ROBERT v 1.0.6 2024/06/06 09:48:42

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.

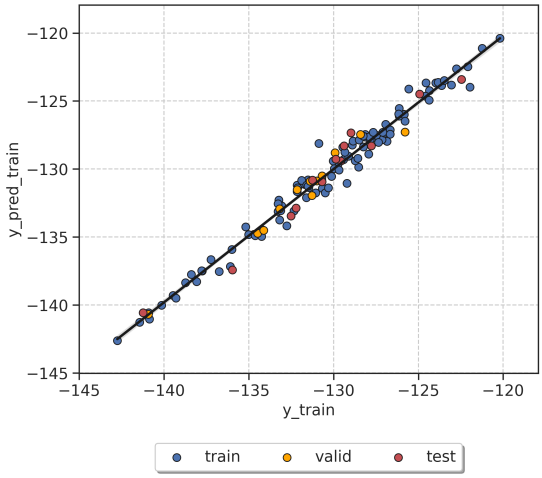
No PFI (all descriptors):

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

MODERATE

The model has a score of 7/10

- ●● The test set shows an R^2 of 0.96
- ● The valid. set has 8.3% of outliers
- ● Using 119:27 points(train+valid.):descriptors
- ●●● The valid. set passes 3 VERIFY tests



Train : R^2 = 0.98, MAE = 0.51, RMSE = 0.68
Valid. : R^2 = 0.96, MAE = 0.57, RMSE = 0.7
Test : R^2 = 0.96, MAE = 0.76, RMSE = 0.87

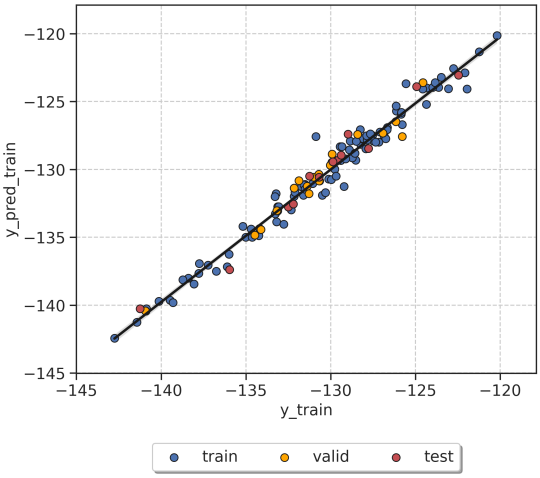
PFI (only important descriptors):

ML model: MVL
Proportion Train:Validation:Test = 77:14:10

STRONG

The model has a score of 9/10

- ●● The test set shows an R^2 of 0.97
- ●● The valid. set has 5.6% of outliers
- ● Using 119:16 points(train+valid.):descriptors
- ●●●● The valid. set passes 4 VERIFY tests



Train : R^2 = 0.98, MAE = 0.55, RMSE = 0.75
Valid. : R^2 = 0.96, MAE = 0.59, RMSE = 0.73
Test : R^2 = 0.97, MAE = 0.68, RMSE = 0.81

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

✔ A ROBERT score of 9 or 10 suggests that the predictive ability of your model is strong, congratulations!

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 (bmc_dft_robert_input_NBOonly.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
- Install scikit-learn-intelex: pip install scikit-learn-intelex==2024.4.0

(if scikit-learn-intelex is not 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 "dG_C5" --csv_name "bmc_dft_robert_input_NBOonly.csv" --names "Cofactor" --ignore "[Cofactor,dG_C4]" --train "[60,65,70,75,80,85,90]" --model "[NN,MVL,RF,GB]"

4. Execution time, Python version and OS:

Originally run in Python 3.10.14 using Linux #1 SMP Tue Oct 19 15:14:17 UTC 2021

Total execution time: 422.57 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: LinearRegression	sklearn model: LinearRegression
random_state: 8	random_state: 8
names: Cofactor	names: Cofactor

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: 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

