

LIFE Project Acronym and Number

ANTARES
LIFE08 ENV/IT/000435

Deliverable Report

Deliverable Name and Number

Deliverable 6
Evaluation on the model performances for the first model
(preliminary report)

Deliverable Date

31/12/2010

Deliverable Data

Associated action	Action 4: List of (Q)SAR models for the ecotoxicological, toxicological, and environmental endpoints for REACH and their review
Beneficiary Responsible	KnowledgeMiner Software
Contact person	Frank Lemke

Introduction

This document summarizes performance results of 21 regression models for the **Bioconcentration factor** in fish. The performance results reported here are based on average experimental log BCF values (y -values) of 701 chemical compounds and corresponding predicted log BCF values of a reported model (\hat{y} -values). This means that the performance tests were run in the output data space, only. This is because this information has been available to us for all considered models as a common basis to run comparative performance tests, while reliable information about the input (descriptor) data used to develop and test a model is only published rarely. Nonetheless, the performance tests we run for each model are important and essential for finding appropriate model candidates to be used in regulatory procedures.

Performance Tests

This section gives a brief description about the performed tests and their resulting scores applied.

Sample size

A model that works on a large number of chemical compounds is more likely being reliable than a model that predicts only few compounds. Also, implementing a QSAR model into REACH regulation would require a model to work on hundreds, thousands, or even ten thousands of compounds as reliable as possible. To take this into account the first score reflects **sample size**:

$$S_N = n / N,$$

with n – number of predicted compounds published for the model or the model works on with some plausibility, and $N = 701$ – number of experimental values available to any model for these performance tests.

Linear regression between predicted and experienced values

This test provides **basic model properties** like model **accuracy** (R^2), **slope** a , and **intercept** b of a linear regression $y = a\hat{y} + b$ on both the *design data* (blue colored information in the graph) the model was developed on and on out-of-sample *test data* (green colored information in the graph). The number of compounds in the design and in the test data set varies from model to model. In cases where the splitting into design and test data was not published or uncertain, a random, roughly 70/30 split of the data was done instead.

Having three basic properties for design and independent out-of-sample data sets, it is possible to get an indication of the generalization power or “robustness” of a model. The less the model properties on the two data sets differ the more robust a model appears.

In result, there are 6 values, 3 static and 3 robustness values, which are compared with the ideal reference model values providing 6 subscores S_i . These 6 subscores are then summarized to a single score S_p

representing basic model properties using this formula:

$$S_P = (3S_1 + 2S_2 + S_3 + S_4 + 0,5S_5 + 0,5S_6) / 8,$$

with S_1 – accuracy on design data, S_2 – accuracy robustness, S_3 – slope on design data, S_4 – slope robustness, S_5 – intercept on design data, S_6 – intercept robustness.

Corresponding to the used linear regression above, $y = a\hat{y} + b$, the regression plot shows the experienced values on the y-axis and the predicted values of the x-axis. We did not consider the reverse case (predicted vs. experienced), $\hat{y} = cy + d$, since from the relation $ac = R^2$ and $a = 1,0$ on the design data, for reference, it follows $c = R^2$ and $d \neq 0$. This provides no new information and can be omitted, therefore.

Over-/underestimation tendencies of a model

For regulatory purposes it makes a huge difference whether a model rather over- or underestimates experienced values. Overestimation is clearly the preferred and acceptable behavior of a QSAR model from a regulators perspective. Since experimental values, which the models are compared to, usually have huge uncertainty themselves this leads to the concept of False Positives and False Negatives known in classification modeling. When assigning an uncertainty ε to the experimental values y an uncertainty interval $y - \varepsilon < y < y + \varepsilon$ is established the predicted values can be compared with. If a predicted value is within this uncertainty interval it is seen as correct prediction. If it is below the lower bound of the uncertainty interval, $y - \varepsilon$, it underestimates the experienced value and is classified as “**False Negative**” (FN). In the opposite case, when the prediction is higher than the upper bound of the interval, $y + \varepsilon$, it overestimates the experimental value and it is “False Positive” (FP). In our tests, we used a rather very small uncertainty of $\pm 10\%$, $\varepsilon = 0,1y$, given that uncertainty of experimental values can easily reach 100% and more. The score for this test is calculated in this way to satisfy the above mentioned regulatory requirements:

$$S_{FN} = 1 - (2FN/n + FP/n) / 3,$$

with n – number of predicted compounds available in the data set for the given model (sample size).

It is possible to optimize the false negative rate by shifting a model by a constant value μ , which changes the intercept $b' = b - \mu$ of the regression equation accordingly. In a second step, we also calculated the optimal false negative score S_{FNopt} for a model based on shifting the entire model by μ_{opt} . This is reported in the results summary for comparison, only.

Error decomposition

The squared model error between experienced and predicted values, $e = (y - \hat{y})^2$, is decomposed into three components:

$$1 = (e_N + e_S + e_B) / e,$$

with e_N – the error rate described by noise in the data, e_S – error rate that describes the deviation of the regression slope from the reference slope of 1, and e_B – rate that expresses how much of the model error is



due to model bias. The **error decomposition** score for this test is calculated accordingly:

$$S_E = (e_N + 2 \cdot e_S - e_B) / 3e.$$

Y-scrambling

One method to check if and to which extent a model is different from a pure chance model is so-called Y-scrambling. Here, the experienced log BCF values are reordered randomly m times and the resulting model accuracy R^2 is measured for every simulation run ($m = 30$ was used in this test). If a model is not just reflecting a random relationship, the observed R^2 is zero, ideally. For each simulation, we also decomposed the error $e = 1 - R^2$ into its components again. For the noise and slope deviation components, e_N and e_S , the same relation as for R^2 is true: the larger the difference between scrambled and original results the more likely the model is not a chance model. The bias component e_B should not change, ideally.

This leads to this **Y-scrambling** score formula:

$$S_S = (\Delta_A + \Delta_N + \Delta_S + 0,25\Delta_B) / 3,25,$$

with

$$\begin{aligned}\Delta_A &= R_0^2 - \frac{1}{m} \sum_{i=1}^m R_i^2 \\ \Delta_N &= 2 \left| \frac{1}{m} \sum_{i=1}^m e_{N,i} - e_{N,0} \right| \\ \Delta_S &= 2 \left| \frac{1}{m} \sum_{i=1}^m e_{S,i} - e_{S,0} \right| \\ \Delta_B &= 1 - \left| e_{B,0} - \frac{1}{m} \sum_{i=1}^m e_{B,i} \right|\end{aligned}$$

where an index of 0 indicates the original, un-scrambled values, correspondingly.

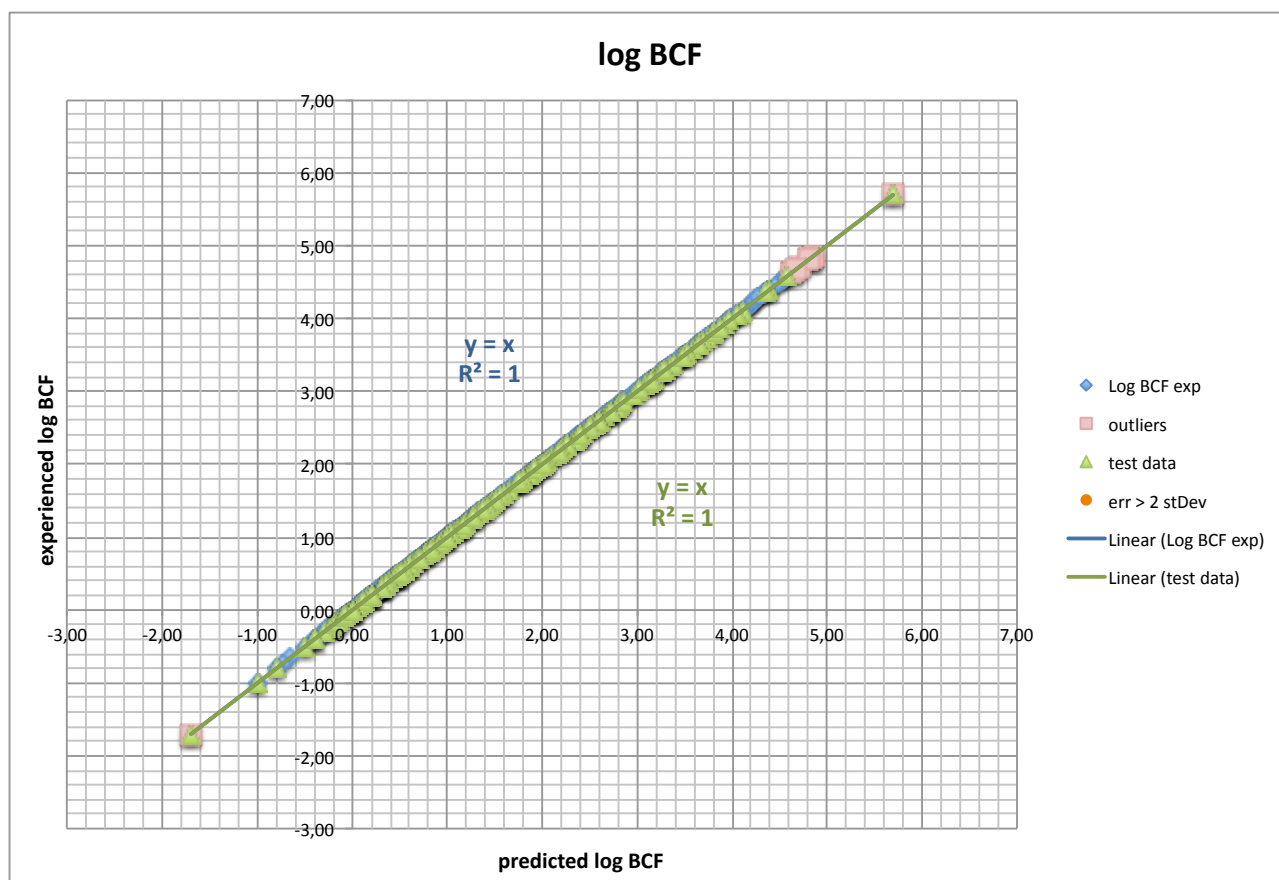


Performance Results

1. Reference Model

The reference graphics and scores for the ideal model are shown below.

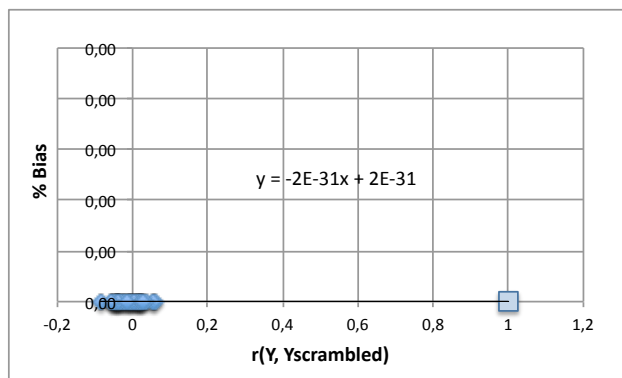
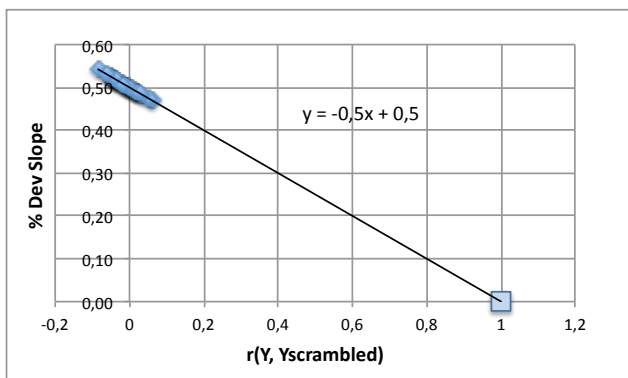
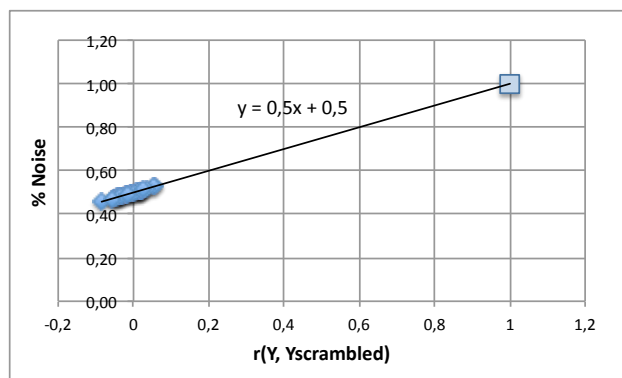
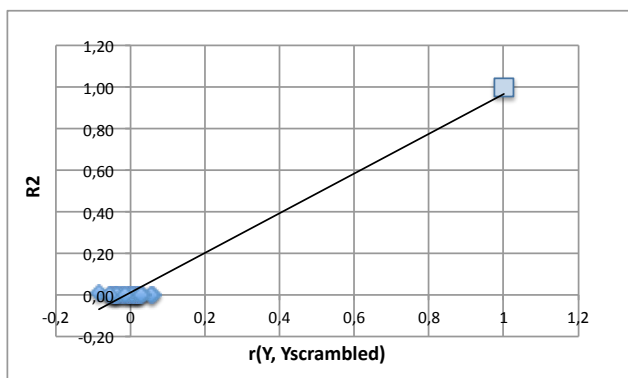
Basic model properties



SCORES

Sample Size: 701	1,00	FN: 0 % FP: 0 %	1,00
Accuracy: 1,0	Slope: 1,0	Intercept: 1,0	1,00
Accuracy Robustness: 1,0	Slope Robustness: 1,0	Intercept Robustness: 1,0	

