuracy = 0.79, F1 score = 0.88, MCC = 0.28
Valid. : Accuracy = 0.92, F1 score = 0.96, MCC = 0.0
Test : Accuracy = 0.75, F1 score = 0.86, MCC = -0.13

**Score thresholds** *(detailed in https://robert.readthedocs.io/en/latest/Score/score.html)*

| **Accuracy** | **Outliers** | **Points:descriptors** | **VERIFY tests** |
|---|---|---|---|
| ●● Accuracy > 0.85 | - Excluded in classif. | ●● > 10:1 p:d ratio | ●● y-shuffle & y-mean |
| ● 0.85 > Accur. > 0.70 | | ● 10:1 > p:d ratio > 3:1 | ● 5-fold CV & onehot |
| - Accur. < 0.70 | | - p:d ratio < 3:1 | - (all tests failed) |

Some tips to improve the score

⚠ Adding meaningful descriptors or replacing/deleting the least useful descriptors used might help. Feature importances are gathered in the SHAP and PFI sections of the /PREDICT/PREDICT_data.dat file.

How to predict new values with these models?

1. Create a CSV database with the new points, including the necessary descriptors.

2. Place the CSV file in the parent folder (i.e., where the module folders were created)

3. Run the PREDICT module as 'python -m robert --predict --csv_test FILENAME.csv'.

4. The predictions will be shown at the end of the resulting PDF report and will be stored in the last column of two CSV files called MODEL_SIZE_test(_No)_PFI.csv, which are in the PREDICT folder.

---

## REPRODUCIBILITY

*This section provides all the instructions to reproduce the results presented.*

**1. Download these files *(the authors should have uploaded the files as supporting information!)*:**

- CSV database (AQME-ROBERT_lipros_data.csv)

**2. Install and adjust the versions of the following Python modules:**

- Install ROBERT and its dependencies: conda install -c conda-forge robert
- Adjust ROBERT version: pip install robert==1.0.5
- scikit-learn-intelex: not installed

*(if scikit-learn-intelex is installed, slightly different results might be obtained)*

**3. Run ROBERT using this command line in the folder with the CSV database:**

python -m robert --y "response" --csv_name "AQME-ROBERT_lipros_data.csv" --type "clas" --names "code_name"

**4. Execution time, Python version and OS:**

Originally run in Python 3.10.14 using Linux #3672-Microsoft Fri Jan 01 08:00:00 PST 2016

Total execution time: 90694.58 seconds *(the number of processors should be specified by the user)*

## TRANSPARENCY

*This section contains important parameters used in scikit-learn models and ROBERT.*

**1. Parameters of the scikit-learn models (same keywords as used in scikit-learn):**

| **No PFI (all descriptors):** | **PFI (only important descriptors):** |
|---|---|
| sklearn model: AdaBoostClassifier | sklearn model: AdaBoostClassifier |
| random_state: 43 | random_state: 43 |
| names: code_name | names: code_name |
| n_estimators: 20 | n_estimators: 20 |
| learning_rate: 0.5 | learning_rate: 0.5 |

**2. ROBERT options for data split (KN or RND), predict type (REG or CLAS) and hyperopt error (RMSE, etc.):**

| **No PFI (all descriptors):** | **PFI (only important descriptors):** |
|---|---|
| split: RND | split: RND |
| type: clas | type: clas |
| error_type: acc | error_type: acc |

## ABBREVIATIONS

*Reference section for the abbreviations used.*

**ACC:** accuracy
**ADAB:** AdaBoost
**CSV:** comma separated values
**CLAS:** classification
**CV:** cross-validation
**F1 score:** balanced F-score
**GB:** gradient boosting
**GP:** gaussian process
**KN:** k-nearest neighbors
**MAE:** root-mean-square error
**MCC:** Matthew's correl. coefficient
**ML:** machine learning
**MVL:** multivariate lineal models
**NN:** neural network
**PFI:** permutation feature importance
**R2:** coefficient of determination
**REG:** Regression
**RF:** random forest
**RMSE:** root mean square error
**RND:** random
**SHAP:** Shapley additive explanations
**VR:** voting regressor

## AQME

*This module performs RDKit conformer generation from SMILES, followed by the creation of 200+ molecular and atomic descriptors using RDKit, xTB and DBSTEP (saved as AQME-ROBERT_FILENAME.csv).*

The complete output (AQME_data.dat) and raw data are stored in the AQME folder.