CURATE

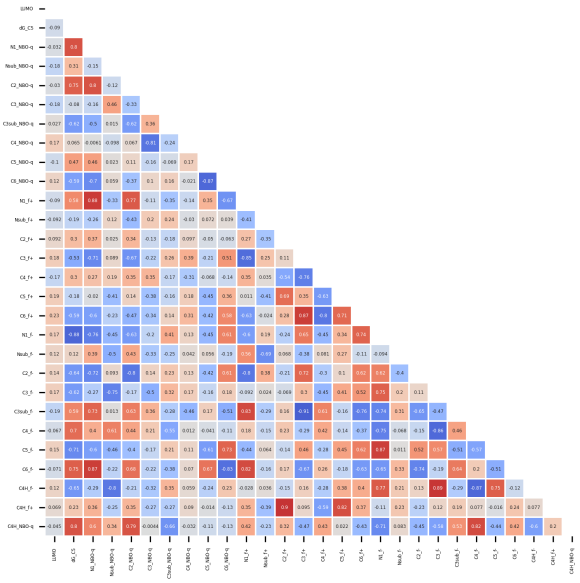
This module takes care of data curation, including filters for correlated descriptors, noise, and duplicates, as well as conversion of categorical descriptors.

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

Time CURATE: 1.01 seconds

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

Pearson's r heatmap



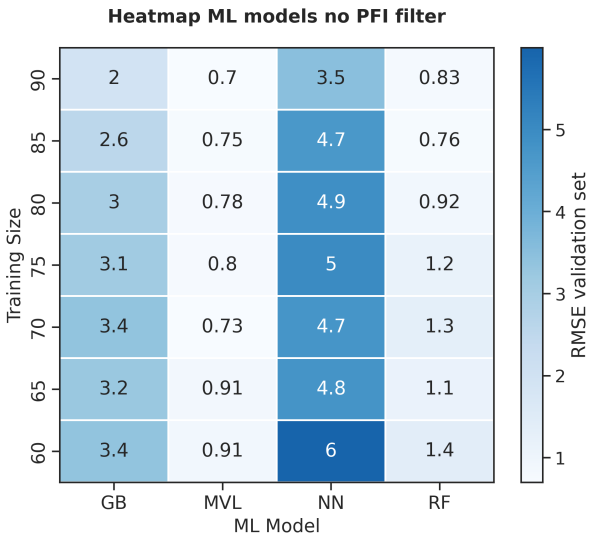
GENERATE

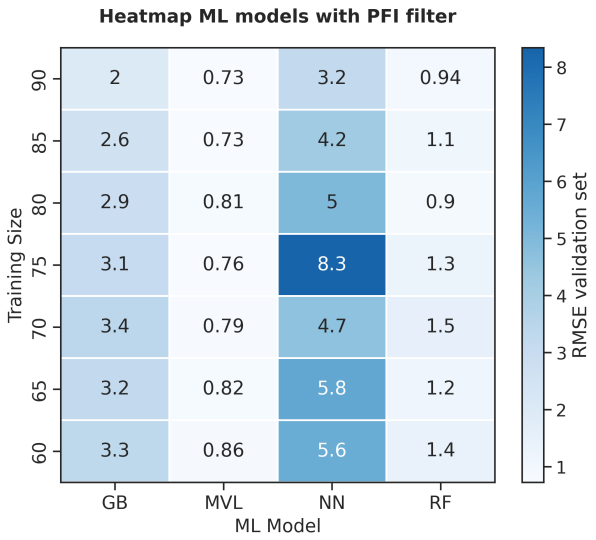
This module carries out a screening of ML models and selects the most accurate one. It includes a comparison of multiple hyperoptimized models and training sizes.

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

Time GENERATE: 414.91 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: 0.68 seconds

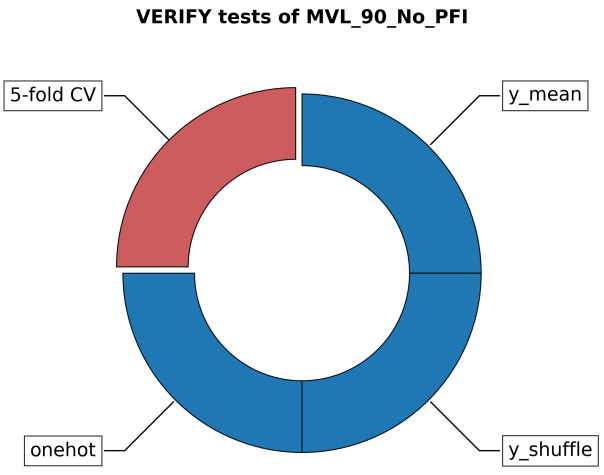
------- Images and summary generated by the VERIFY module -------