Error decomposition and Y-scrambling



SCORES

Error Decomposition

Noise: 1,0

Deviation Slope: 0,0

Bias: 0,0

1,00

Y-Scrambling

Accuracy: 1,0

Noise: 1,0

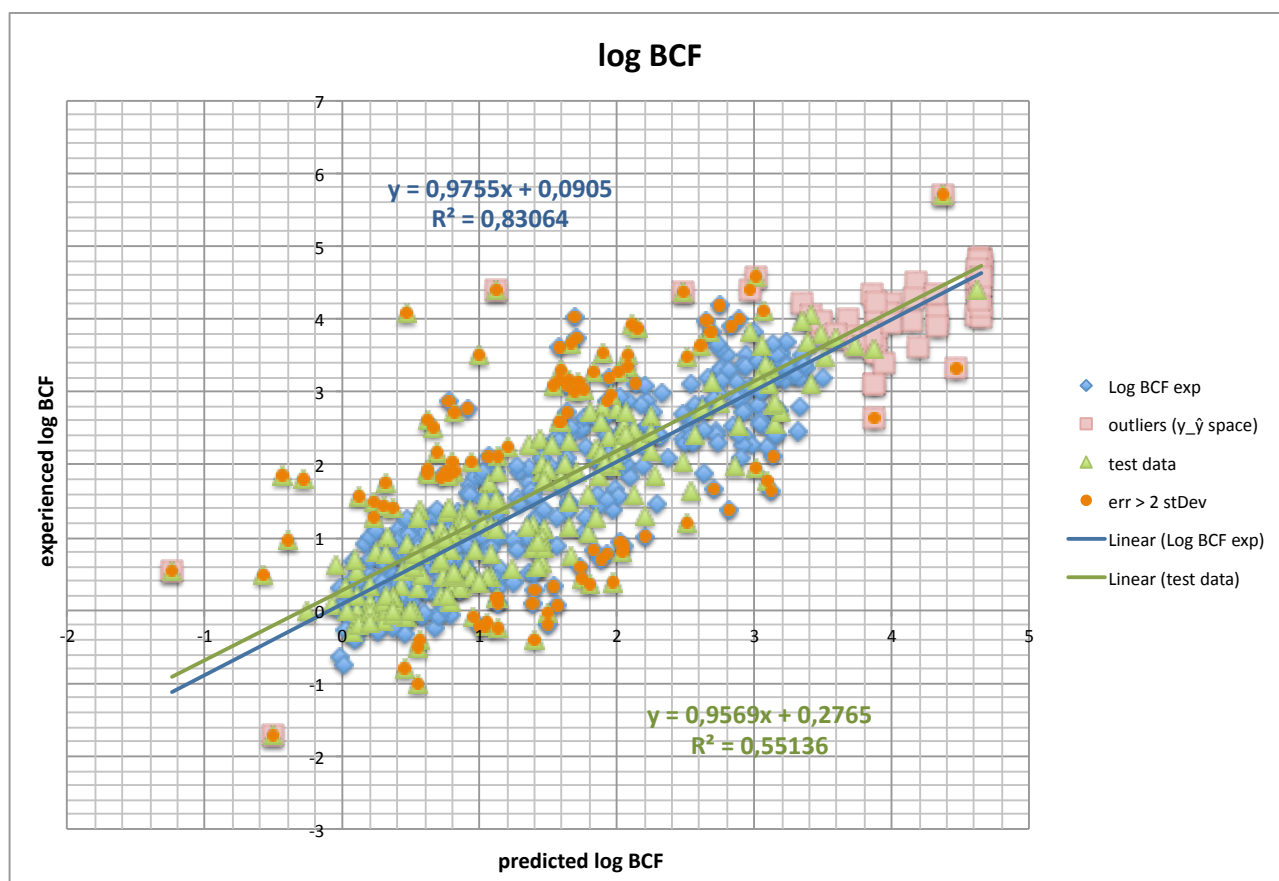
Deviation Slope: 1,0

Bias: 1,0

1,00

2. CAESAR Model

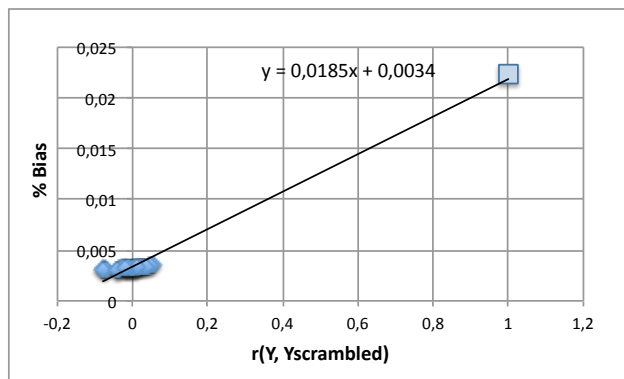
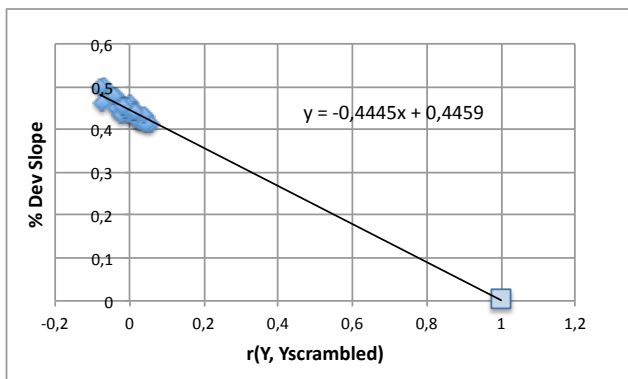
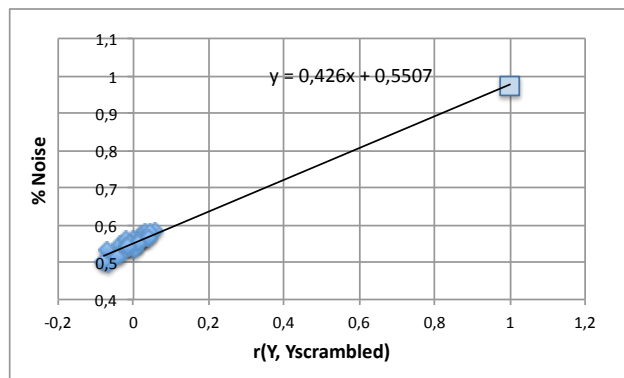
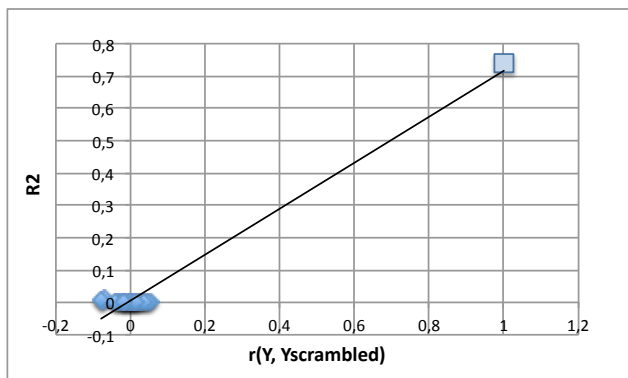
Basic model properties



SCORES

Sample Size: 701	1,00	FN: 42 % FP: 36 %	0,60
Accuracy: 0,74	Slope: 0,97	Intercept: 0,82	0,79
Accuracy Robustness: 0,66	Slope Robustness: 0,99	Intercept Robustness: 0,82	

Error decomposition and Y-scrambling



SCORES

Error Decomposition

Noise: 0,97

Deviation Slope: 0,0

Bias: 0,02

0,98

Y-Scrambling

Accuracy: **0,74**

Noise: **0,85**

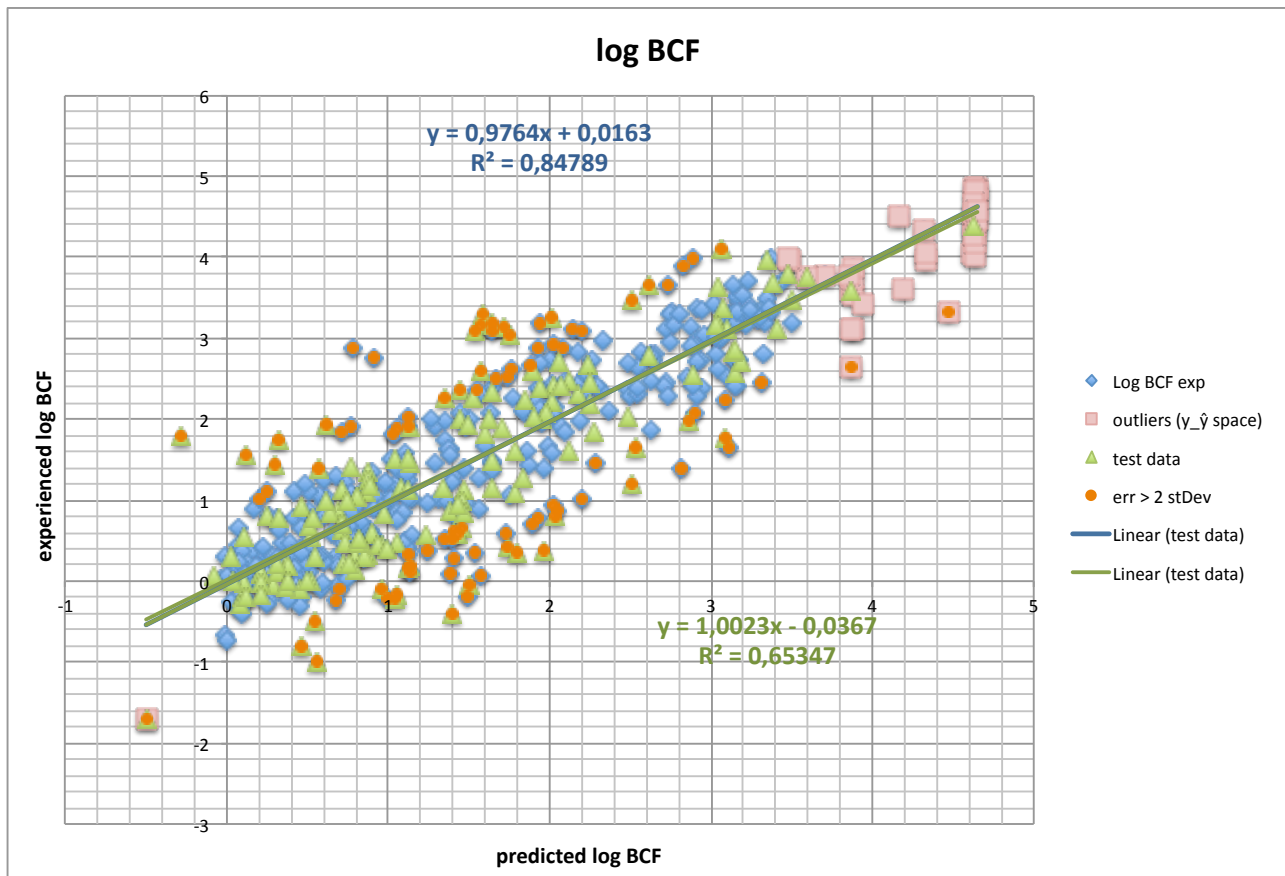
Deviation Slope: **0,89**

Bias: **0,98**

0,84

3. CAESAR valid

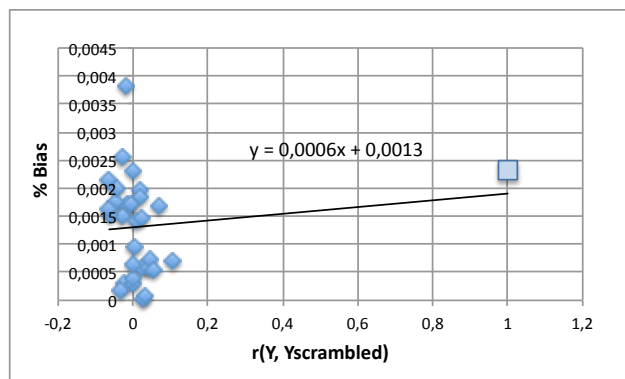
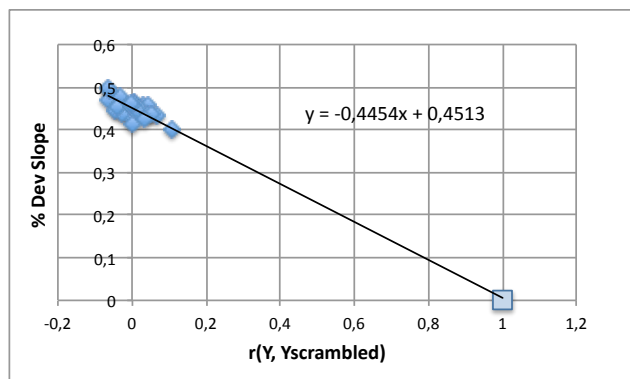
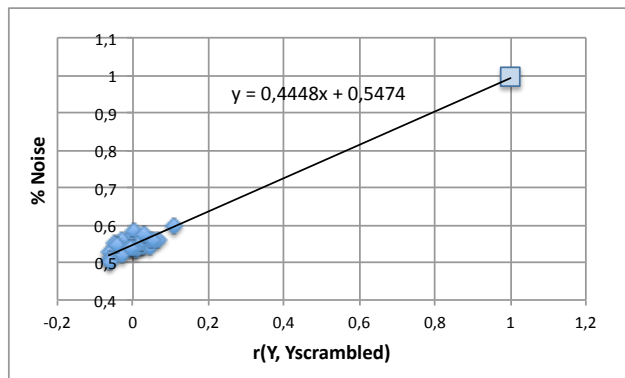
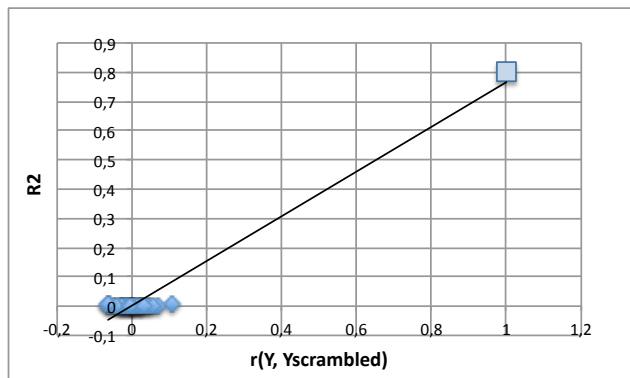
Basic model properties



SCORES

Sample Size: 531	0,76	FN: 34 % FP: 43 %	0,63
Accuracy: 0,80	Slope: 0,99	Intercept: 0,99	0,86
Accuracy Robustness: 0,76	Slope Robustness: 0,98	Intercept Robustness: 0,96	

Error decomposition and Y-scrambling



SCORES

Error Decomposition

Noise: 0,996

Deviation Slope: 0,0

Bias: 0,0

1,00

Y-Scrambling

Accuracy: 0,80

Noise: 0,90

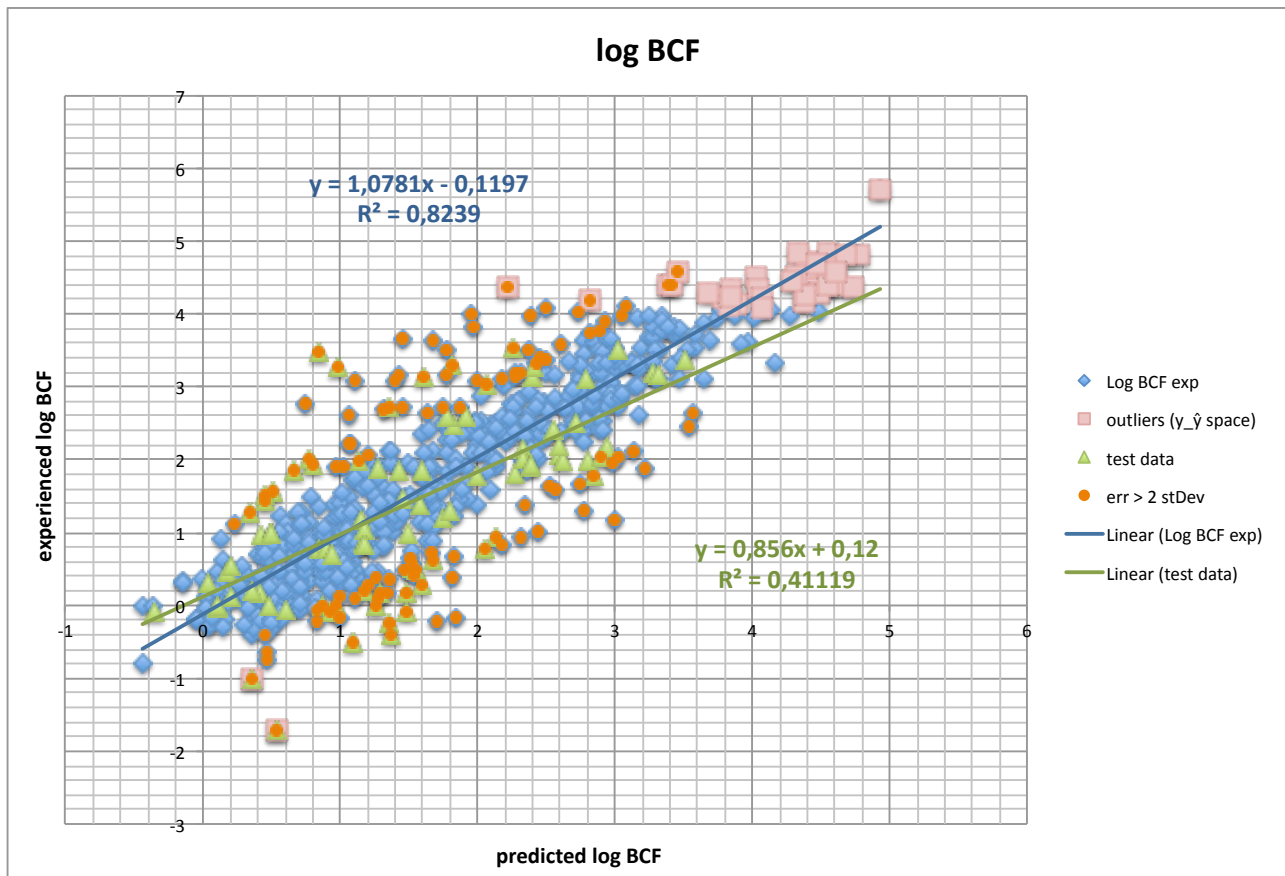
Deviation Slope: 0,90

Bias: 1,00

0,88

4. TEST Model

Basic model properties

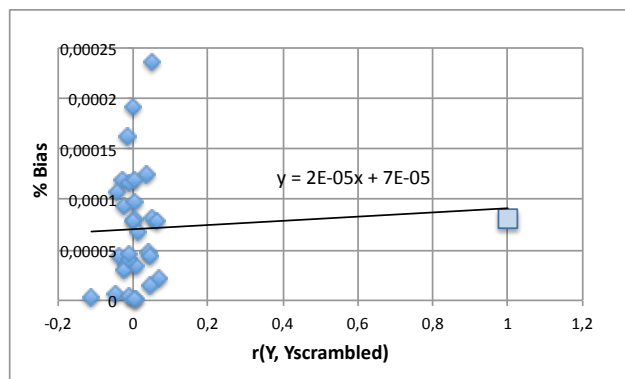
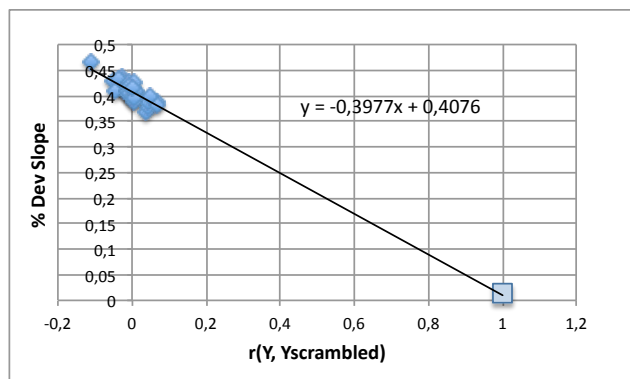
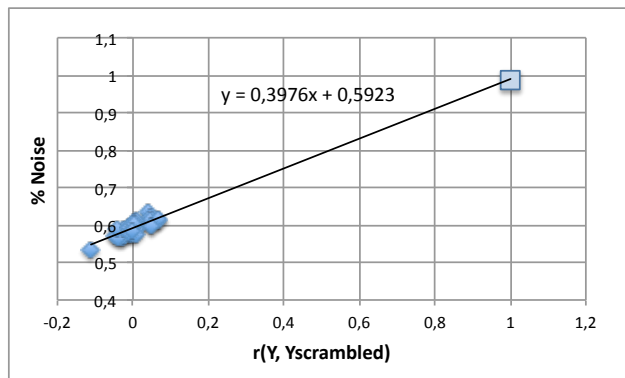
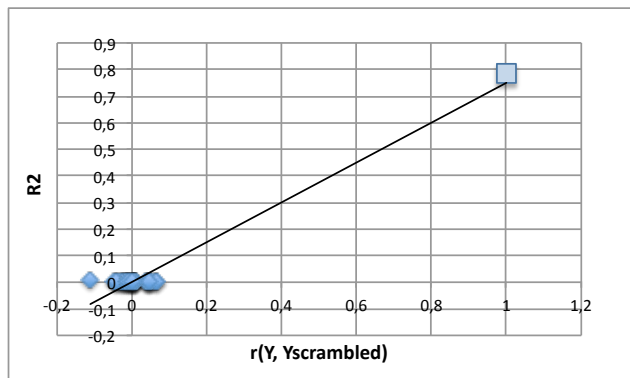