Time AQME: 89553.05 seconds

## CURATE

*This module takes care of data curation, including filters for correlated descriptors, noise, and duplicates, as well as conversion of categorical descriptors.*

The complete output (CURATE_data.dat) and curated database are stored in the CURATE folder.

Time CURATE: 1.52 seconds

**------- Images generated by the CURATE module -------**

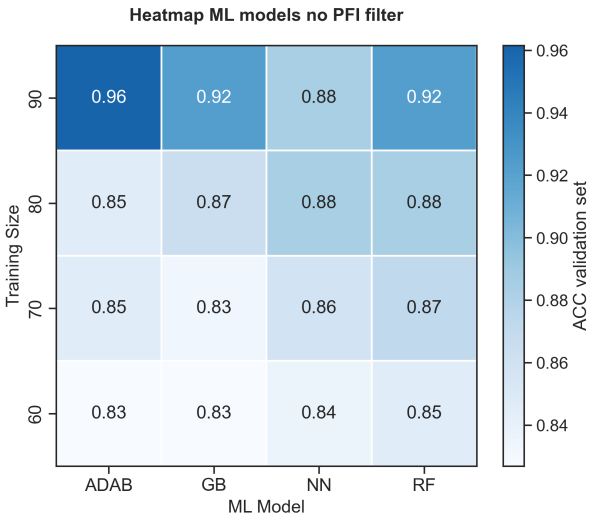
## GENERATE

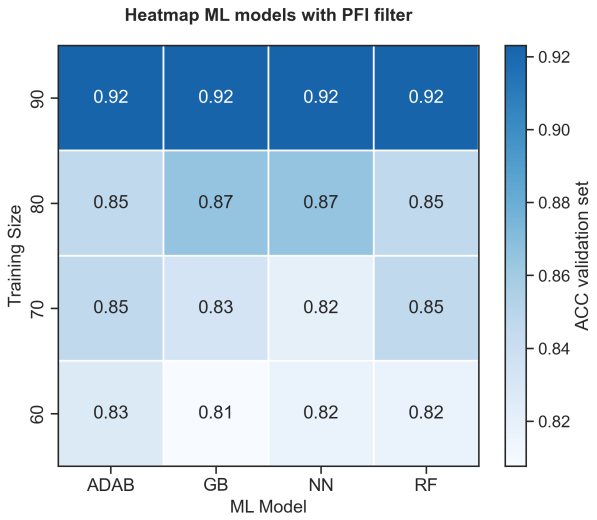
*This module carries out a screening of ML models and selects the most accurate one. It includes a comparison of multiple hyperoptimized models and training sizes.*

The complete output (GENERATE_data.dat) and heatmaps are stored in the GENERATE folder.

Time GENERATE: 1122.47 seconds

**------- Images generated by the GENERATE module -------**




## VERIFY

*Determination of predictive ability of models using four tests: 5-fold CV, y-mean (error against the mean y baseline), y-shuffle (predict with shuffled y values), and one-hot (predict using one-hot encoding instead of the X values).*

The complete output (VERIFY_data.dat) and donut plot are stored in the VERIFY folder.

