ROBERT v 1.0.4 2023/10/01 14:29:42

How to cite: ROBERT v 1.0.4, Dalmau, D.; Alegre-Requena, J. V., 2023. https://github.com/jvalegre/robert

ROBERT SCORE

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

No PFI (all descriptors):

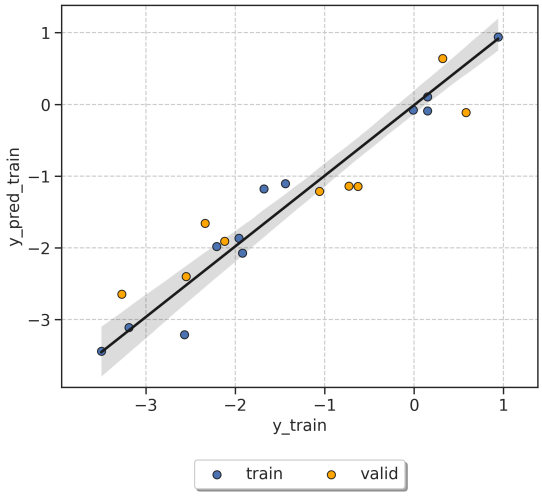
ML model: NN

Proportion Train:Validation = 57:43

WEAK

The model has a score of 5/10

- ●● The valid. set shows an R^2 of 0.89
- - The valid. set has 33.3% of outliers
- - Using 21:91 points(train+valid.):descriptors
- ●●● The valid. set passes 3 VERIFY tests



Train : R^2 = 0.96, MAE = 0.2, RMSE = 0.28

Valid. : R^2 = 0.89, MAE = 0.42, RMSE = 0.47

PFI (only important descriptors):

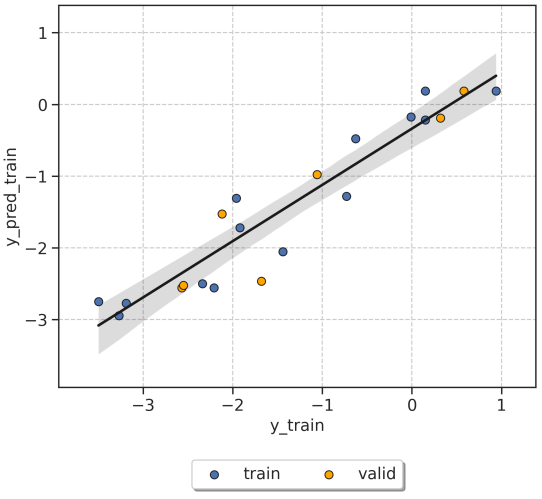
ML model: RF

Proportion Train:Validation = 67:33

MODERATE

The model has a score of 8/10

- ●● The valid. set shows an R^2 of 0.88
- ●● The valid. set has 0.0% of outliers
- ● Using 21:4 points(train+valid.):descriptors
- ●●● The valid. set passes 3 VERIFY tests



Train : R^2 = 0.91, MAE = 0.39, RMSE = 0.45

Valid. : R^2 = 0.88, MAE = 0.34, RMSE = 0.45

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

R^2	Outliers	Points:descriptors	VERIFY tests
●● R^2 > 0.85	●● < 7.5% of outliers	●● > 10:1 p:d ratio	Up to ●●●● (tests pass)
● 0.85 > R^2 > 0.70	● 7.5% < outliers < 15%	● 10:1 > p:d ratio > 3:1	- (all tests failed)
- R^2 < 0.70	- > 15% of outliers	- p:d ratio < 3:1	

Some tips to improve the score

⚠ The model uses only 21 datapoints, adding meaningful datapoints might help to improve the model.

⚠ One 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. For example, avoid using many points with similar y values and only a few points with distant y values.

