

Tutorial

Straightforwarding Scoring Suite (3S)

Version 1.0.0

Author: Milad Rayka



STRAIGHTFORWARDING SCORING SUITE

3S

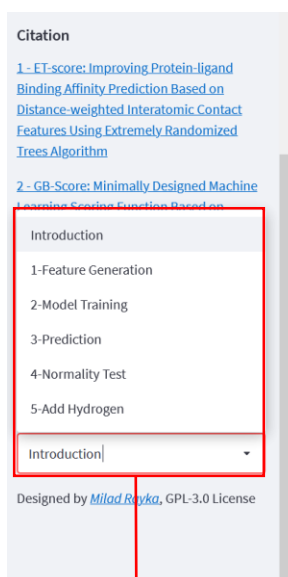
Table of contents

1 – Introduction	3
2 – Feature Generation	3
3 – Model Training.....	5
4 - Prediction	6
5 - Normality Test	7
6 - Add Hydrogen.....	12

1 – Introduction

Straightforwarding Scoring Suite (3S) is a collection of several tools to ease the procedure of designing a machine learning scoring function. These tools are designed based on our recent papers which we introduced a new scheme of feature generation based on distance-weighted interatomic contact and using Gradient Boosting Trees as a machine learning algorithm.

So far, this suite contains five tools: 1-Feature Generation 2-Model Training 3-Prediction 4-Normality Test 5-Add Hydrogen.



Different operations can be chosen from this side panel.

3S Web Application



Straightforwarding Scoring Suite (3S) is a collection of several tools to ease the procedure of designing a machine learning scoring function. These tools are designed based on our recent papers which we introduced a new scheme of feature generation based on distance-weighted interatomic contact and using Gradient Boosting Trees as a machine learning algorithm.

2 – Feature Generation

In this mode, features for different structure of complexes based on aforementioned method are generated.

3S Web Application

Feature Generation

The screenshot shows the 'Feature Generation' section of the 3S Web Application. It includes several input fields and expandable sections. Annotations with red arrows point to specific elements:

- Theoretical Information**: An expander button with a '+' icon. Annotation: "This expander provides information about feature generation method."
- Caution about directory** and **Caution about structures**: Two expandable sections. Annotation: "Some cautions about complex structures."
- Enter directory of your complex structures:** A text input field with "Example/structures" as a placeholder. A help icon (?) is circled, with an arrow pointing to a "Help" button. Annotation: "Help"
- Enter exponent of weighting function:** A numeric input field with "2.00". A help icon (?) is circled. Annotation: "Parameters for feature generation method. See *theoretical information* expander."
- Optimized value for n is 2.**: A blue informational message.
- Enter distance cutoff:** A numeric input field with "12.00". A help icon (?) is circled. Annotation: "Parameters for feature generation method. See *theoretical information* expander."
- Prefered value for (d_{cutoff}) is 12 Å.**: A blue informational message.
- Single pdb file for complex.**: A checkbox. Annotation: "Check this, if ligand and protein structures are in a single pdb file."
- Enter output filename in .csv:** A text input field with "sample.csv". A help icon (?) is circled. Annotation: "Output filename in csv format."
- Start feature generations**: A button. Annotation: "Click on this to start operation."

3 – Model Training

In this mode, a machine learning scoring function (Gradient Boosting Trees) is designed for a dataset of provided complex structures.

Model Training

Theoretical Information

Enter path of features data (.csv):

Example/files/x_set.csv

Caution: Your target data csv file should has two columns with the following names: **pdbid** and **binding_affinity**.

Enter path of target data (.csv):

Example/files/y_set.csv

Caution: Your test set csv file should has two columns with the following names: **pdbid** and **binding_affinity**.

Enter path of test set pdbid (.csv):

Example/files/test_set_pdbid.csv

Enter variance threshold:

0.01

Enter correlation threshold:

0.95

☐ Use GPU accelerator during training.

Enter output filename in .joblib:

saved_model.joblib

Start training operation

Click on this to start operation.

Output trained model name.

Parameters for preprocessing step. More information is provided in *theoretical information* expander.

4-Prediction

Binding affinity of complexes are predicted using a ML-Score.

First, use feature generation mode for generate features.

Binding Affinity Prediction

Binding affinity of complexes are predicted using a ML-Score.

Caution 1: All complexes have to turn to numerical representation using Feature Generation mode.

Caution 2: If you want to use GB-Score as a predictor, ligand and protein structures should be hydrogenated. Use Add Hydrogen mode for this. Then, features should be generated using $n=2$, $d_{cutoff}=12$ Å. Also, columns_pdbbind_2019.txt, mean_pdbbind_2019.csv, and std_pdbbind_2019.csv files should be used for Features name, Mean of features, and STD of features respectively.

If you want to use GB-Score, follow this Caution.

Saved model during training mode.

A ML saved model	?
Example/model/gb_score.joblib	
Features name	?
Example/files/columns_pdbbind_2019.txt	
Mean of features	?
Example/files/mean_pdbbind_2019.csv	
STD of features	?
Example/files/std_pdbbind_2019.csv	
Generated features of complex	?
Example/files/x_test_set.csv	
Prediction filename	?
pred_test_set.csv	
Experimental binding affinity value of test set.	?
None	

These files are generated during model training mode.