Time VERIFY: 2.18 seconds

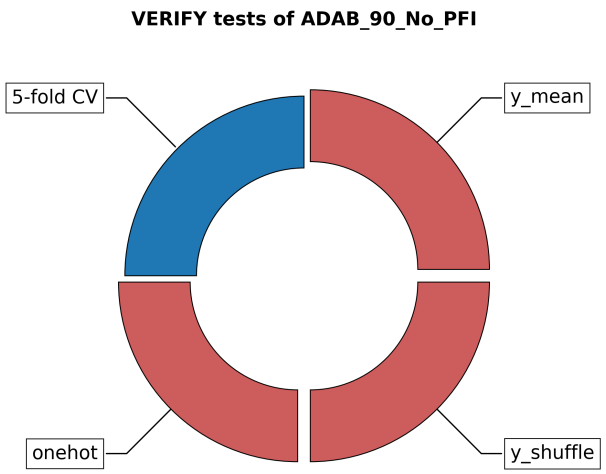
**------- Images and summary generated by the VERIFY module -------**

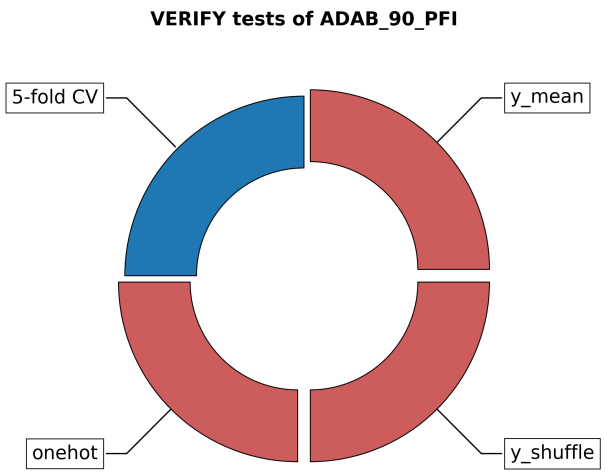
**No PFI (all descriptors):**

Original ACC (valid. set) 0.96 - 25% thres. = 0.72
- o 5-fold CV: PASSED, ACC = 0.76, higher than thres.
- x y_mean: FAILED, ACC = 0.92, higher than thres.
- x y_shuffle: FAILED, ACC = 0.81, higher than thres.
- x onehot: FAILED, ACC = 0.92, higher than thres.

**PFI (only important descriptors):**

Original ACC (valid. set) 0.92 - 25% thres. = 0.69
- o 5-fold CV: PASSED, ACC = 0.76, higher than thres.
- x y_mean: FAILED, ACC = 0.92, higher than thres.
- x y_shuffle: FAILED, ACC = 0.92, higher than thres.
- x onehot: FAILED, ACC = 0.92, higher than thres.



## PREDICT

*This module predicts and plots the results of training and validation sets from GENERATE, as well as from external test sets (if any). Feature importances from SHAP and PFI, and outlier analysis are also represented.*

The complete output (PREDICT_data.dat) and heatmaps are stored in the PREDICT folder.

Time PREDICT: 15.36 seconds

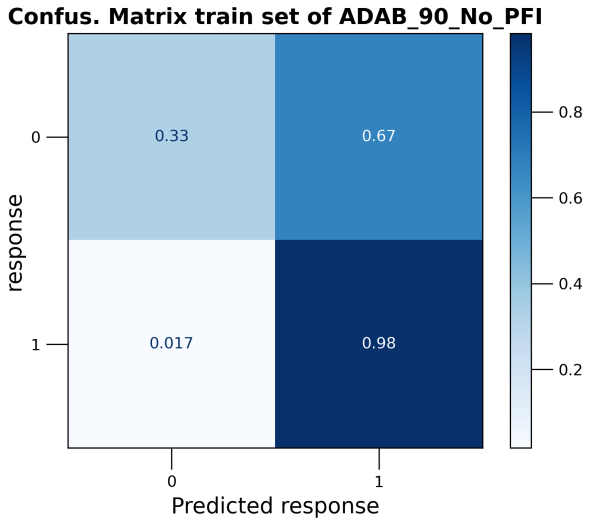
**------- Images and summary generated by the PREDICT module -------**