⚠ 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 stored in the last column of two CSV files called MODEL_SIZE_test(_No)_PFI.csv, which are stored 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!)*:

- Report with results (ROBERT_report.pdf)
- CSV database (solubility_short.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==1.0.4
- Install scikit-learn-intelex: pip install scikit-learn-intelex==2023.2.1

(if scikit-learn-intelex is not installed, slightly different results might be obtained)

- Install AQME and its dependencies: conda install -c conda-forge aqme
- Adjust AQME version: pip install aqme==1.5.1
- Install xTB: conda install -c conda-forge xtb
- Adjust xTB version (if possible): conda install -c conda-forge xtb=6.6.1

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

python -m robert --aqme --y "solubility" --csv_name "solubility_short.csv"

4. Execution time, Python version and OS:

Originally run in Python 3.10.12 using Linux #1 SMP Tue Dec 21 19:02:23 UTC 2021

Total execution time: 86.86 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: MLPRegressor	sklearn model: RandomForestRegressor
random_state: 19	random_state: 43
names: code_name	names: code_name
batch_size: 4	n_estimators: 100
hidden_layer_sizes: [16, 16]	max_depth: 5
learning_rate_init: 0.01	max_features: 0.5
max_iter: 200	min_samples_split: 2
validation_fraction: 0.1	min_samples_leaf: 1
alpha: 0.0001	min_weight_fraction_leaf: 0
shuffle: True	ccp_alpha: 0
tol: 0.0001	oob_score: False
early_stopping: False	max_samples: 0.75
beta_1: 0.999	
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):	**PFI (only important descriptors):**
split: KN	split: KN
type: reg	type: reg
error_type: rmse	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

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: 25.75 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: 1.36 seconds

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

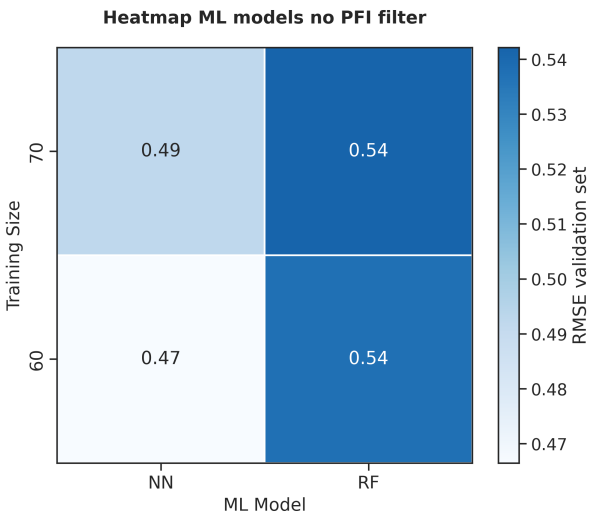
GENERATE

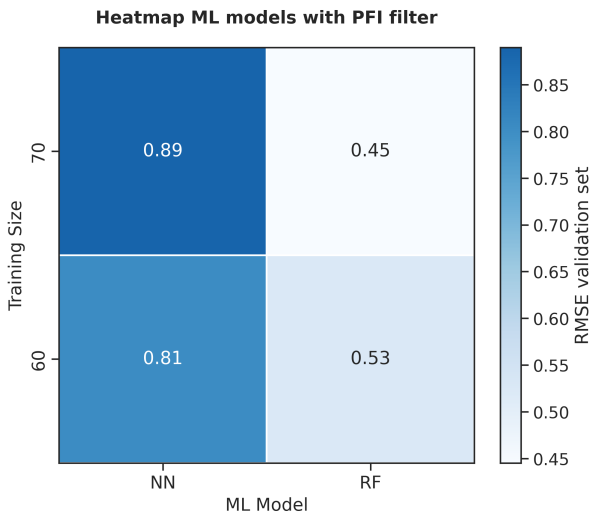
This module carries out a screening of ML models and selects the most accurate one. It includes a comparison of multiple hyperoptimized models and training sizes.