Use feature generation mode for this.

Output file name.

Start predicting operation

Click on this to start operation.

Experimental binding affinity file.

5-Normality Test

In this mode, if the test data has binding label, normality property of errors is analysed.

Investigating Normality of an Error Distribution

Theoretical Information

+

Caution: Your csv file should have two columns with the following names: Target and Predicted.

Loading data.

Load your data (.csv)

?

Example/files/prediction_test_set.csv

Shows data.

	Unnamed: 0	Target	Predicted
0	4llx	2.8900	3.55
1	5c28	5.6600	4.26
2	3uu0	7.9600	7.32
3	3ui7	9.0000	7.46
4	5c2h	11.0900	8.35
5	2v00	3.6600	4.70
6	3wz8	5.8200	6.20
7	3pww	7.3200	6.96
8	3prs	7.8200	7.41
9	<		>

Some information about data.

Mean of error: 0.058

Root mean square deviation (RMSD): 1.2

Mean absolute error (MAE): 0.925

Shapiro-Wilk (W) Test

Theoretical Information

+

W

0.9936

Pr

0.273

So, error distribution is normal ($Pr > 0.05$).

Skewness and Kurtosis Tests

Theoretical Information

+

Skew

-0.0149

Kurtosis

3.3

Error distribution is asymmetric and leptokurtic.

95th quantile of the absolute errors distribution

Theoretical Information

+

Q95

2.5812

Tests about
normality of
data.

Normal Quantile-Quantile Plot

Theoretical Information

+

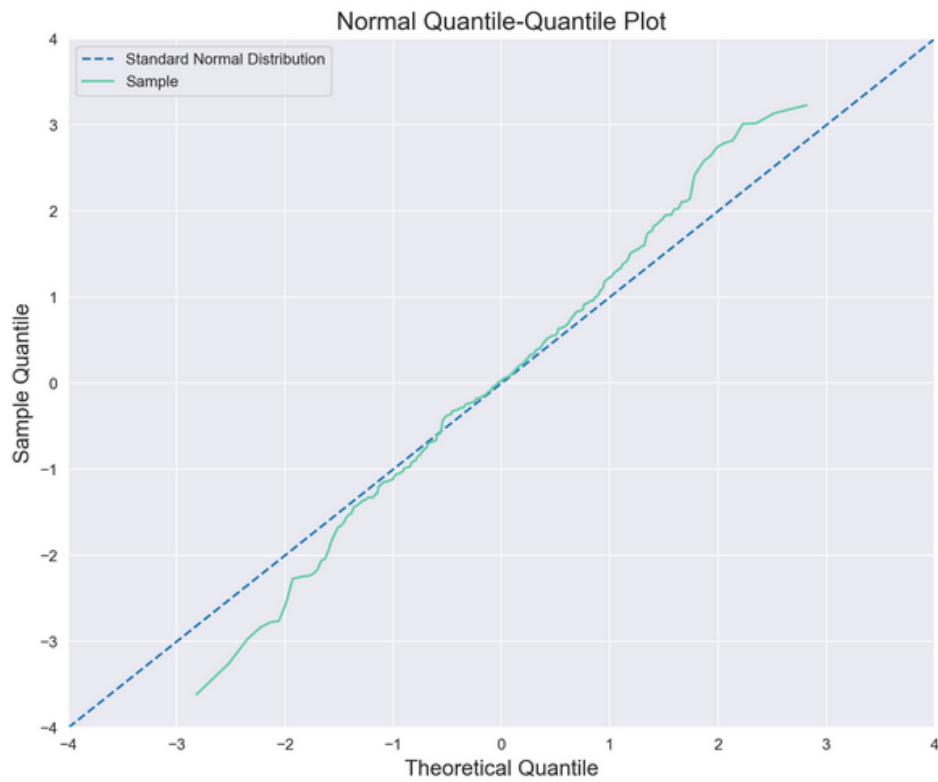
Save the plot and
change the color
of plot.

☐ Save QQ-plot in .png

Choose color for QQ-plot:



Show QQ-plot.



Caution: To download plot, first check *Save QQ-plot in .png* then the download button appears.

Histogram of Error Distribution

Theoretical Information

+

☐ Save histogram plot in .png.