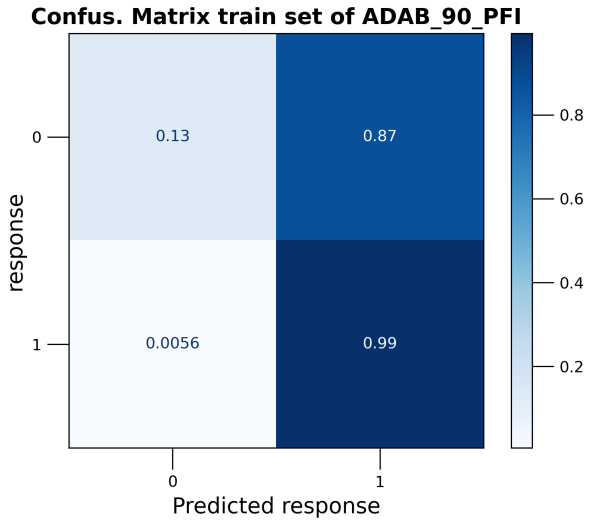
**No PFI (all descriptors):**

Prediction metrics and descriptors
- Points Train:Validation:Test = 233:26:28
- Proportion Train:Validation:Test = 81:9:10
- Number of descriptors = 115
- Proportion (train+valid.) points:descriptors = 259:115
- Train : Accuracy = 0.83, F1 score = 0.9, MCC = 0.46
- Valid. : Accuracy = 0.96, F1 score = 0.98, MCC = 0.8
- Test : Accuracy = 0.79, F1 score = 0.88, MCC = -0.09

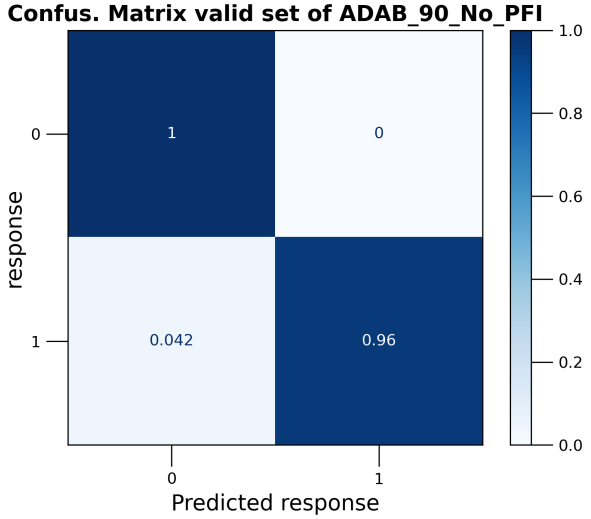
**PFI (only important descriptors):**

Prediction metrics and descriptors
- Points Train:Validation:Test = 233:26:28
- Proportion Train:Validation:Test = 81:9:10
- Number of descriptors = 3
- Proportion (train+valid.) points:descriptors = 259:3
- Train : Accuracy = 0.79, F1 score = 0.88, MCC = 0.28
- Valid. : Accuracy = 0.92, F1 score = 0.96, MCC = 0.0
- Test : Accuracy = 0.75, F1 score = 0.86, MCC = -0.13