SCORES

Sample Size: 686	0,98	FN: 35 % FP: 41 %	0,63
Accuracy: 0,78	Slope: 0,97	Intercept: 1,00	0,75
Accuracy Robustness: 0,50	Slope Robustness: 0,80	Intercept Robustness: 0,76	



Error decomposition and Y-scrambling



SCORES

Error Decomposition

Noise: 0,99

Deviation Slope: 0,01

Bias: 0,0

0,99

Y-Scrambling

Accuracy: **0,78**

Noise: **0,79**

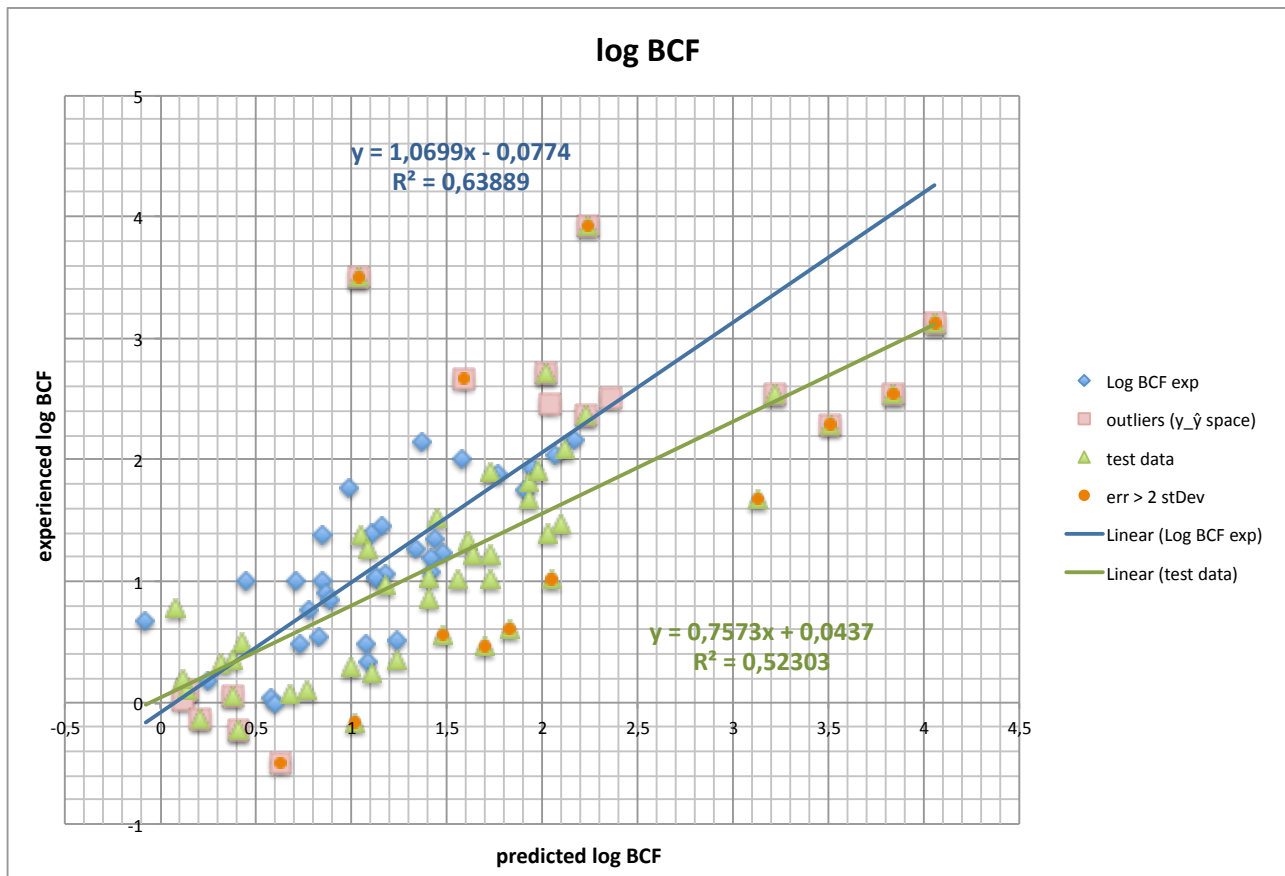
Deviation Slope: **0,79**

Bias: 1,00

0,80

5. Fu Acid Model

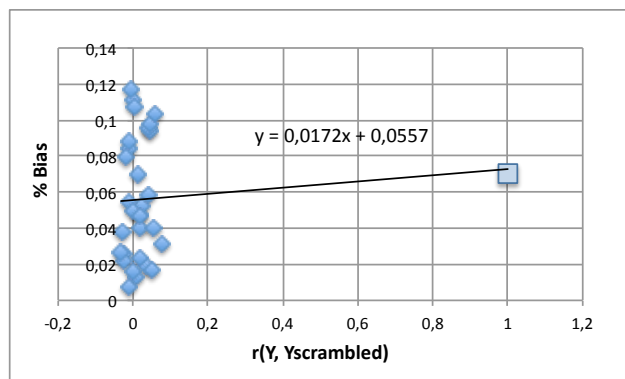
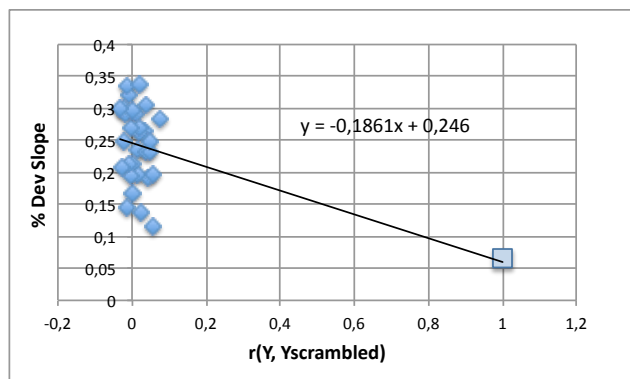
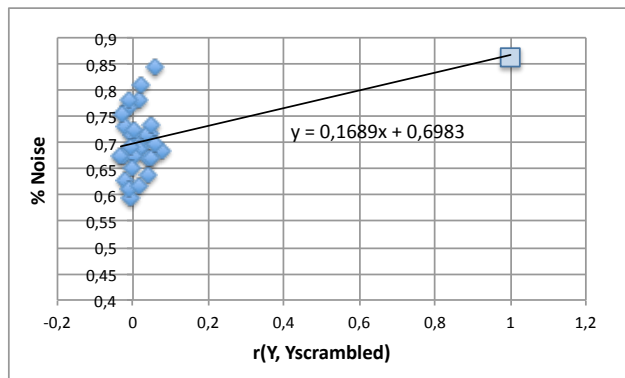
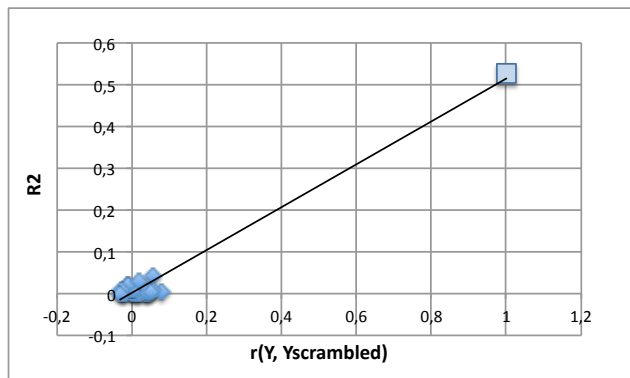
Basic model properties



SCORES

Sample Size: 85	0,12	FN: 24 % FP: 52 %	0,67
Accuracy: 0,53	Slope: 0,91	Intercept: 0,99	0,72
Accuracy Robustness: 0,81	Slope Robustness: 0,70	Intercept Robustness: 0,89	

Error decomposition and Y-scrambling



SCORES

Error Decomposition

Noise: 0,86

Deviation Slope: 0,07

Bias: 0,07

0,91

Y-Scrambling

Accuracy: 0,52

Noise: 0,33

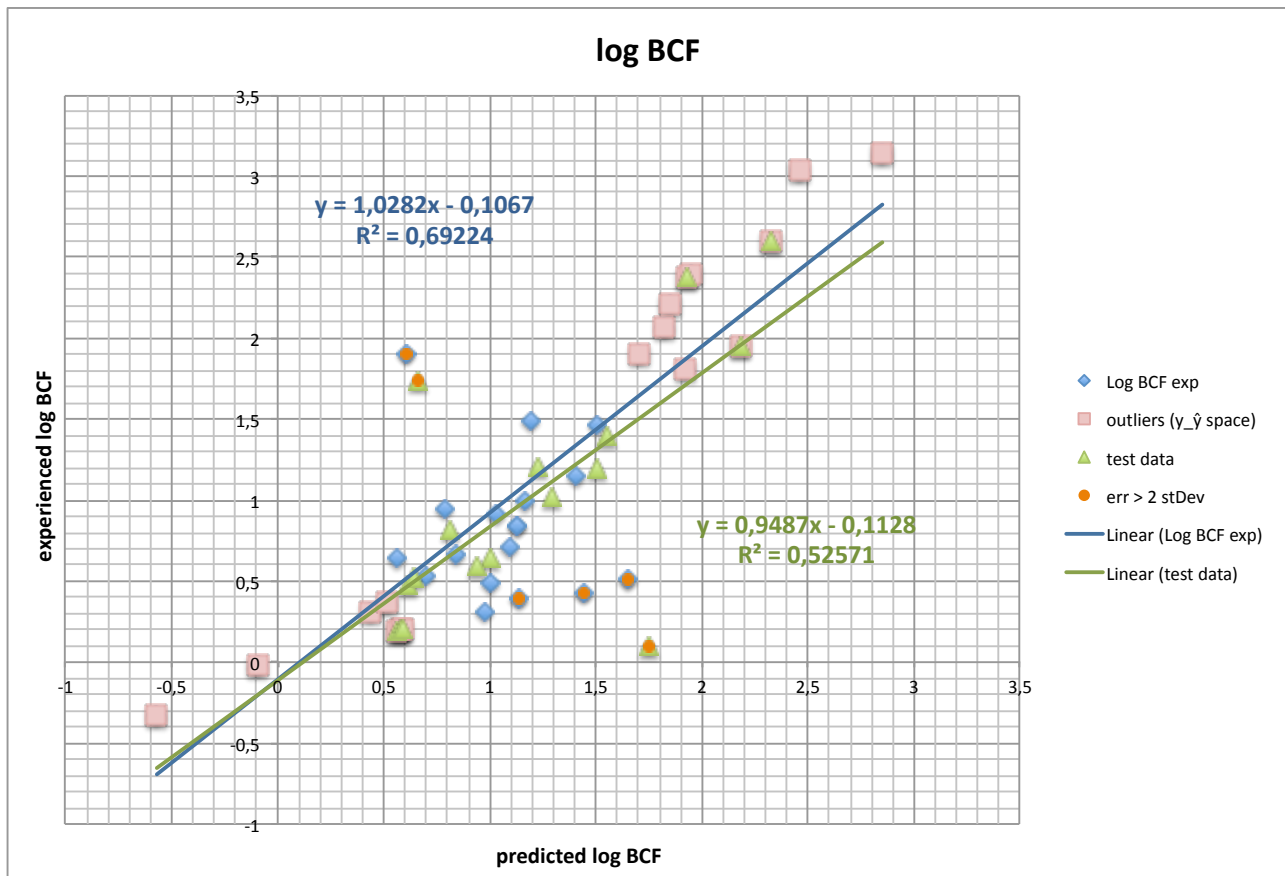
Deviation Slope: 0,35

Bias: 0,99

0,44

6. Fu Base Model

Basic model properties

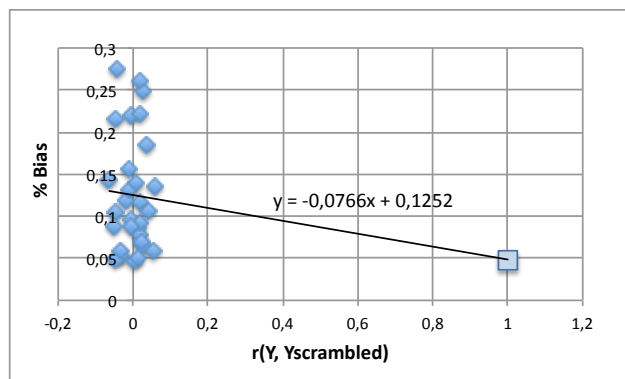
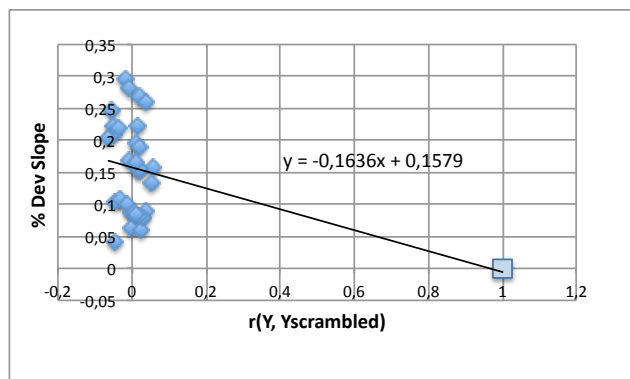
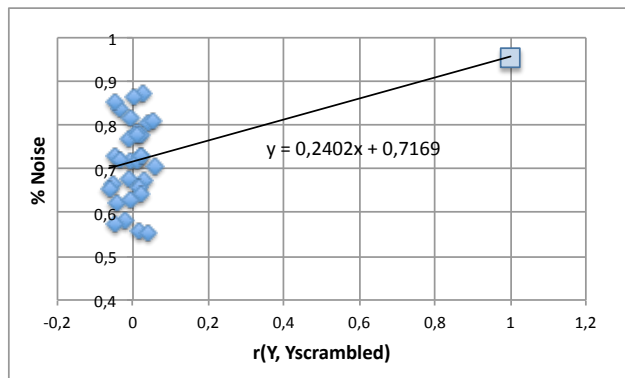
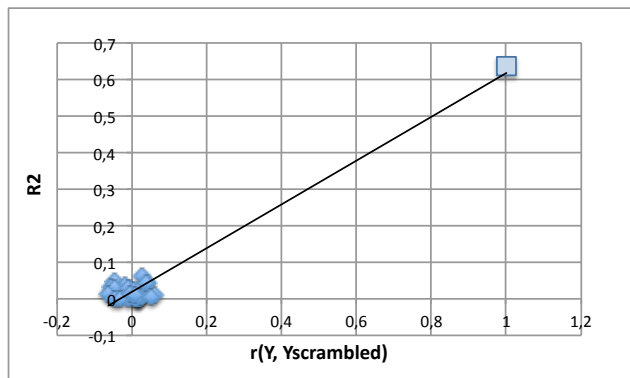


