

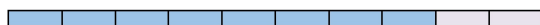


ROBERT v 1.0.6 2025/02/04 10:00:14

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: NN

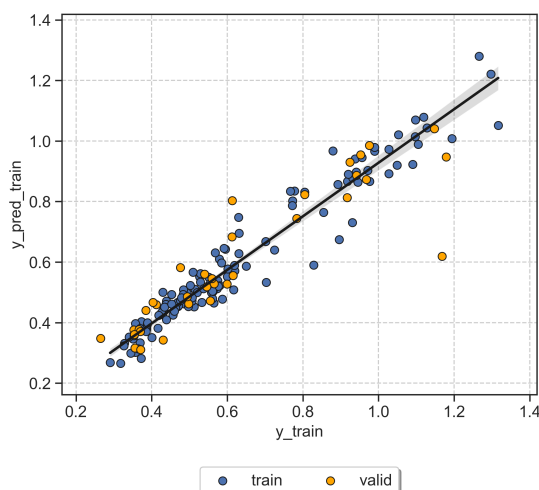
Proportion Train:Validation = 80:20



MODERATE

The model has a score of 8/10

- The valid. set shows an R^2 of 0.81
- The valid. set has 8.8% of outliers
- Using 166:8 points(train+valid.):descriptors
- The valid. set passes 4 VERIFY tests

Train : $R^2 = 0.94$, MAE = 0.047, RMSE = 0.068Valid. : $R^2 = 0.81$, MAE = 0.071, RMSE = 0.12**PFI (only important descriptors):**

ML model: NN

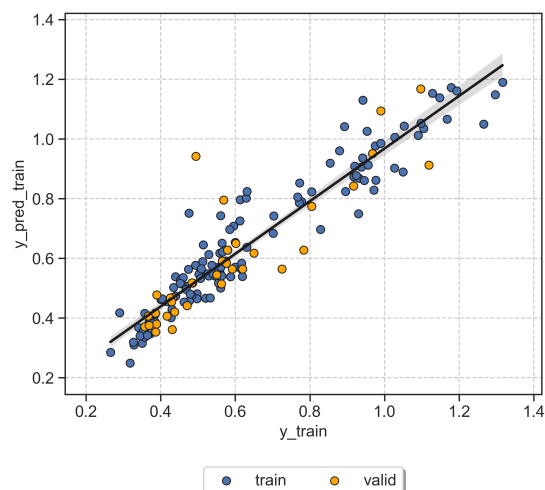
Proportion Train:Validation = 80:20



MODERATE

The model has a score of 7/10

- The valid. set shows an R^2 of 0.76
- The valid. set has 8.8% of outliers
- Using 166:6 points(train+valid.):descriptors
- The valid. set passes 3 VERIFY tests

Train : $R^2 = 0.92$, MAE = 0.056, RMSE = 0.077Valid. : $R^2 = 0.76$, MAE = 0.067, RMSE = 0.11**Score thresholds** (detailed in <https://robert.readthedocs.io/en/latest/Score/score.html>)

R^2 _____

- $R^2 > 0.85$
- $0.85 > R^2 > 0.70$
- $R^2 < 0.70$

Outliers _____

- $< 7.5\%$ of outliers
- $7.5\% < \text{outliers} < 15\%$
- $> 15\%$ of outliers

Points:descriptors _____

- $> 10:1$ p:d ratio
- $10:1 > \text{p:d ratio} > 3:1$
- p:d ratio $< 3:1$

VERIFY tests _____

Up to •••• (tests pass)

- (all tests failed)