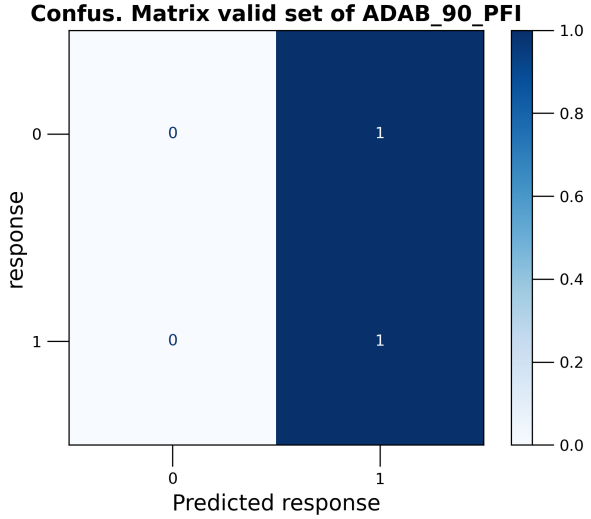


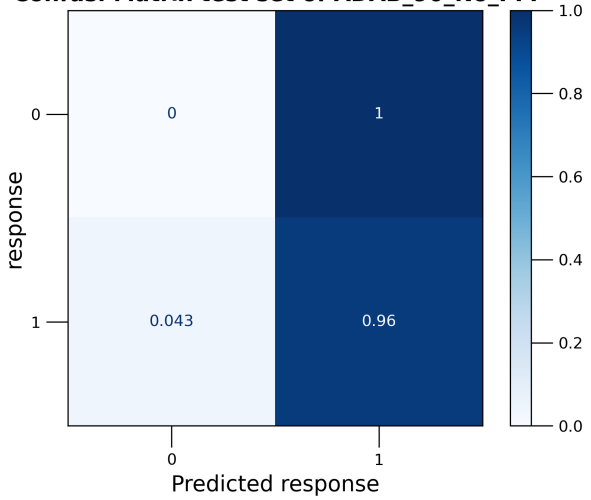
**Confus. Matrix valid set of ADAB_90_No_PFI**



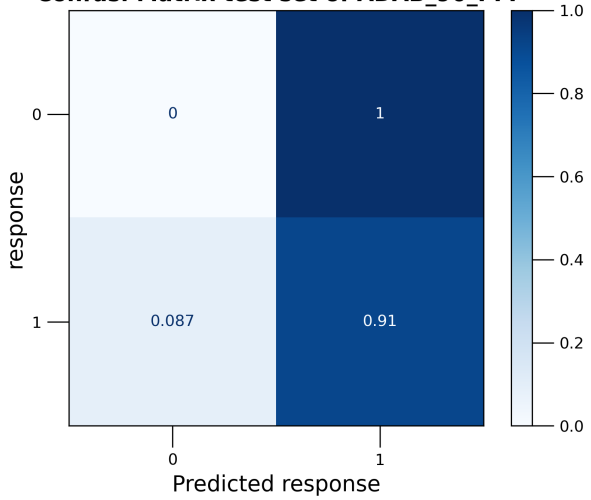

**Confus. Matrix valid set of ADAB_90_PFI**



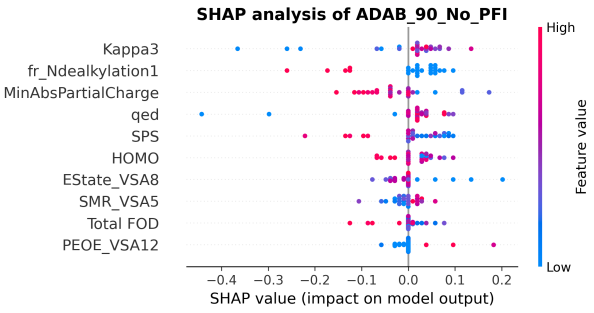

**Confus. Matrix test set of ADAB_90_No_PFI**



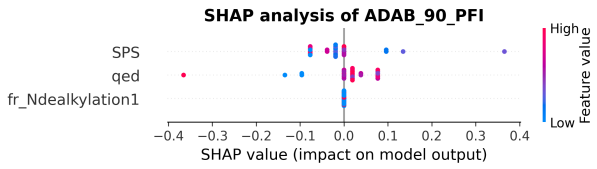

**Confus. Matrix test set of ADAB_90_PFI**



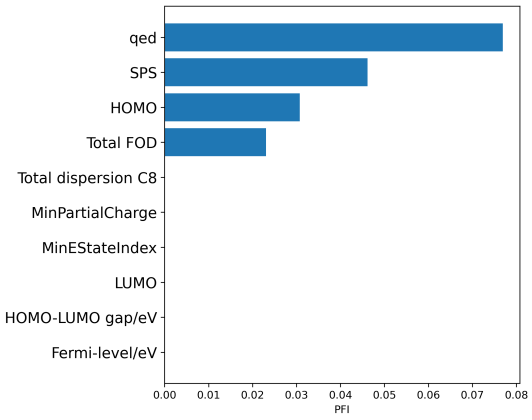

**SHAP analysis of ADAB_90_No_PFI**




**SHAP analysis of ADAB_90_PFI**



**Permutation feature importances (PFIs) of ADAB_90_No_PFI**




**Permutation feature importances (PFIs) of ADAB_90_PFI**