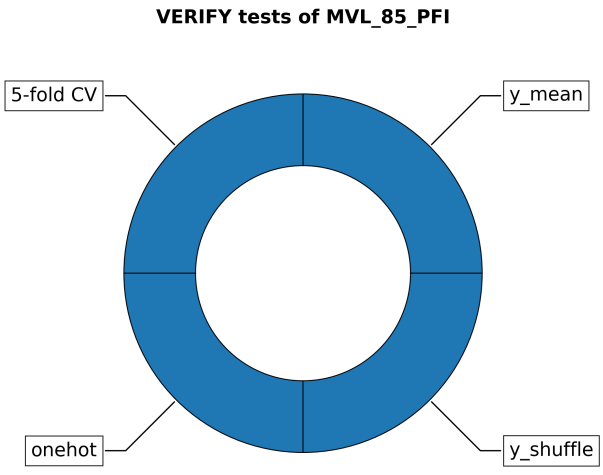
No PFI (all descriptors):

Original RMSE (valid. set) 0.7 + 25% thres. = 0.87
- x 5-fold CV: FAILED, RMSE = 0.97, higher than thres.
- o y_mean: PASSED, RMSE = 3.5, higher than thres.
- o y_shuffle: PASSED, RMSE = 5.6, higher than thres.
- o onehot: PASSED, RMSE = 3.9, higher than thres.

PFI (only important descriptors):

Original RMSE (valid. set) 0.73 + 25% thres. = 0.91
- o 5-fold CV: PASSED, RMSE = 0.84, lower than thres.
- o y_mean: PASSED, RMSE = 3.7, higher than thres.
- o y_shuffle: PASSED, RMSE = 6.0, higher than thres.
- o onehot: PASSED, RMSE = 3.7, 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: 5.97 seconds

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

No PFI (all descriptors):

Prediction metrics and descriptors
- Points Train:Validation:Test = 107:12:13
- Proportion Train:Validation:Test = 81:9:10
- Number of descriptors = 27
- Proportion (train+valid.) points:descriptors = 119:27
- Train : R^2 = 0.98, MAE = 0.51, RMSE = 0.68
- Valid. : R^2 = 0.96, MAE = 0.57, RMSE = 0.7
- Test : R^2 = 0.96, MAE = 0.76, RMSE = 0.87

Outliers (max. 10 shown)
Train: 4 outliers out of 107 datapoints (3.7%)
- BMC_009 (5.0 SDs)
- BMC_010 (2.1 SDs)
- BMC_039 (2.9 SDs)
- BMC_046 (3.3 SDs)

Validation: 1 outliers out of 12 datapoints (8.3%)
- BMC_045 (2.1 SDs)

Test: 2 outliers out of 13 datapoints (15.4%)
- BMC_078 (2.1 SDs)
- BMC_092 (2.5 SDs)

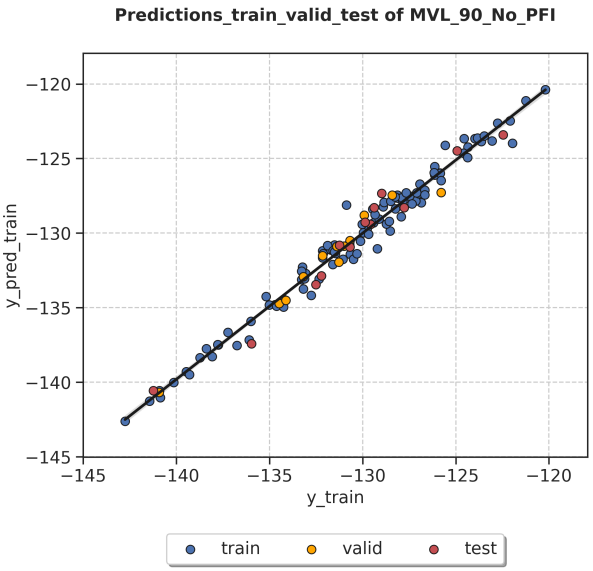
PFI (only important descriptors):