Some tips to improve the score

- ⚠ Two of your models have more than 7.5% of outliers (5% is expected for a normal distribution with the t-value of 2 that ROBERT uses), using a more homogeneous distribution of results might help.
- ⚠ Replacing or deleting the least useful descriptors used might help to improve the model. 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 (smooth_sin_rGO_sin_DMF_clima_noise_ratio_m16.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==1.0.6`
- 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 --model "[NN]" --train "[80]" --ignore "[indice,area,cell]" --names "indice" --y "PCE" --auto_test  
"False" --csv_name "smooth_sin_rGO_sin_DMF_clima_noise_ratio_m16.csv"
```

4. Execution time, Python version and OS:

Originally run in Python 3.10.12 using Darwin Darwin Kernel Version 22.5.0: Thu Jun 8 22:22:19 PDT 2023; root:xnu-8796.121.3~7/RELEASE_ARM_T8020

Total execution time: 22.99 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):

sklearn model: MLPRegressor
 random_state: 70
 names: indice
 batch_size: 4
 hidden_layer_sizes: [16, 16]
 learning_rate_init: 0.01
 max_iter: 50
 validation_fraction: 0.2
 alpha: 0.0001
 shuffle: True
 tol: 0.0001
 early_stopping: False
 beta_1: 0.9
 beta_2: 0.999
 epsilon: 1e-08

PFI (only important descriptors):

sklearn model: MLPRegressor
 random_state: 19
 names: indice
 batch_size: 8
 hidden_layer_sizes: [32]
 learning_rate_init: 0.001
 max_iter: 200
 validation_fraction: 0.1
 alpha: 0.0001
 shuffle: True
 tol: 0.0001
 early_stopping: False
 beta_1: 0.8
 beta_2: 0.999
 epsilon: 1e-08

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

No PFI (all descriptors):

split: RND
 type: reg
 error_type: rmse

PFI (only important descriptors):

split: RND
 type: reg
 error_type: rmse



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

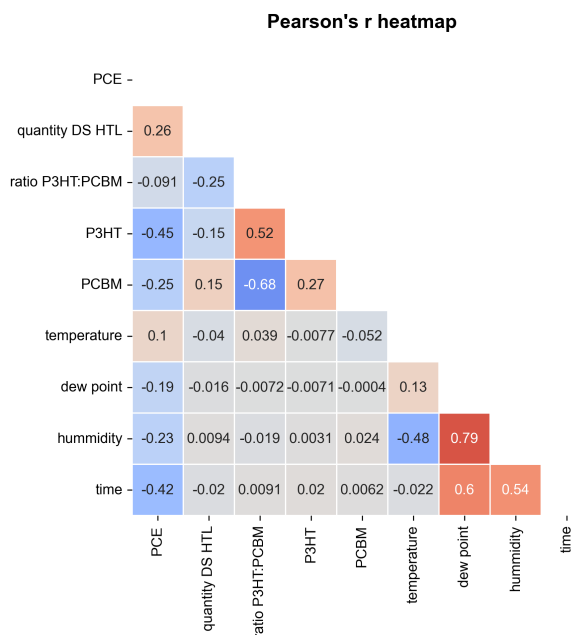
**CURATE**

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: 0.3 seconds

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

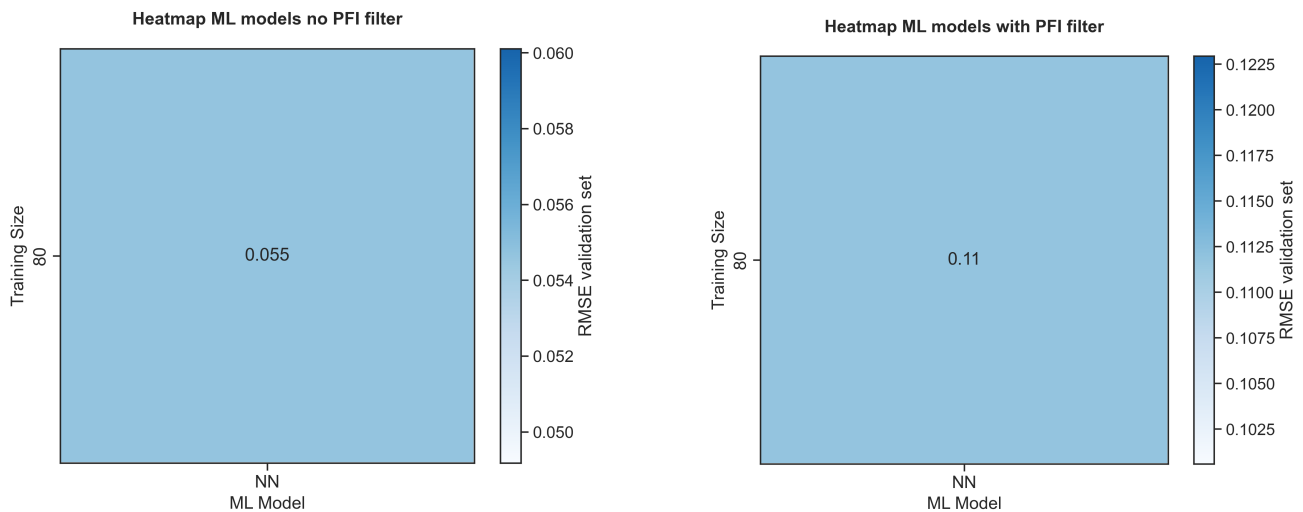
**GENERATE**

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: 16.23 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: 2.66 seconds