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

Time GENERATE: 50.14 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: 1.41 seconds

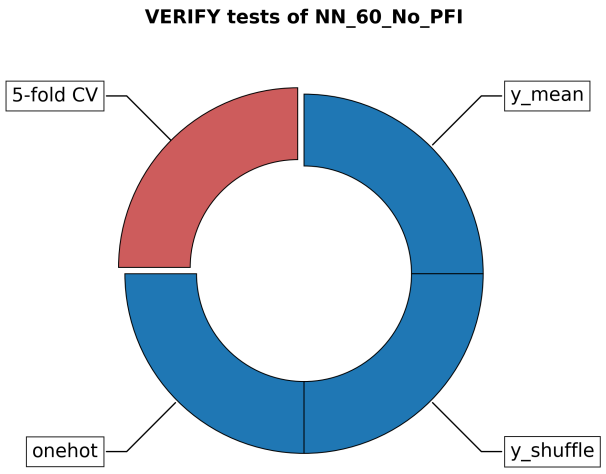
------- Images and summary generated by the VERIFY module -------

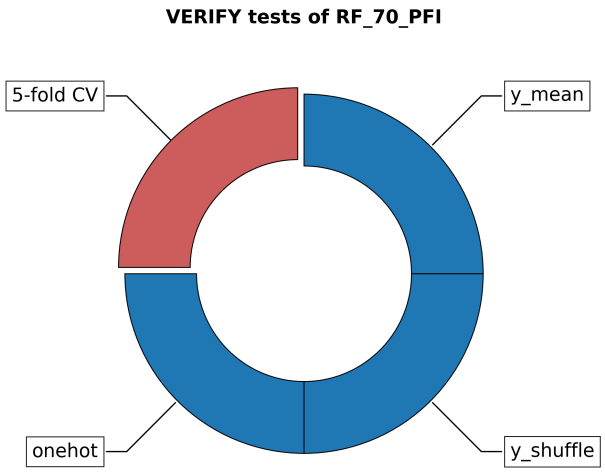
No PFI (all descriptors):

Original RMSE (valid. set) 0.47 + 25% thres. = 0.58
x 5-fold CV: FAILED, RMSE = 0.63, higher than thres.
o y_mean: PASSED, RMSE = 1.3, higher than thres.
o y_shuffle: PASSED, RMSE = 1.7, higher than thres.
o onehot: PASSED, RMSE = 0.69, higher than thres.

PFI (only important descriptors):

Original RMSE (valid. set) 0.45 + 25% thres. = 0.56
x 5-fold CV: FAILED, RMSE = 0.79, higher than thres.
o y_mean: PASSED, RMSE = 1.2, higher than thres.
o y_shuffle: PASSED, RMSE = 1.3, higher than thres.
o onehot: PASSED, RMSE = 0.98, 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: 8.2 seconds

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

No PFI (all descriptors):

Prediction metrics and descriptors
- Points Train:Validation = 12:9
- Proportion Train:Validation = 57:43
- Number of descriptors = 91
- Proportion (train+valid.) points:descriptors = 21:91
- Train : R^2 = 0.96, MAE = 0.2, RMSE = 0.28
- Valid. : R^2 = 0.89, MAE = 0.42, RMSE = 0.47

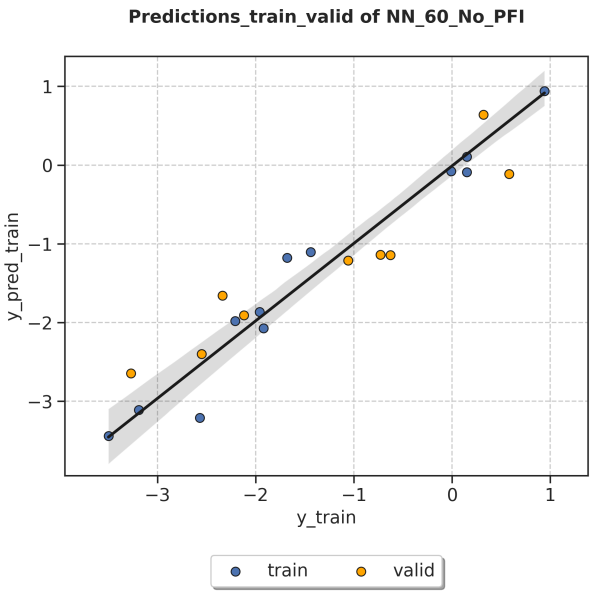
Outliers (max. 10 shown)
Train: 1 outliers out of 12 datapoints (8.3%)
- mol_1069 (2.3 SDs)