SCORES

Sample Size: 46	0,07	FN: 30 % FP: 59 %	0,60
Accuracy: 0,64	Slope: 0,99	Intercept: 0,90	0,78
Accuracy Robustness: 0,75	Slope Robustness: 0,92	Intercept Robustness: 0,99	

Alternative Non-Testing methods Assessed for REACH Substances

Error decomposition and Y-scrambling



SCORES

Error Decomposition

Noise: 0,95

Deviation Slope: 0,0

Bias: 0,05

0,97

Y-Scrambling

Accuracy: **0,62**

Noise: **0,47**

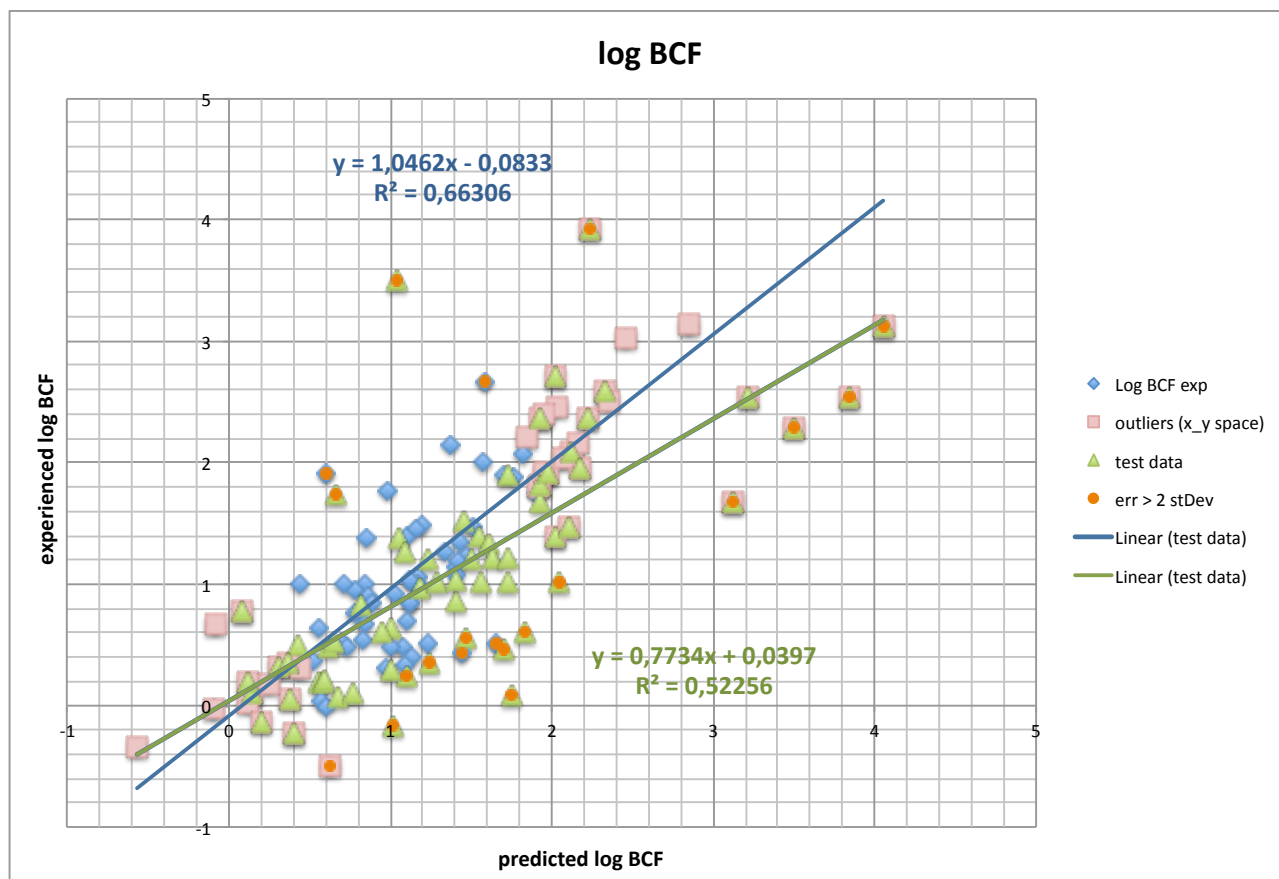
Deviation Slope: **0,32**

Bias: **0,92**

0,50

7. Fu Model

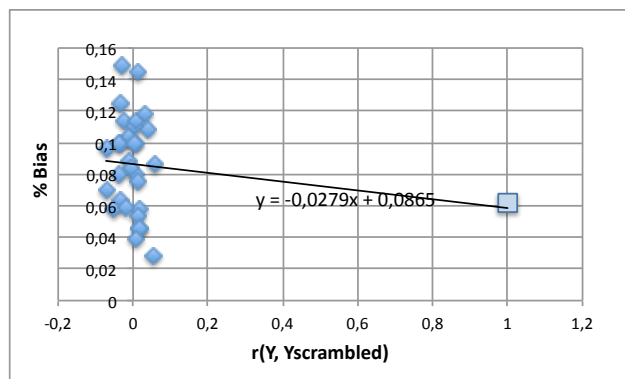
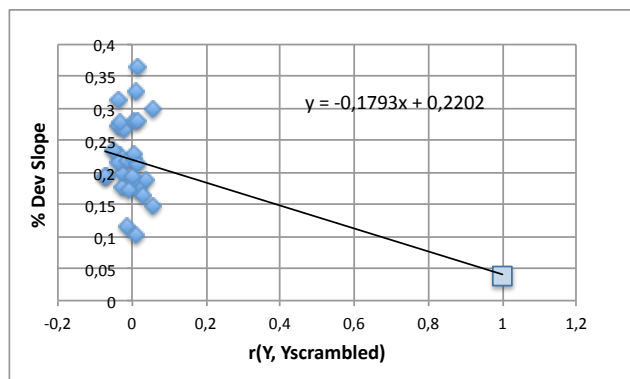
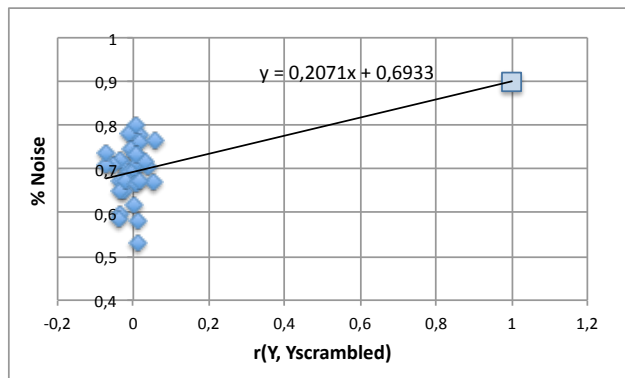
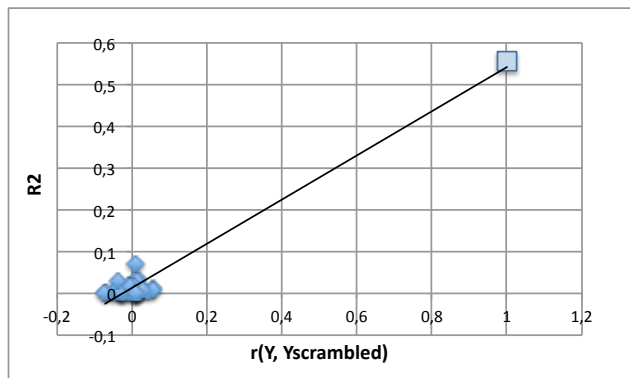
Basic model properties



SCORES

Sample Size: 131	0,19	FN: 26 % FP: 54 %	0,67
Accuracy: 0,56	Slope: 0,91	Intercept: 0,98	0,73
Accuracy Robustness: 0,79	Slope Robustness: 0,73	Intercept Robustness: 0,88	

Error decomposition and Y-scrambling



SCORES

Error Decomposition

Noise: 0,90

Deviation Slope: 0,04

Bias: 0,06

0,93

Y-Scrambling

Accuracy: 0,55

Noise: 0,41

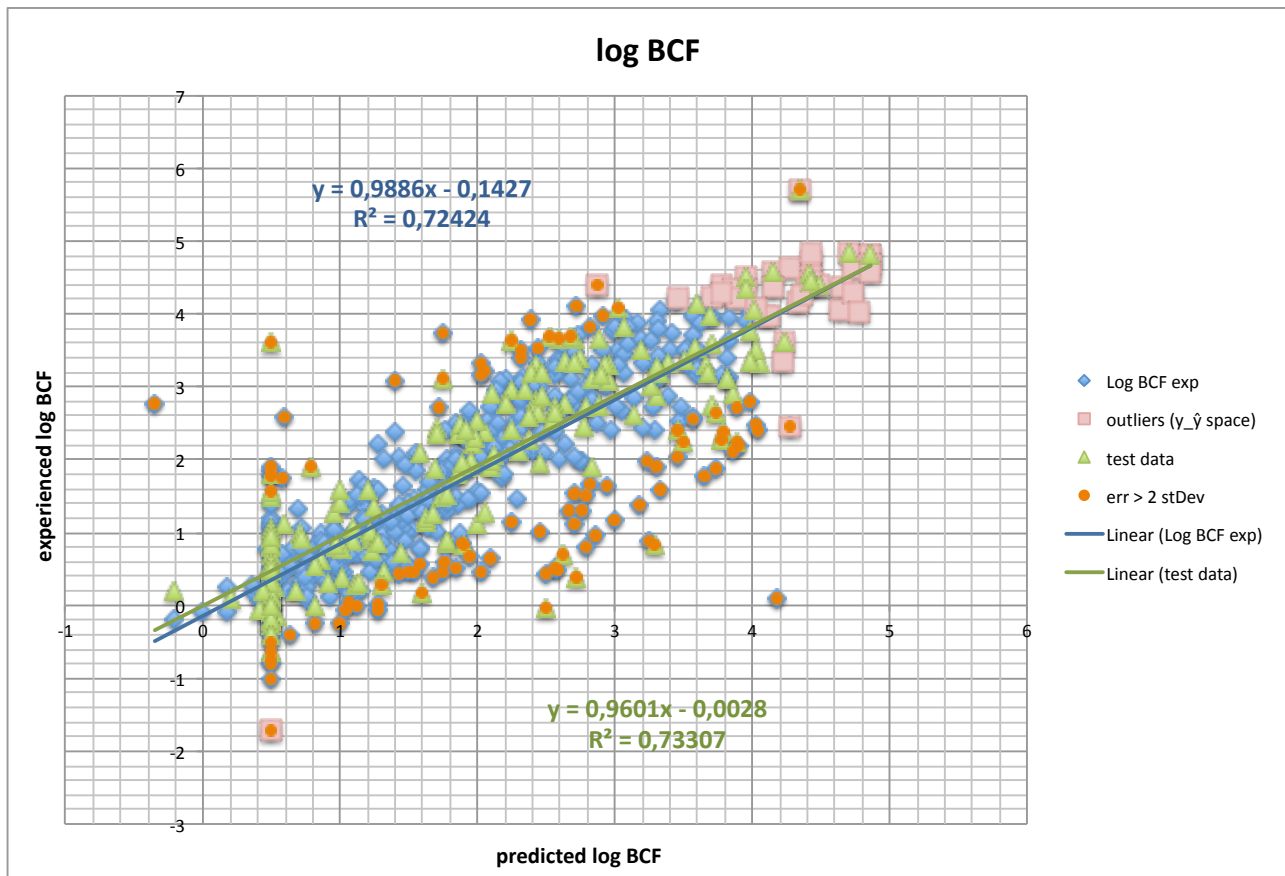
Deviation Slope: 0,37

Bias: 0,98

0,48

8. Episuite Regression Model

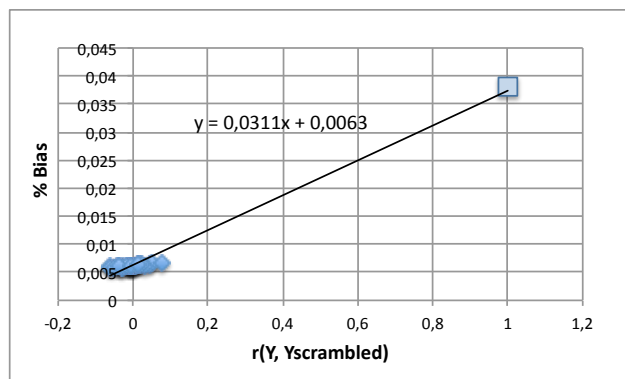
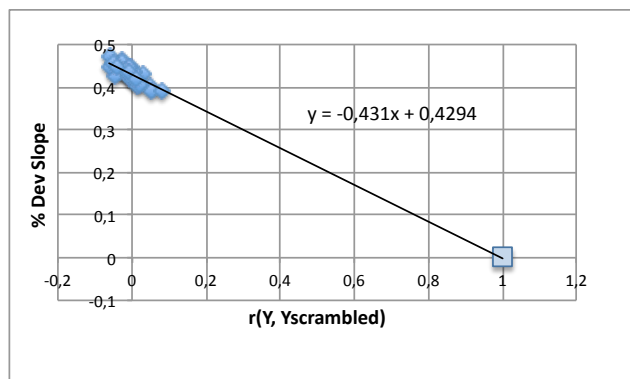
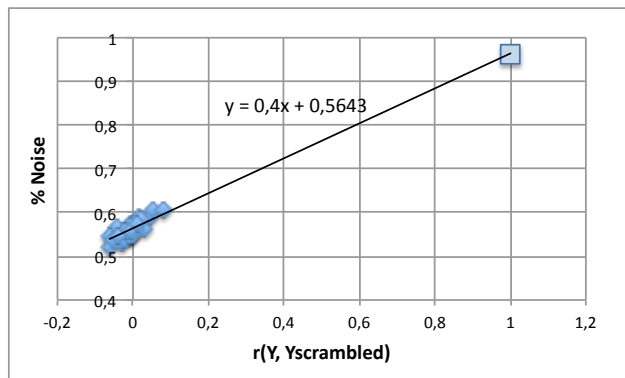
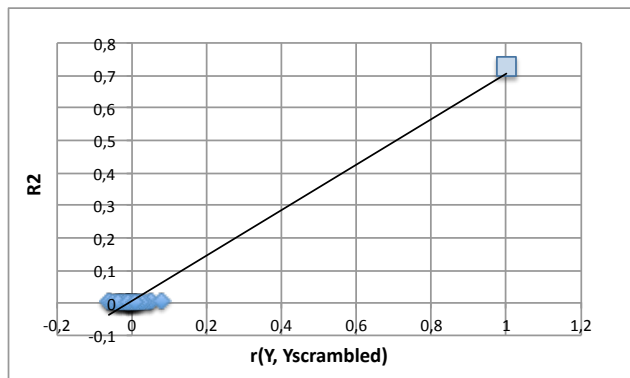
Basic model properties



SCORES

Sample Size: 701	1,00	FN: 28 % FP: 50 %	0,65
Accuracy: 0,73	Slope: 0,97	Intercept: 0,93	0,87
Accuracy Robustness: 0,99	Slope Robustness: 0,97	Intercept Robustness: 0,86	