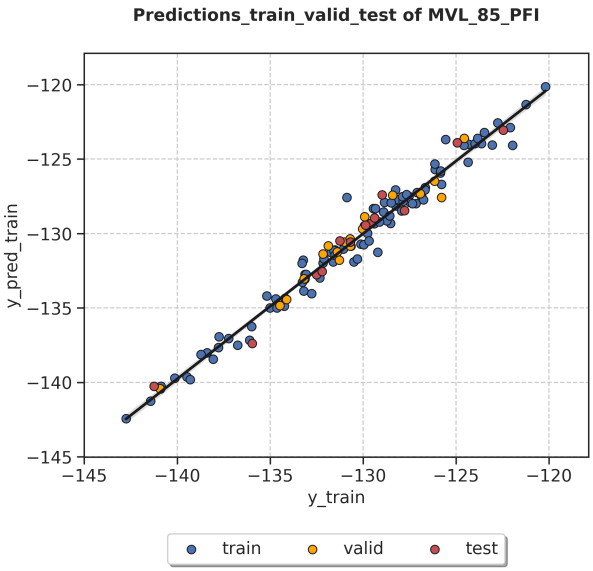
Prediction metrics and descriptors
- Points Train:Validation:Test = 101:18:13
- Proportion Train:Validation:Test = 77:14:10
- Number of descriptors = 16
- Proportion (train+valid.) points:descriptors = 119:16
- Train : R^2 = 0.98, MAE = 0.55, RMSE = 0.75
- Valid. : R^2 = 0.96, MAE = 0.59, RMSE = 0.73
- Test : R^2 = 0.97, MAE = 0.68, RMSE = 0.81

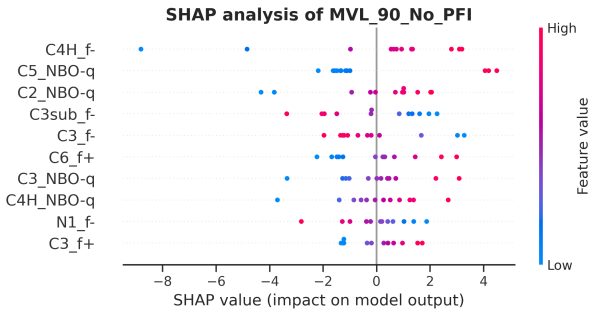
Outliers (max. 10 shown)
Train: 4 outliers out of 101 datapoints (4.0%)
- BMC_009 (5.3 SDs)
- BMC_010 (2.6 SDs)
- BMC_039 (2.9 SDs)
- BMC_046 (3.0 SDs)

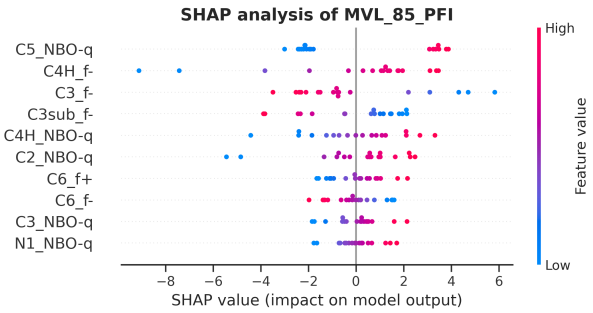
Validation: 1 outliers out of 18 datapoints (5.6%)
- BMC_045 (2.4 SDs)

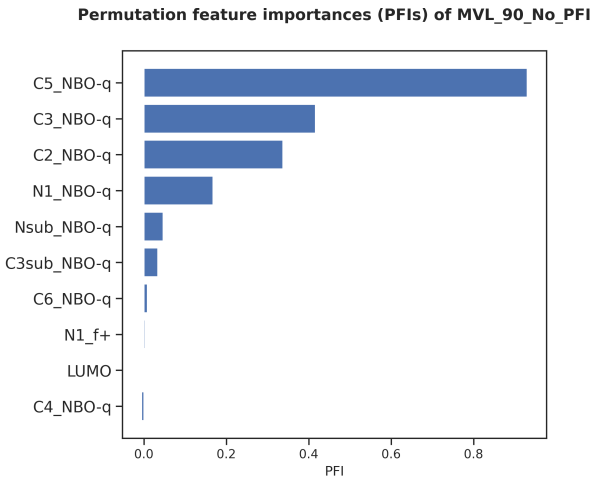
Test: 0 outliers out of 13 datapoints (0.0%)