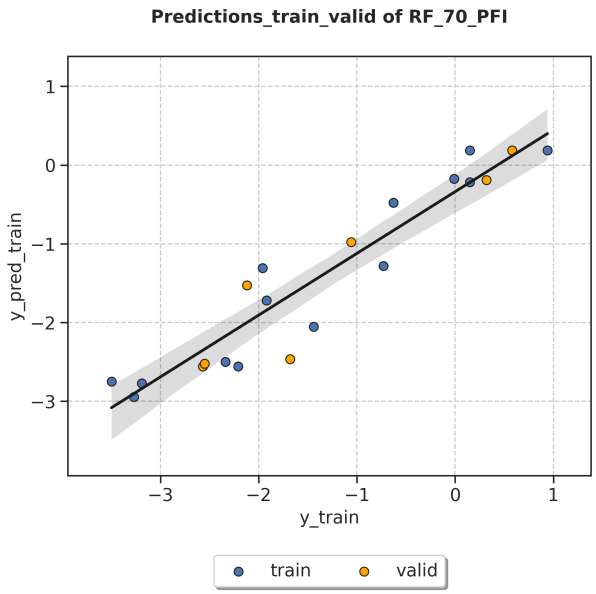
Validation: 3 outliers out of 9 datapoints (33.3%)
- mol_100 (2.2 SDs)
- mol_106 (2.6 SDs)
- mol_1076 (2.5 SDs)

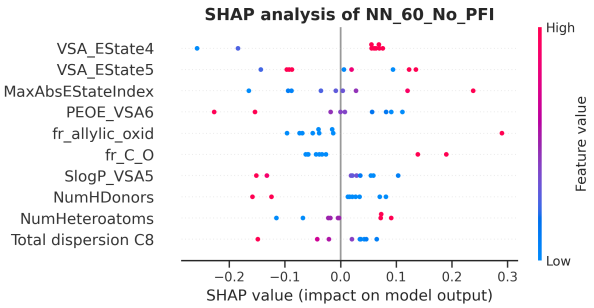
PFI (only important descriptors):

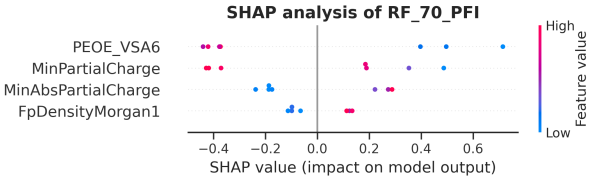
Prediction metrics and descriptors
- Points Train:Validation = 14:7
- Proportion Train:Validation = 67:33
- Number of descriptors = 4
- Proportion (train+valid.) points:descriptors = 21:4
- Train : R^2 = 0.91, MAE = 0.39, RMSE = 0.45
- Valid. : R^2 = 0.88, MAE = 0.34, RMSE = 0.45

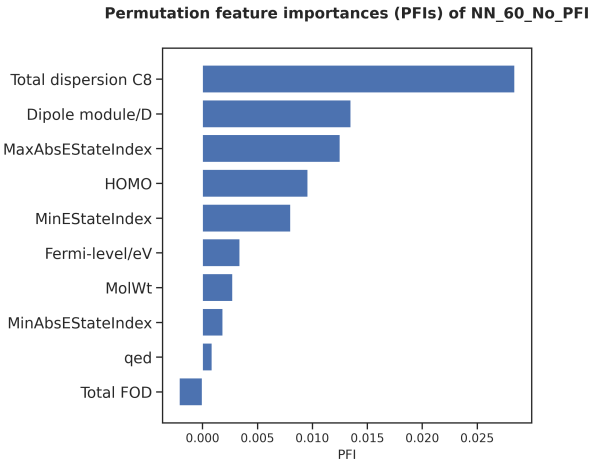
Outliers (max. 10 shown)
Train: 0 outliers out of 14 datapoints (0.0%)
Validation: 0 outliers out of 7 datapoints (0.0%)