Error decomposition and Y-scrambling



SCORES

Error Decomposition

Noise: 0,96

Deviation Slope: 0,0

Bias: 0,04

0,97

Y-Scrambling

Accuracy: **0,73**

Noise: **0,80**

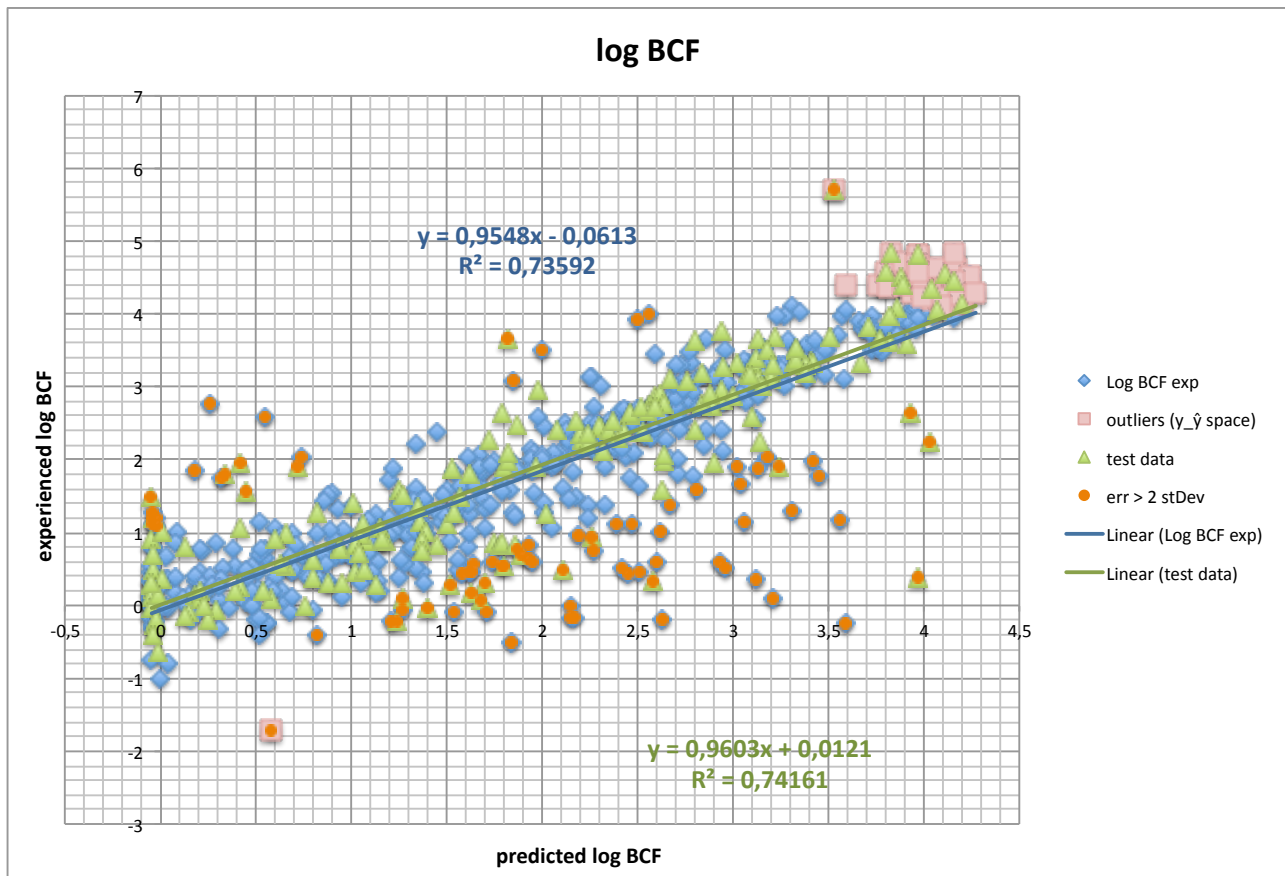
Deviation Slope: **0,86**

Bias: **0,97**

0,81

9. Episuite Arnod-Gobas Up Model

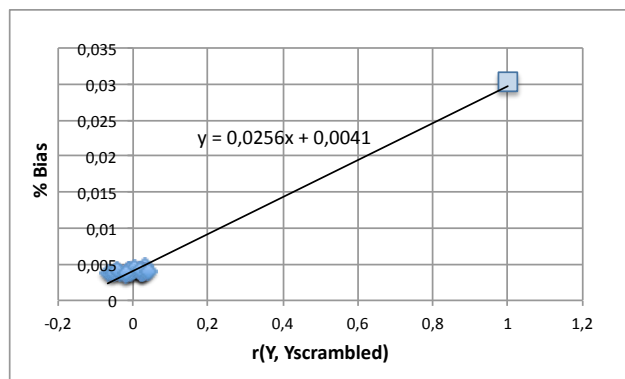
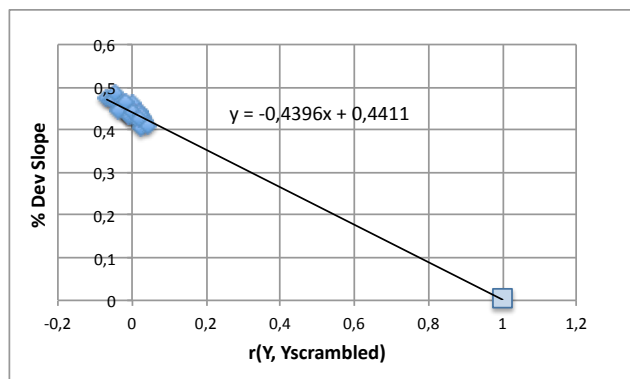
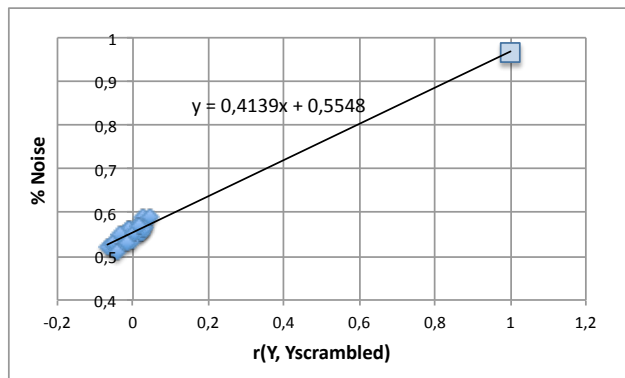
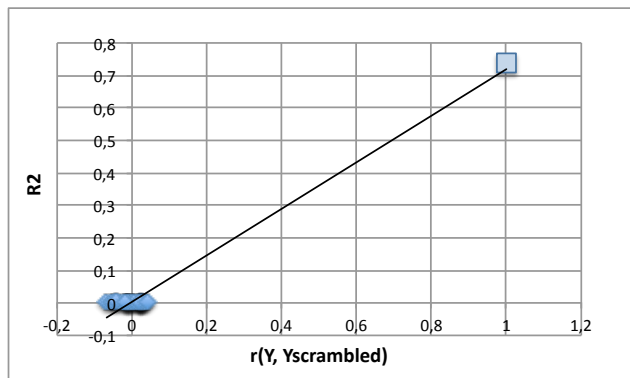
Basic model properties



SCORES

Sample Size: 697	0,99	FN: 25 % FP: 39 %	0,70
Accuracy: 0,74	Slope: 0,96	Intercept: 1,00	0,89
Accuracy Robustness: 0,99	Slope Robustness: 0,99	Intercept Robustness: 0,99	

Error decomposition and Y-scrambling



SCORES

Error Decomposition

Noise: 0,97

Deviation Slope: 0,0

Bias: 0,03

0,98

Y-Scrambling

Accuracy: **0,74**

Noise: **0,82**

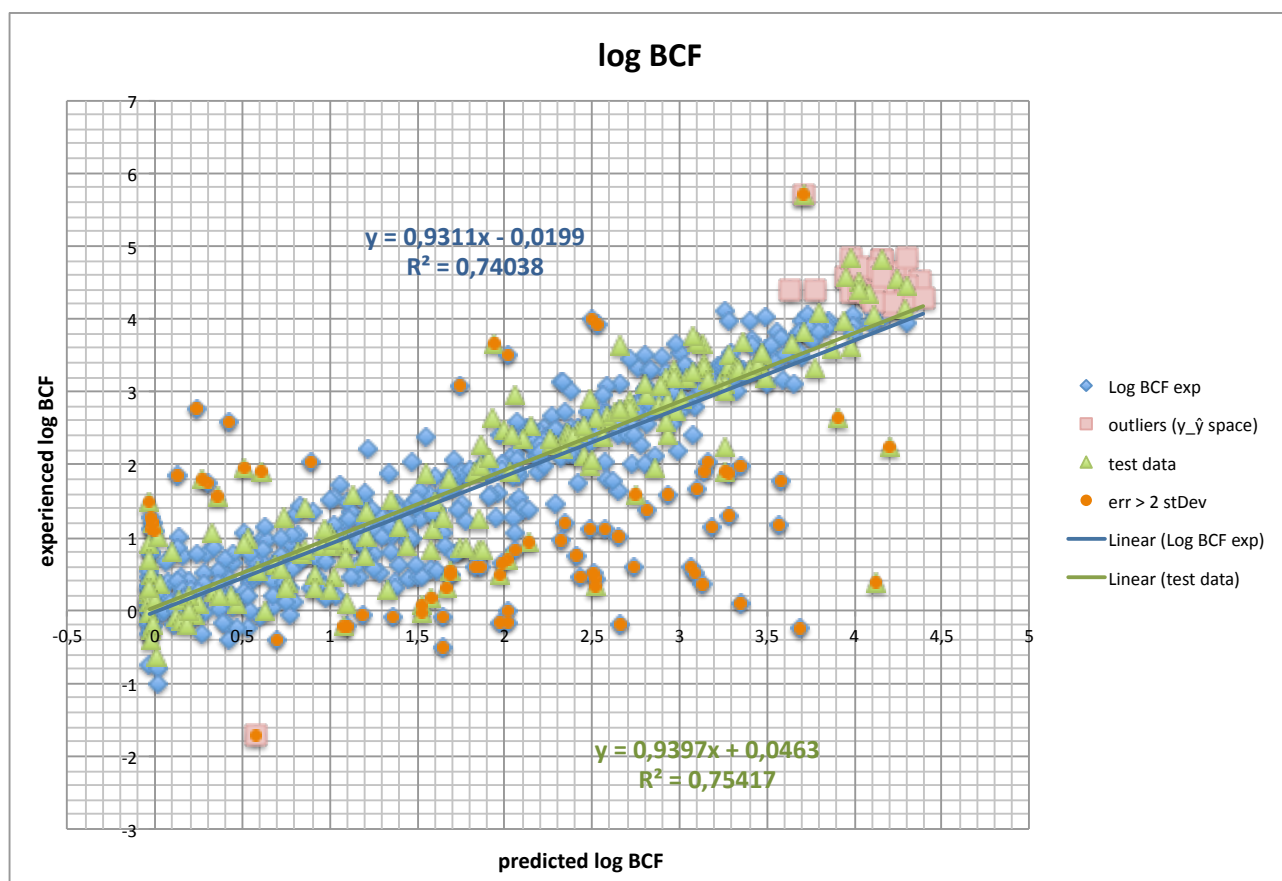
Deviation Slope: **0,87**

Bias: **0,97**

0,82

10. Episuite Arnod-Gobas Mid Model

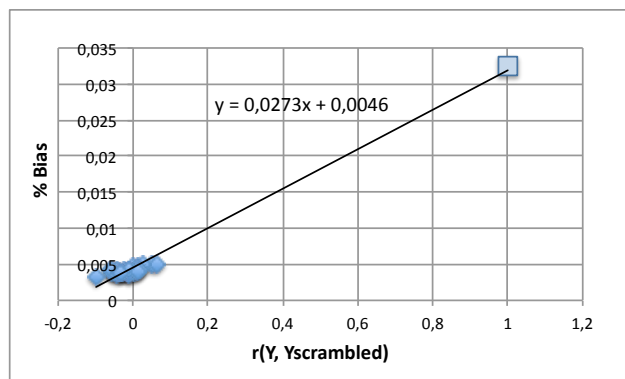
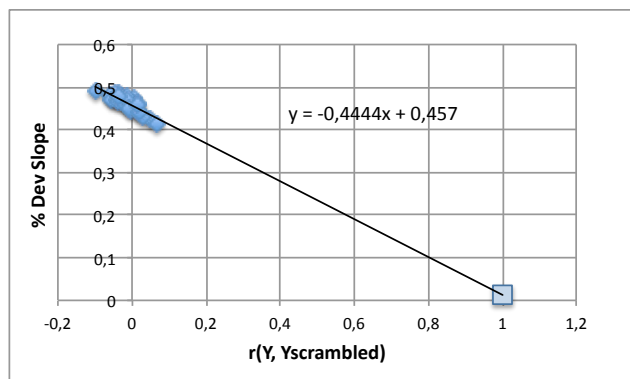
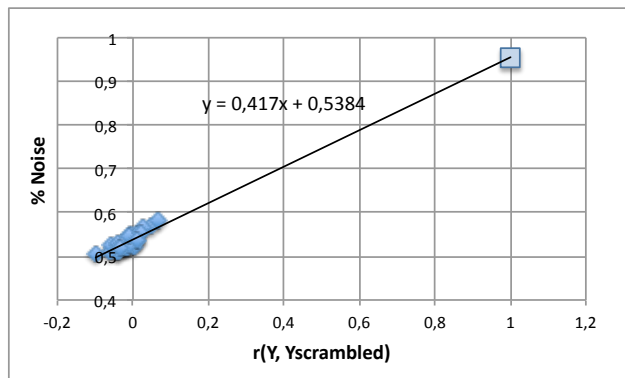
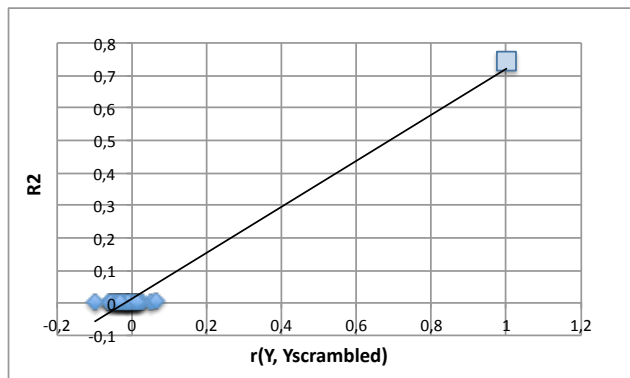
Basic model properties



SCORES

Sample Size: 697	0,99	FN: 26 % FP: 38 %	0,70
Accuracy: 0,74	Slope: 0,94	Intercept: 0,99	0,89
Accuracy Robustness: 0,99	Slope Robustness: 0,99	Intercept Robustness: 0,93	