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

No PFI (all descriptors):

Original RMSE (valid. set) 0.12 + 25% thres. = 0.15

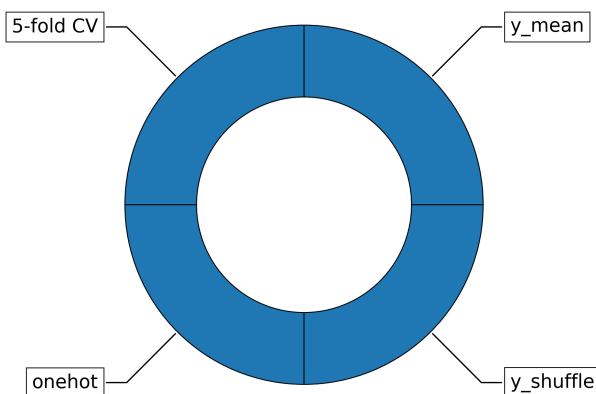
- o 5-fold CV: PASSED, RMSE = 0.09, lower than thres.
- o y_mean: PASSED, RMSE = 0.26, higher than thres.
- o y_shuffle: PASSED, RMSE = 0.37, higher than thres.
- o onehot: PASSED, RMSE = 0.28, higher than thres.

PFI (only important descriptors):

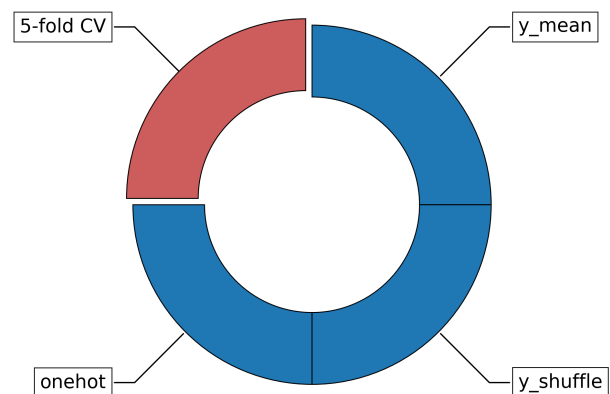
Original RMSE (valid. set) 0.11 + 25% thres. = 0.14

- o 5-fold CV: FAILED, RMSE = 0.16, higher than thres.
- o y_mean: PASSED, RMSE = 0.21, higher than thres.
- o y_shuffle: PASSED, RMSE = 0.24, higher than thres.
- o onehot: PASSED, RMSE = 0.21, higher than thres.

VERIFY tests of NN_80_No_PFI



VERIFY tests of NN_80_PFI



**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: 3.8 seconds

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

No PFI (all descriptors):Prediction metrics and descriptors

- Points Train:Validation = 132:34
- Proportion Train:Validation = 80:20
- Number of descriptors = 8
- Proportion (train+valid.) points:descriptors = 166:8
- Train : $R^2 = 0.94$, MAE = 0.047, RMSE = 0.068
- Valid. : $R^2 = 0.81$, MAE = 0.071, RMSE = 0.12

Outliers (max. 10 shown)

Train: 7 outliers out of 132 datapoints (5.3%)

- 5 (2.4 SDs)
- 9 (3.5 SDs)
- 13 (3.1 SDs)
- 14 (3.9 SDs)
- 72 (4.4 SDs)
- 82 (2.8 SDs)
- 163 (2.5 SDs)

Validation: 3 outliers out of 34 datapoints (8.8%)

- 1 (1e+01 SDs)
- 30 (2.9 SDs)
- 77 (3.8 SDs)

PFI (only important descriptors):Prediction metrics and descriptors

- Points Train:Validation = 132:34
- Proportion Train:Validation = 80:20
- Number of descriptors = 6
- Proportion (train+valid.) points:descriptors = 166:6
- Train : $R^2 = 0.92$, MAE = 0.056, RMSE = 0.077
- Valid. : $R^2 = 0.76$, MAE = 0.067, RMSE = 0.11

Outliers (max. 10 shown)

Train: 8 outliers out of 132 datapoints (6.1%)

- 4 (3.1 SDs)
- 13 (2.4 SDs)
- 30 (2.4 SDs)
- 67 (4.2 SDs)
- 75 (2.2 SDs)
- 80 (2.6 SDs)
- 121 (2.5 SDs)
- 152 (2.4 SDs)

Validation: 3 outliers out of 34 datapoints (8.8%)

- 7 (2.9 SDs)
- 93 (3.3 SDs)
- 161 (7.5 SDs)