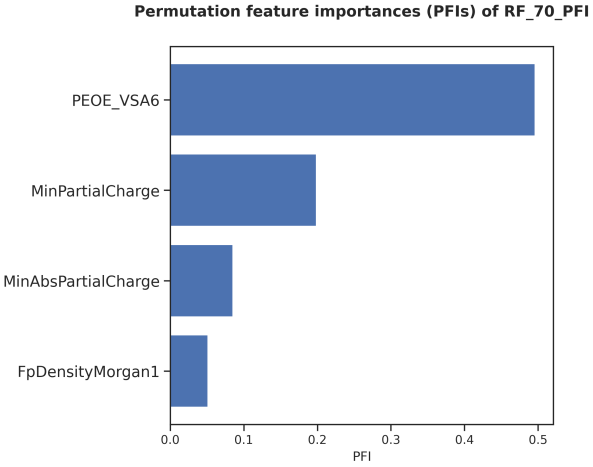


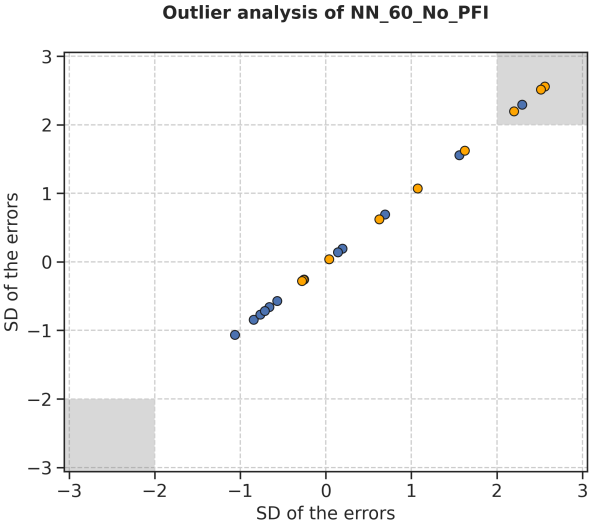


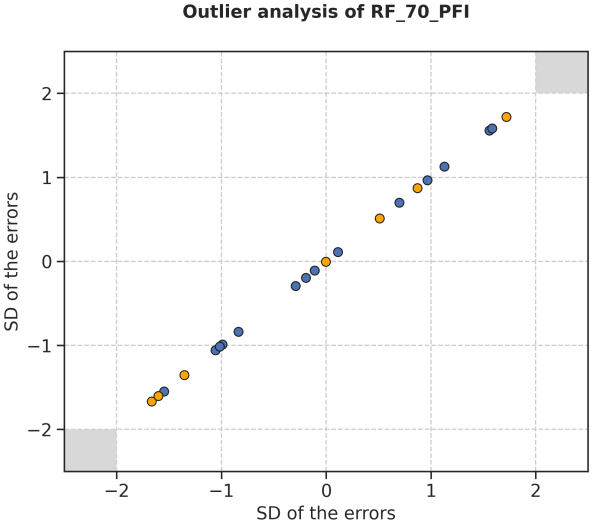

SHAP analysis of NN_60_No_PFI
VSA_EState4
VSA_EState5
MaxAbsEStateIndex
PEOE_VSA6
fr_allylic_oxid
fr_C_O
SlogP_VSA5
NumHDonors
NumHeteroatoms
Total dispersion C8
−0.2 −0.1 0.0 0.1 0.2 0.3
SHAP value (impact on model output)
High
Feature value
Low


SHAP analysis of RF_70_PFI
PEOE_VSA6
MinPartialCharge
MinAbsPartialCharge
FpDensityMorgan1
−0.4 −0.2 0.0 0.2 0.4 0.6
SHAP value (impact on model output)
High
Feature value
Low


Permutation feature importances (PFIs) of NN_60_No_PFI
Total dispersion C8
Dipole module/D
MaxAbsEStateIndex
HOMO
MinEStateIndex
Fermi-level/eV
MolWt
MinAbsEStateIndex
qed
Total FOD
0.000 0.005 0.010 0.015 0.020 0.025
PFI


Permutation feature importances (PFIs) of RF_70_PFI
PEOE_VSA6
MinPartialCharge
MinAbsPartialCharge
FpDensityMorgan1
0.0 0.1 0.2 0.3 0.4 0.5
PFI


Outlier analysis of NN_60_No_PFI
SD of the errors
−3 −2 −1 0 1 2 3
SD of the errors


Outlier analysis of RF_70_PFI
SD of the errors
−2 −1 0 1 2
SD of the errors