Error decomposition and Y-scrambling



SCORES

Error Decomposition

Noise: 0,95

Deviation Slope: 0,01

Bias: 0,04

0,97

Y-Scrambling

Accuracy: **0,74**

Noise: **0,84**

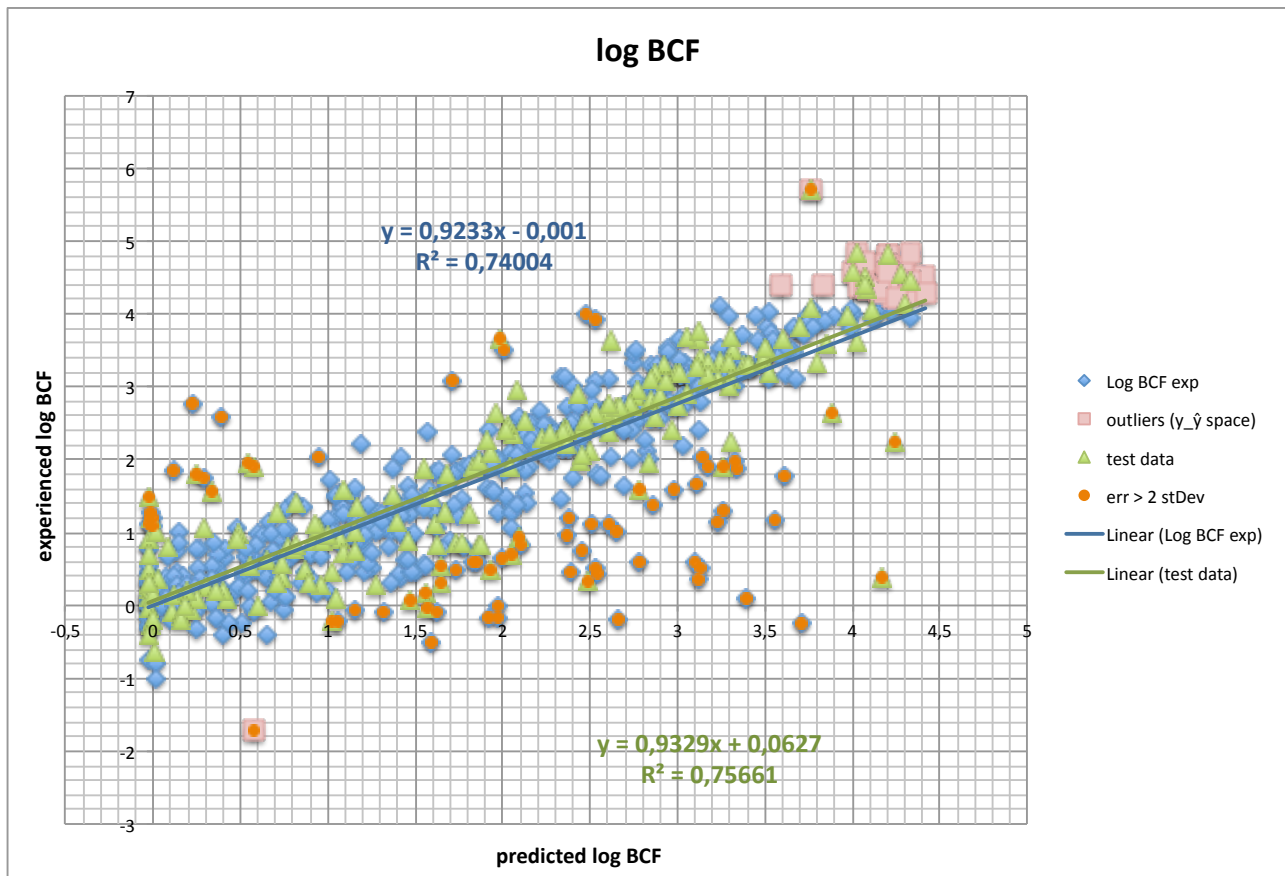
Deviation Slope: **0,90**

Bias: **0,97**

0,84

11. Episuite Arnod-Gobas Low Model

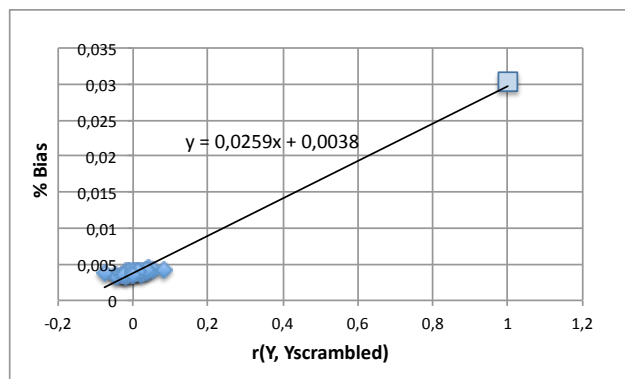
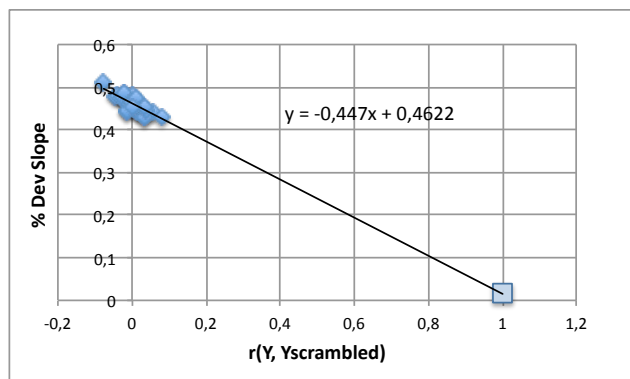
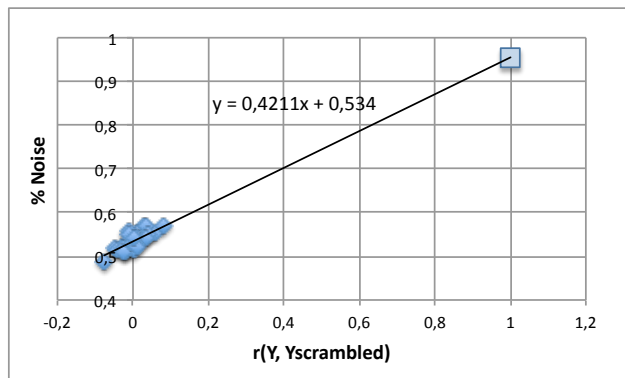
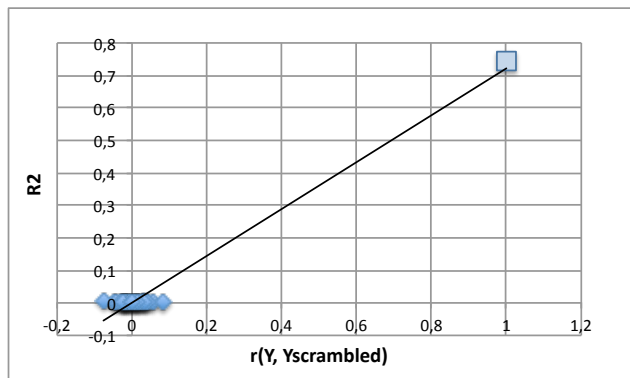
Basic model properties



SCORES

Sample Size: 697	0,99	FN: 27 % FP: 37 %	0,70
Accuracy: 0,74	Slope: 0,93	Intercept: 0,97	0,88
Accuracy Robustness: 0,99	Slope Robustness: 0,99	Intercept Robustness: 0,94	

Error decomposition and Y-scrambling



SCORES

Error Decomposition

Noise: 0,95

Deviation Slope: 0,02

Bias: 0,03

0,97

Y-Scrambling

Accuracy: **0,74**

Noise: **0,83**

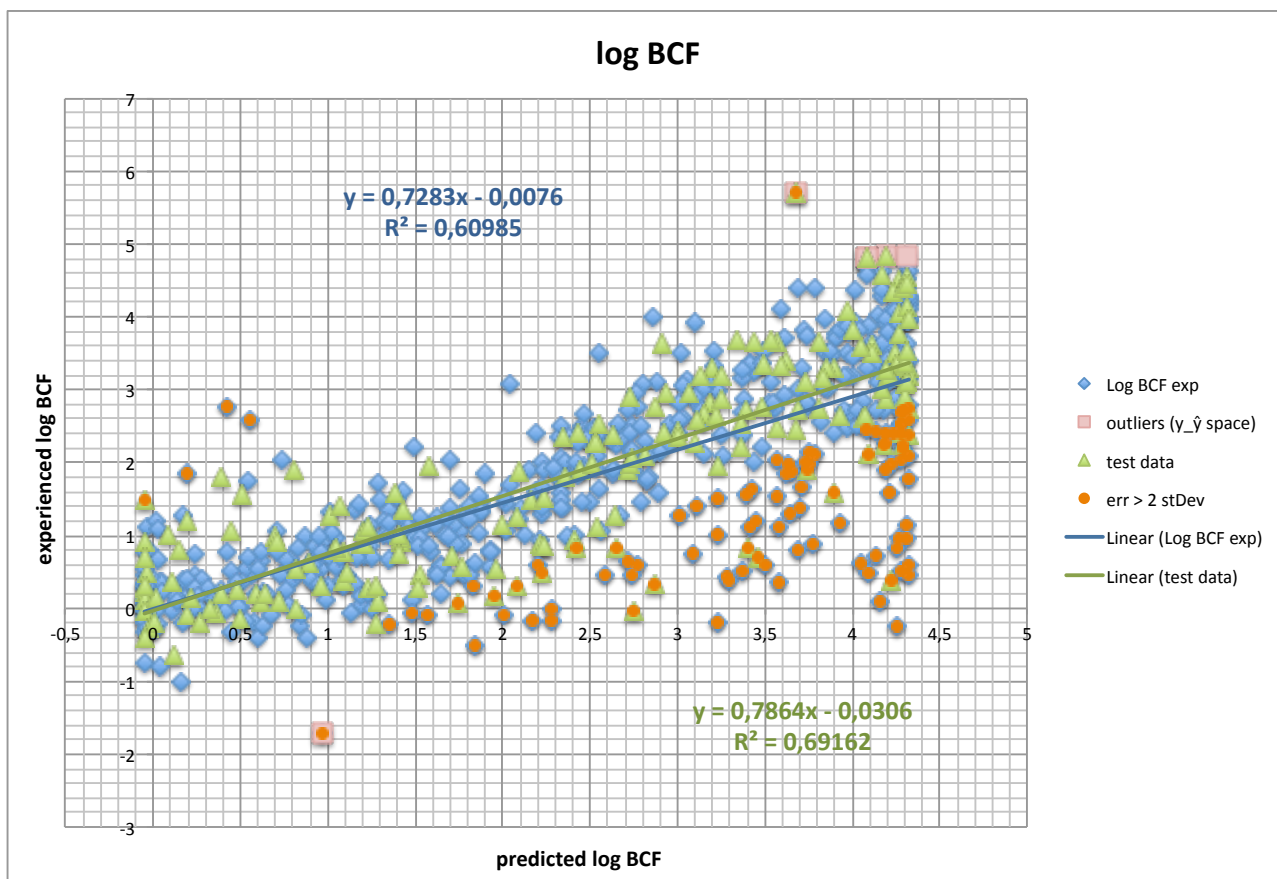
Deviation Slope: **0,89**

Bias: **0,97**

0,83

12. Episuite Arnod-Gobas Up KM=0 Model

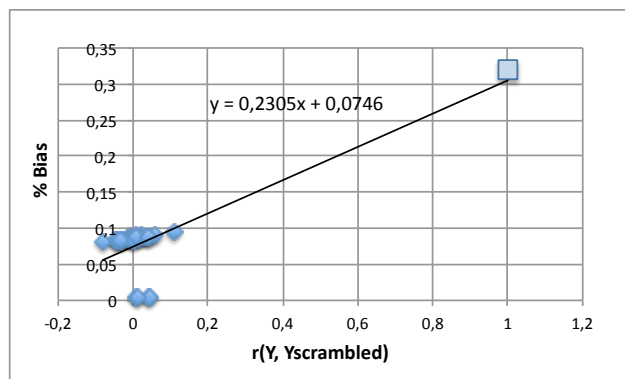
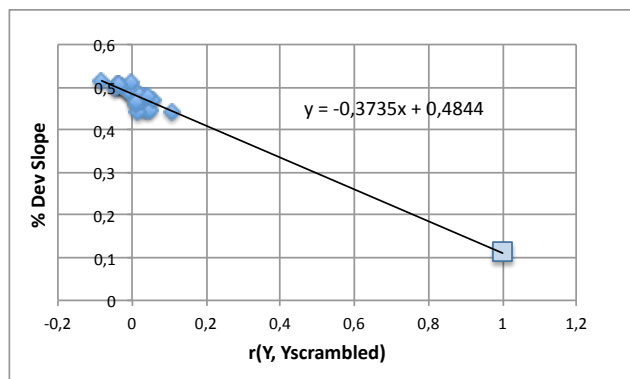
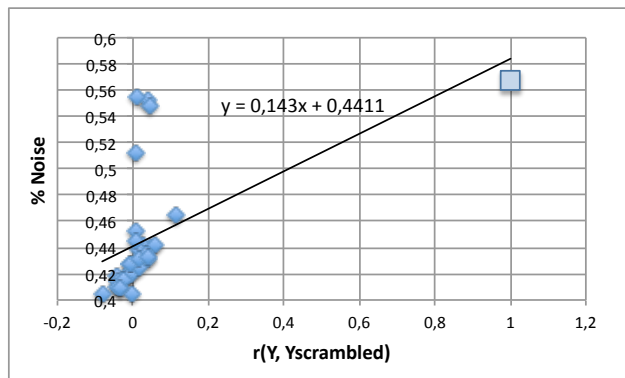
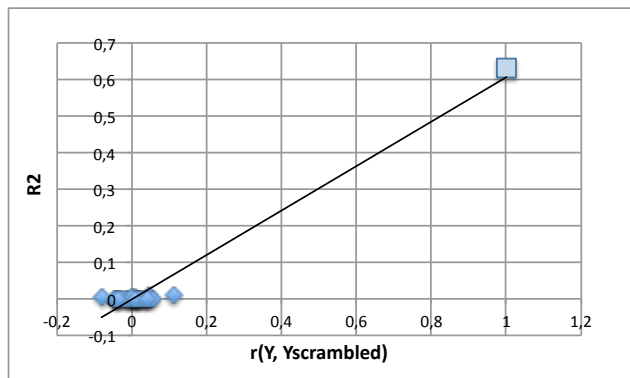
Basic model properties



SCORES

Sample Size: 697	0,99	FN: 14 % FP: 67 %	0,68
Accuracy: 0,63	Slope: 0,76	Intercept: 0,98	0,79
Accuracy Robustness: 0,88	Slope Robustness: 0,92	Intercept Robustness: 0,98	