Choose color for histogram plot:

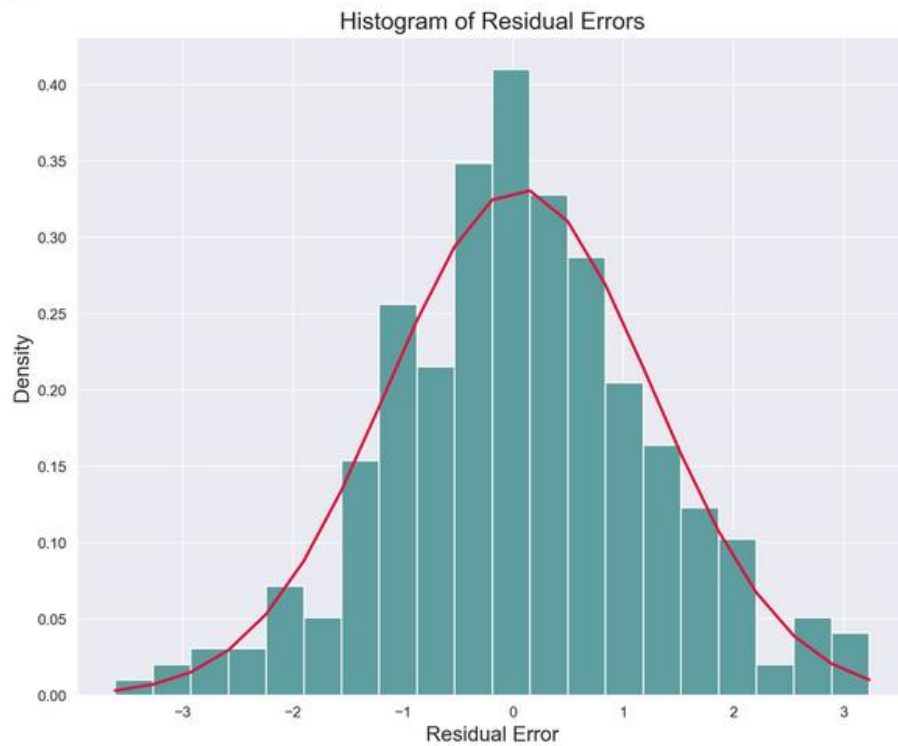


Choose color for gaussian plot:



Save and change the colors of plot.

Shows histogram of residual errors.



Caution: To download plot, first check *Save histogram plot in .png* then the download button appears.

Outliers Plot

Theoretical Information +

Choose a specific percentile



☐ Save outliers plot in .png

Choose color for non-outliers:



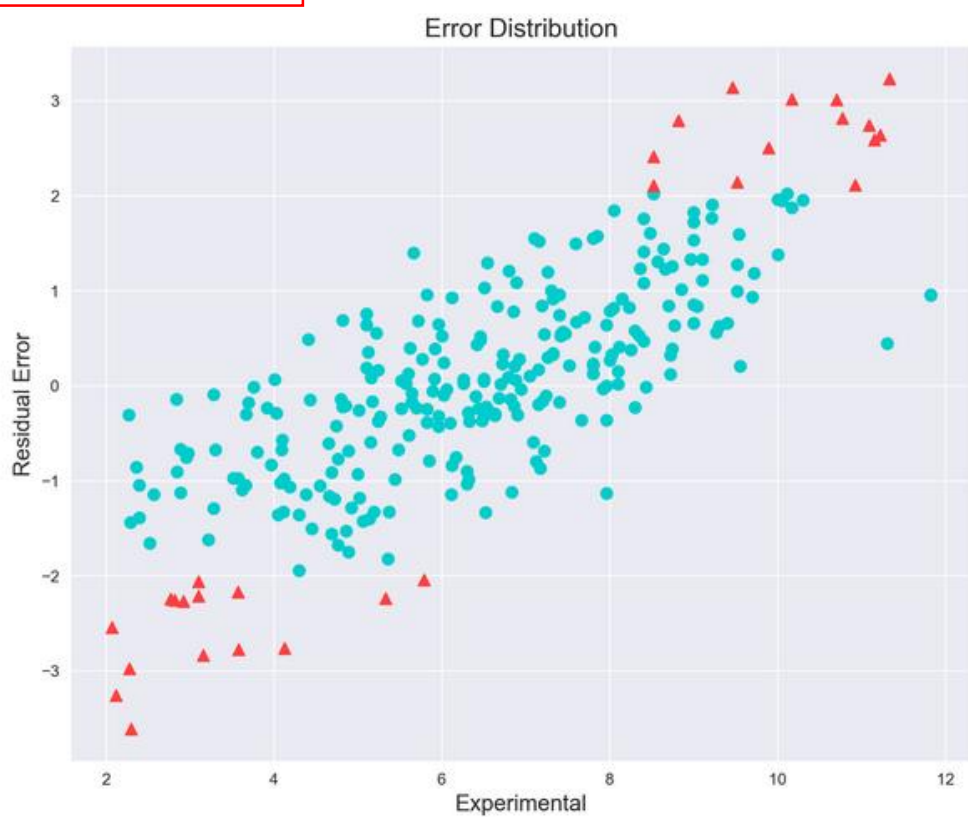
Choose color for outliers:



Save and change the colors of the plot.

Slider for choosing percentiles.

Outliers plot.



Caution: To download plot, first check *save outliers plot in .png* then the download button appears.

6-Add Hydrogen

Add hydrogens to ligand and protein at pH=7.4 using [PDB2PQR](#) and [Openbabel](#).

Caution about directory +

Enter directory of your complex structures: ?

Example/structures

Start operation

Click on this to start the operation.

Directory of structures files.