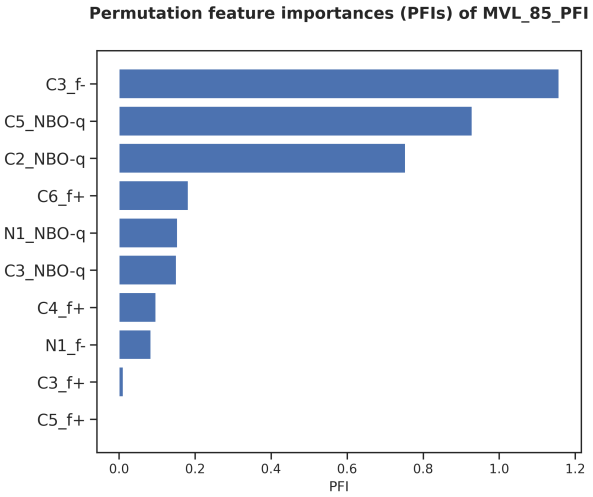


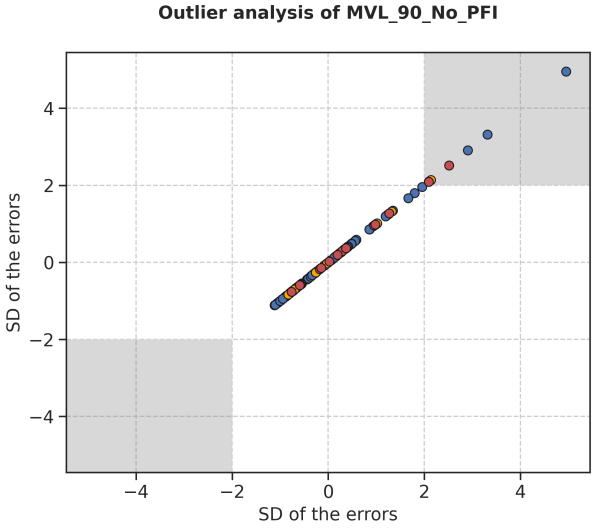


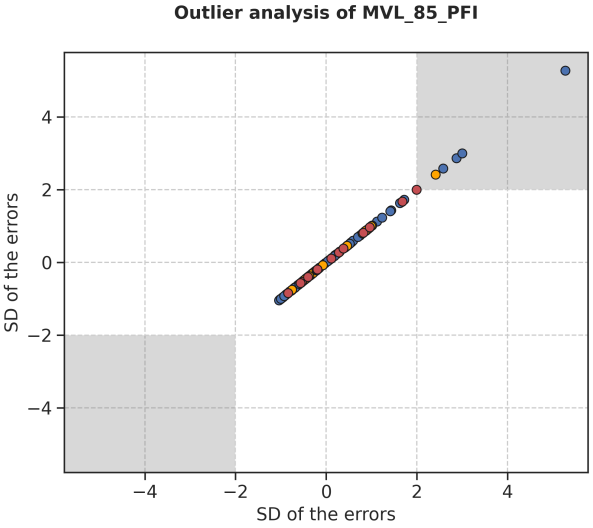

SHAP analysis of MVL_90_No_PFI
C4H_f-
C5_NBO-q
C2_NBO-q
C3sub_f-
C3_f-
C6_f+
C3_NBO-q
C4H_NBO-q
N1_f-
C3_f+
−8 −6 −4 −2 0 2 4
SHAP value (impact on model output)
High
Feature value
Low


SHAP analysis of MVL_85_PFI
C5_NBO-q
C4H_f-
C3_f-
C3sub_f-
C4H_NBO-q
C2_NBO-q
C6_f+
C6_f-
C3_NBO-q
N1_NBO-q
−8 −6 −4 −2 0 2 4 6
SHAP value (impact on model output)
High
Feature value
Low


Permutation feature importances (PFIs) of MVL_90_No_PFI
C5_NBO-q
C3_NBO-q
C2_NBO-q
N1_NBO-q
Nsub_NBO-q
C3sub_NBO-q
C6_NBO-q
N1_f+
LUMO
C4_NBO-q
0.0 0.2 0.4 0.6 0.8
PFI


Permutation feature importances (PFIs) of MVL_85_PFI
C3_f-
C5_NBO-q
C2_NBO-q
C6_f+
N1_NBO-q
C3_NBO-q
C4_f+
N1_f-
C3_f+
C5_f+
0.0 0.2 0.4 0.6 0.8 1.0 1.2
PFI


Outlier analysis of MVL_90_No_PFI
SD of the errors
4
2
0
−2
−4
−4 −2 0 2 4
SD of the errors


Outlier analysis of MVL_85_PFI
SD of the errors
4
2
0
−2
−4
−4 −2 0 2 4
SD of the errors