Error decomposition and Y-scrambling



SCORES

Error Decomposition

Noise: 0,57

Deviation Slope: 0,11

Bias: 0,32

0,71

Y-Scrambling

Accuracy: 0,63

Noise: 0,25

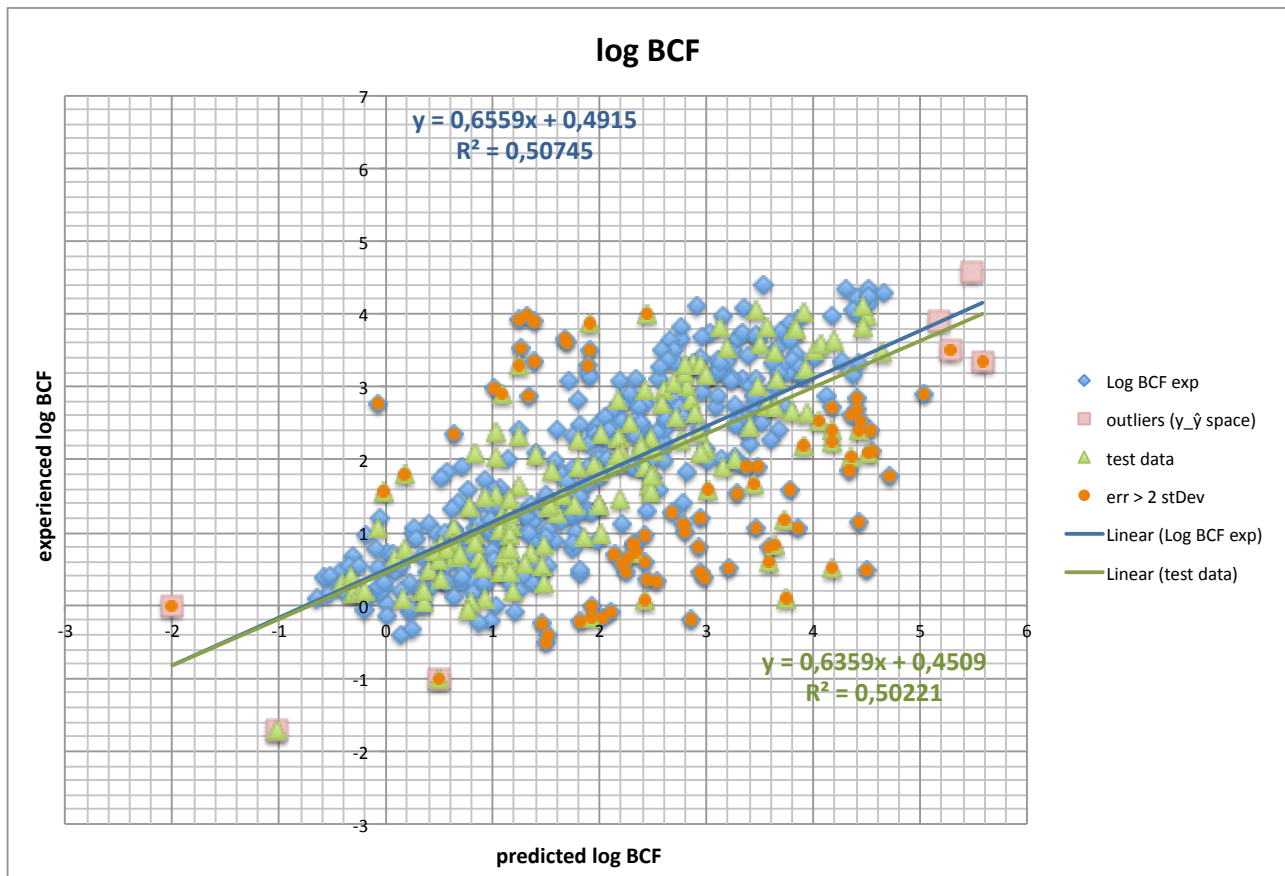
Deviation Slope: 0,74

Bias: 0,76

0,56

13. Mackay valid Model

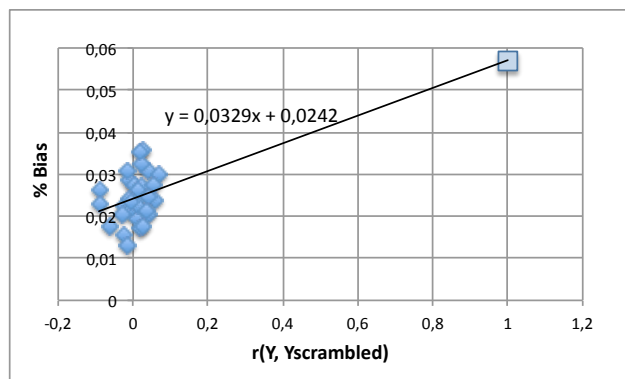
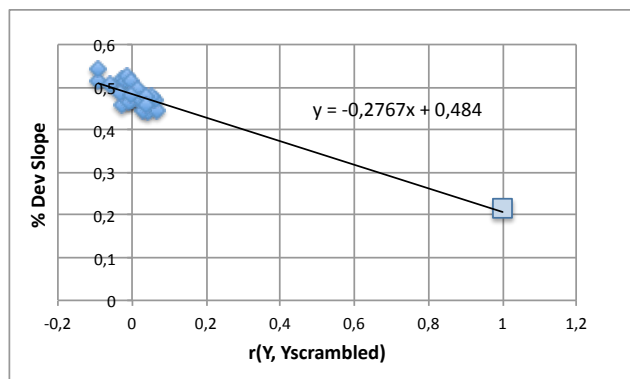
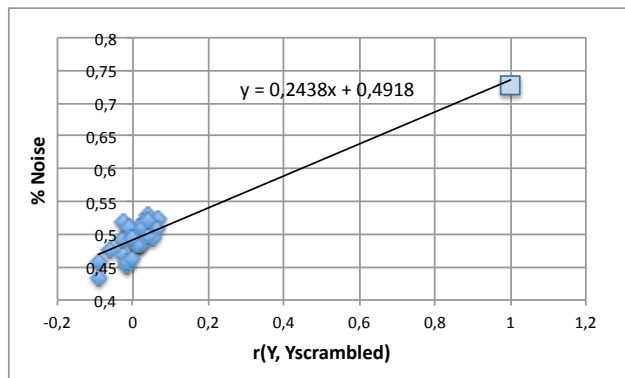
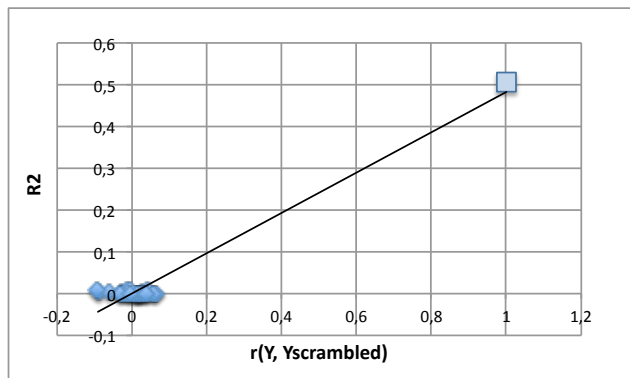
Basic model properties



SCORES

Sample Size: 549	0,78	FN: 32 % FP: 47 %	0,63
Accuracy: 0,51	Slope: 0,65	Intercept: 0,53	0,74
Accuracy Robustness: 1,00	Slope Robustness: 0,97	Intercept Robustness: 0,96	

Error decomposition and Y-scrambling



SCORES

Error Decomposition

Noise: 0,73

Deviation Slope: 0,22

Bias: 0,05

0,82

Y-Scrambling

Accuracy: 0,50

Noise: 0,47

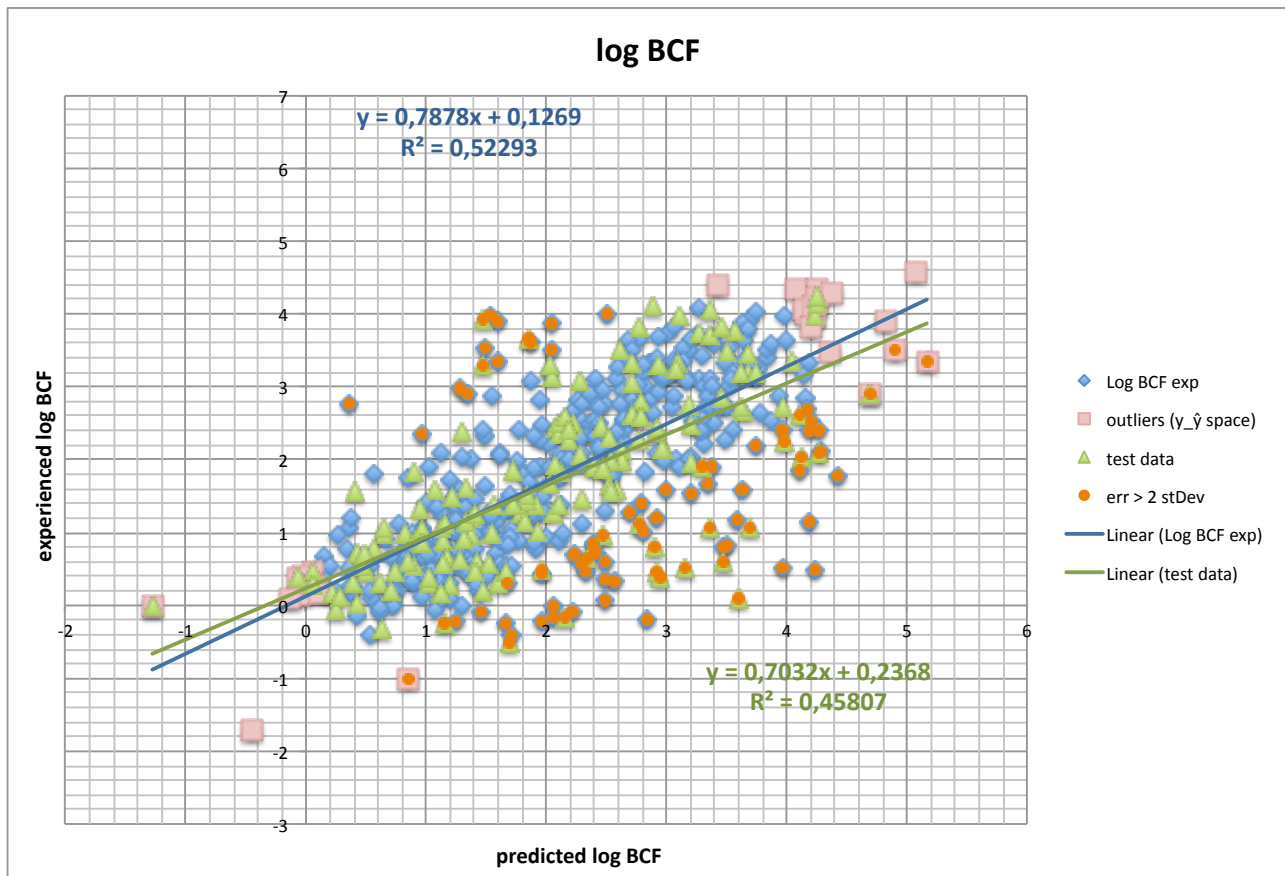
Deviation Slope: 0,53

Bias: 0,97

0,54

14. Veith valid Model

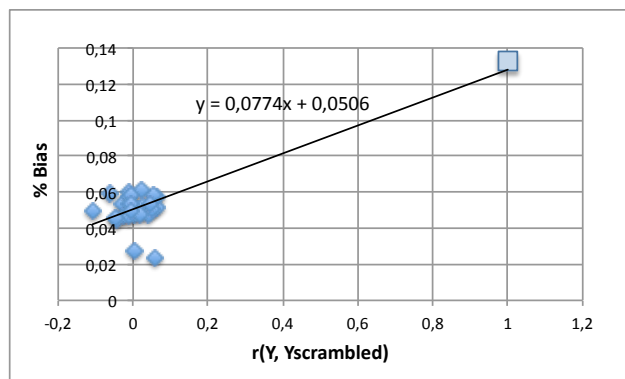
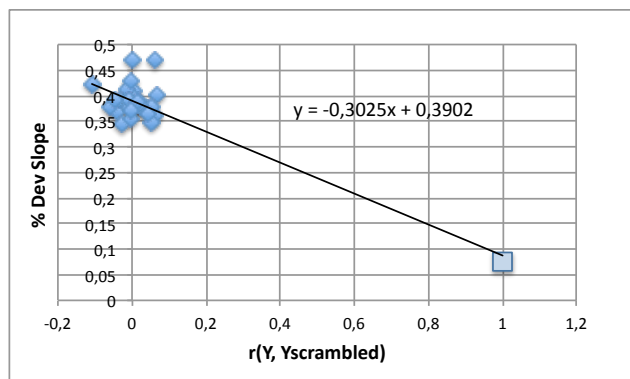
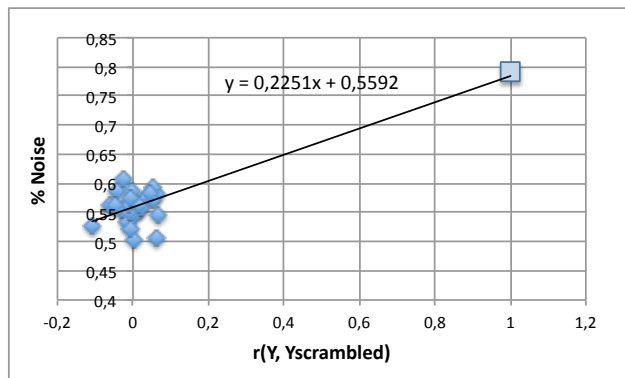
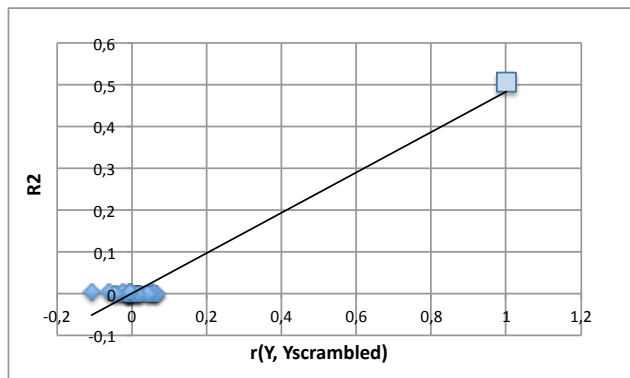
Basic model properties



SCORES

Sample Size: 549	0,78	FN: 23 % FP: 56 %	0,66
Accuracy: 0,51	Slope: 0,75	Intercept: 0,82	0,72
Accuracy Robustness: 0,88	Slope Robustness: 0,89	Intercept Robustness: 0,89	

Error decomposition and Y-scrambling



SCORES

Error Decomposition

Noise: 0,79

Deviation Slope: 0,08

Bias: 0,13

0,86

Y-Scrambling

Accuracy: 0,50

Noise: 0,46

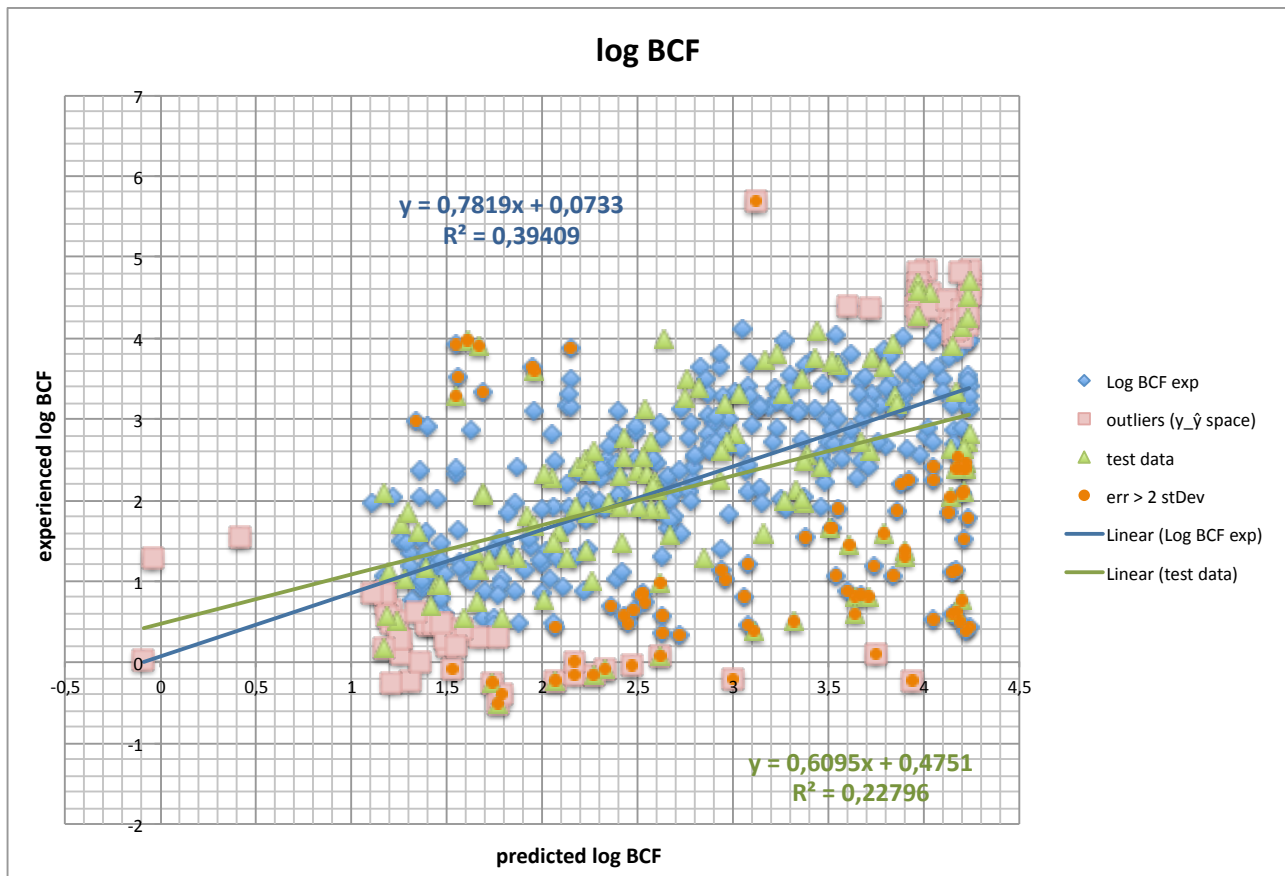
Deviation Slope: 0,63

Bias: 0,92

0,56

15. Bintein valid Model

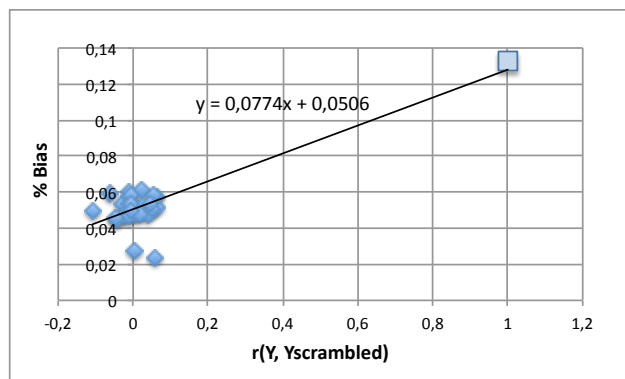
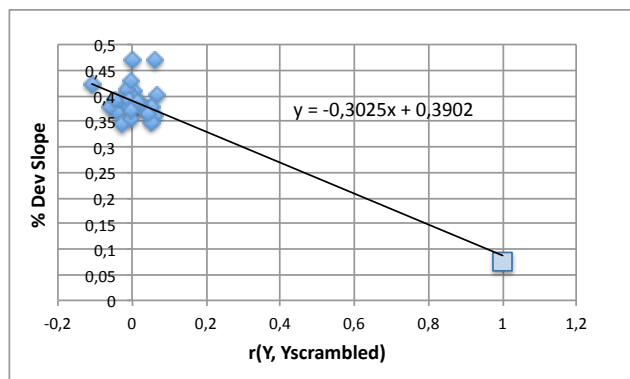
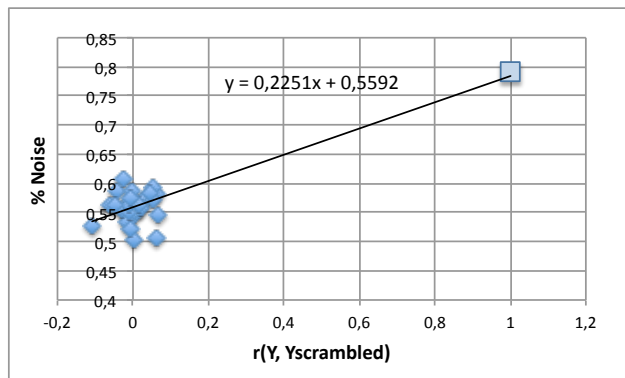
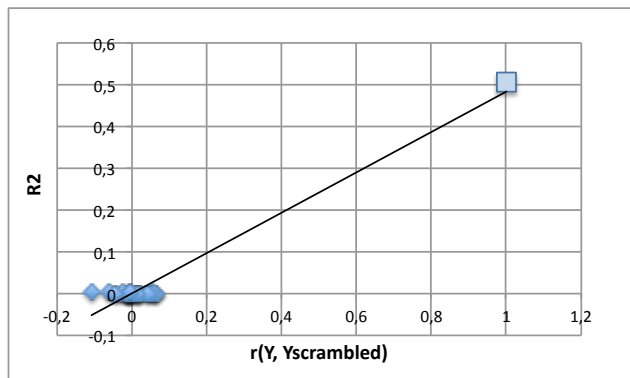
Basic model properties



SCORES

Sample Size: 514	0,73	FN: 16 % FP: 60 %	0,69
Accuracy: 0,35	Slope: 0,70	Intercept: 0,86	0,56
Accuracy Robustness: 0,56	Slope Robustness: 0,78	Intercept Robustness: 0,85	

Error decomposition and Y-scrambling



SCORES

Error Decomposition

Noise: 0,79

Deviation Slope: 0,08

Bias: 0,13

0,86

Y-Scrambling

Accuracy: 0,50

Noise: 0,46

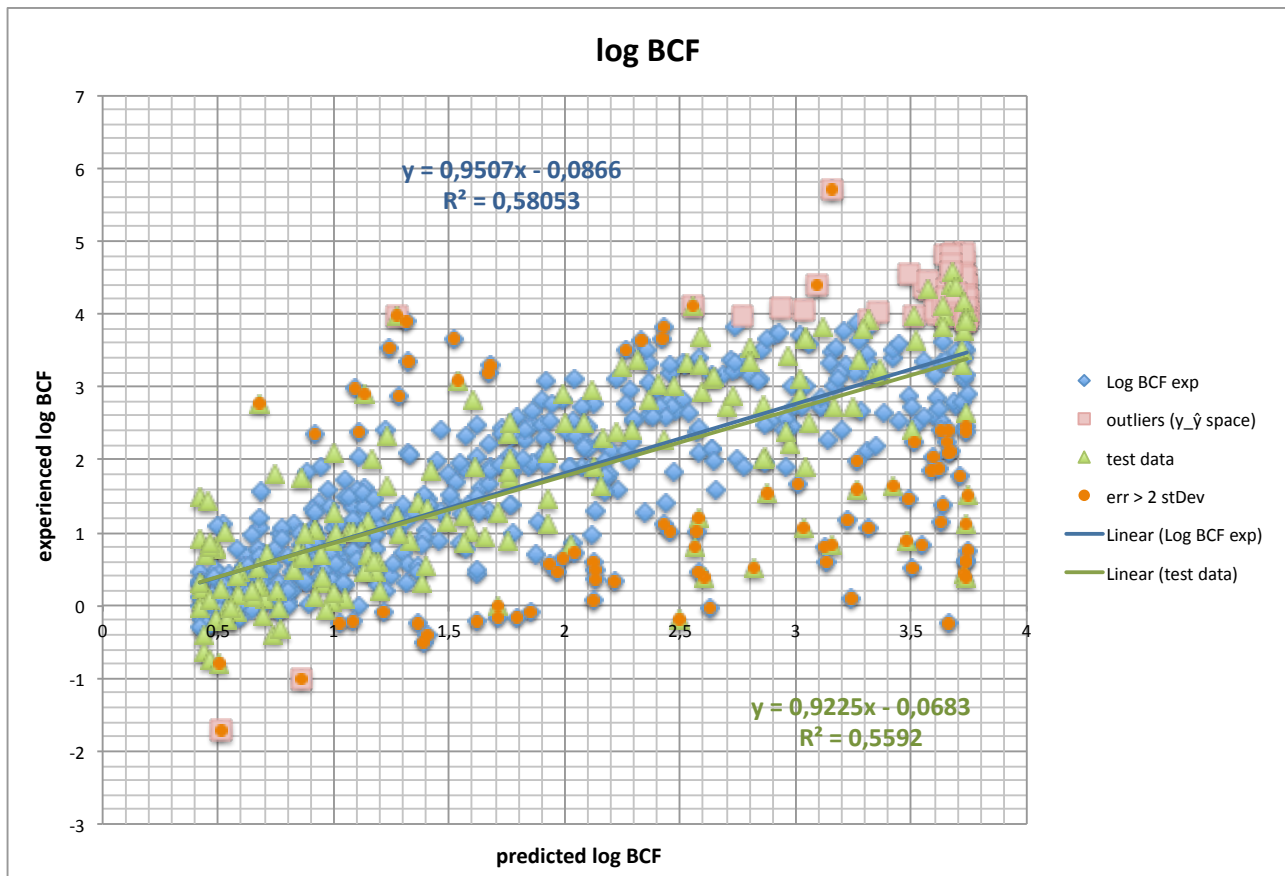
Deviation Slope: 0,63

Bias: 0,92

0,56

16. Dimitrov valid Model

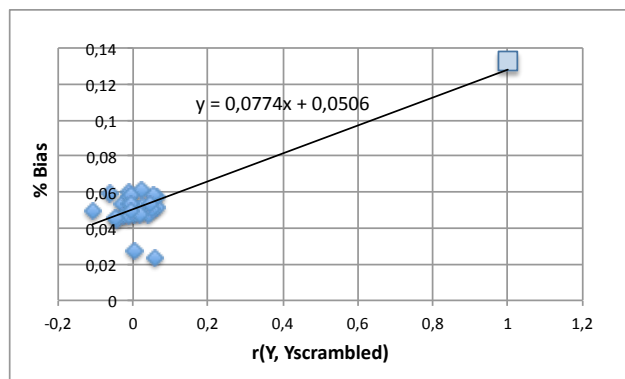
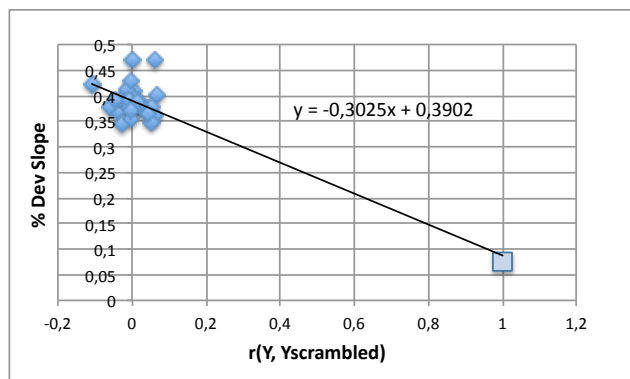
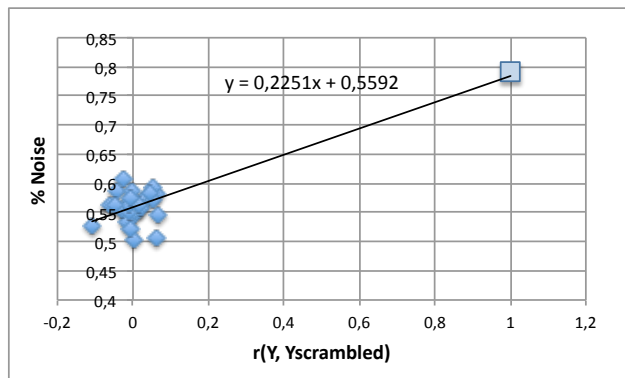
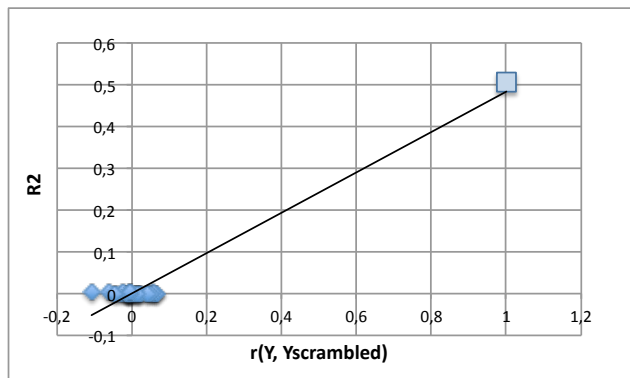
Basic model properties



SCORES

Sample Size: 687	0,98	FN: 31 % FP: 50 %	0,62
Accuracy: 0,58	Slope: 0,94	Intercept: 0,92	0,81
Accuracy Robustness: 0,97	Slope Robustness: 0,97	Intercept Robustness: 0,98	

Error decomposition and Y-scrambling



SCORES

Error Decomposition

Noise: 0,95

Deviation Slope: 0,0

Bias: 0,05

0,97

Y-Scrambling

Accuracy: **0,57**

Noise: **0,69**

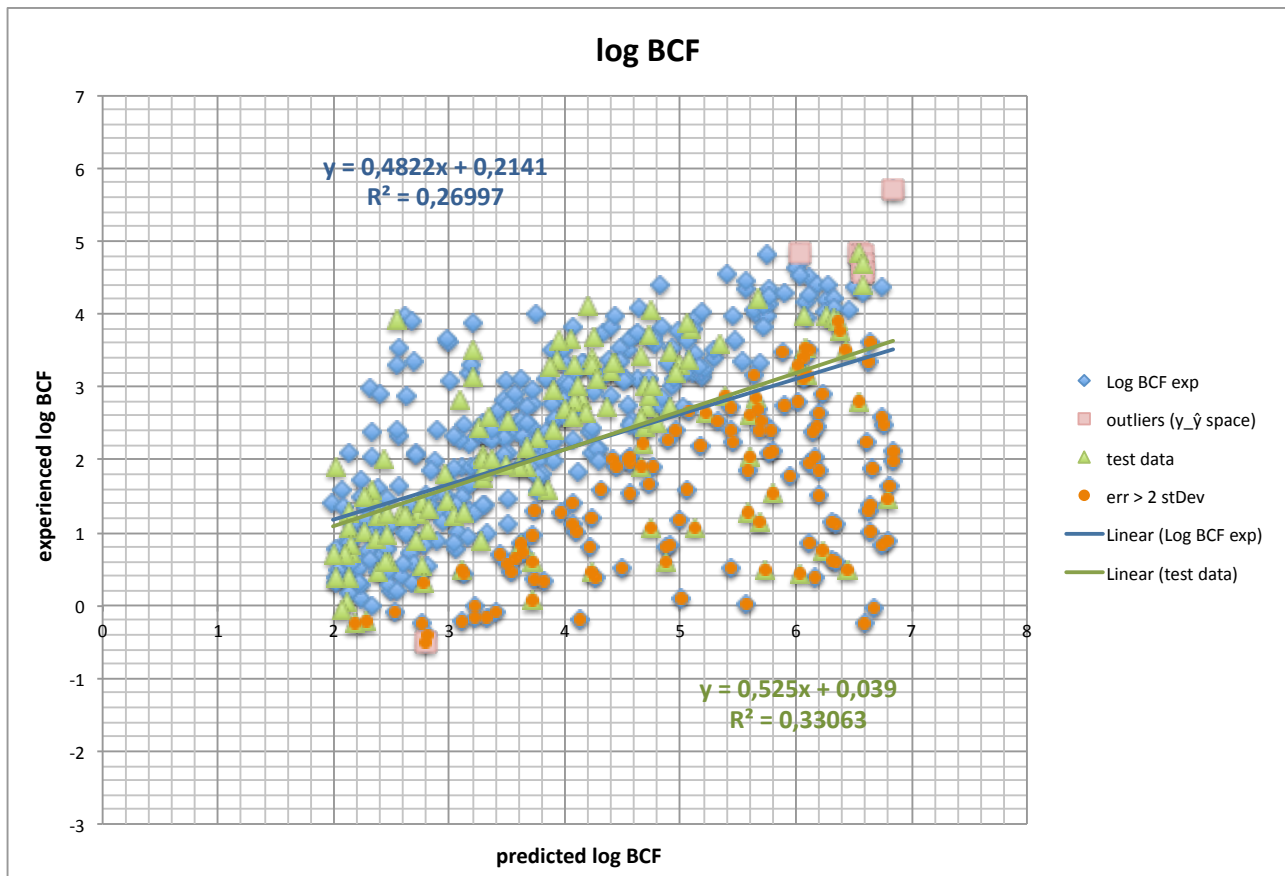
Deviation Slope: **0,74**

Bias: **0,97**

0,69

17. Nendza max valid Model

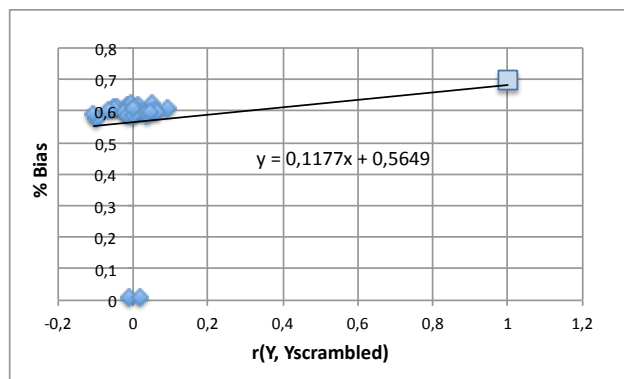
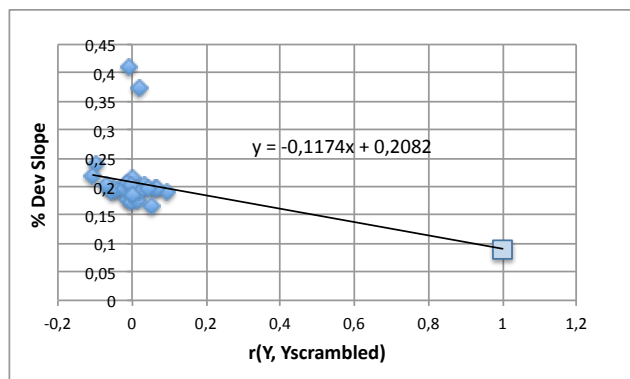
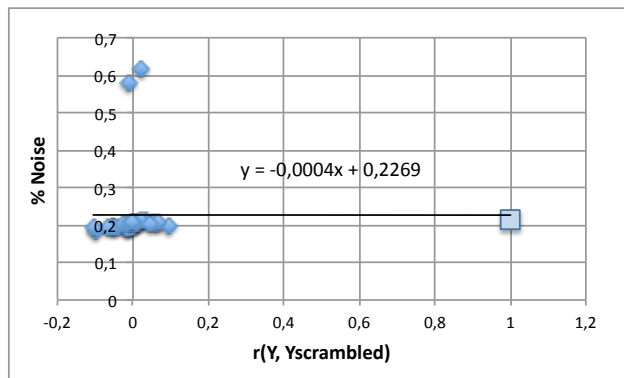
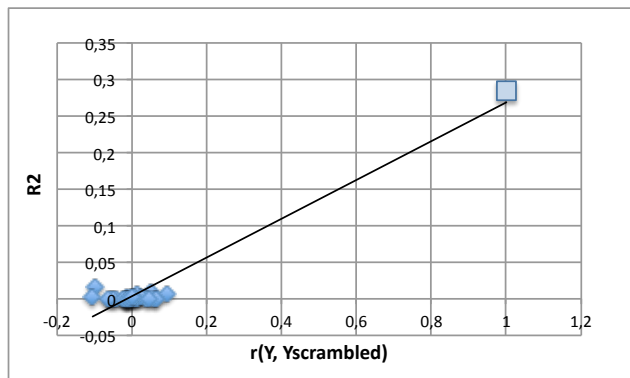
Basic model properties



SCORES

Sample Size: 531	0,76	FN: 2 % FP: 95 %	0,67
Accuracy: 0,28	Slope: 0,50	Intercept: 0,88	0,60
Accuracy Robustness: 0,82	Slope Robustness: 0,92	Intercept Robustness: 0,82	

Error decomposition and Y-scrambling



SCORES

Error Decomposition

Noise: 0,21

Deviation Slope: 0,09

Bias: 0,70

0,48

Y-Scrambling

Accuracy: **0,28**

Noise: **0,03**

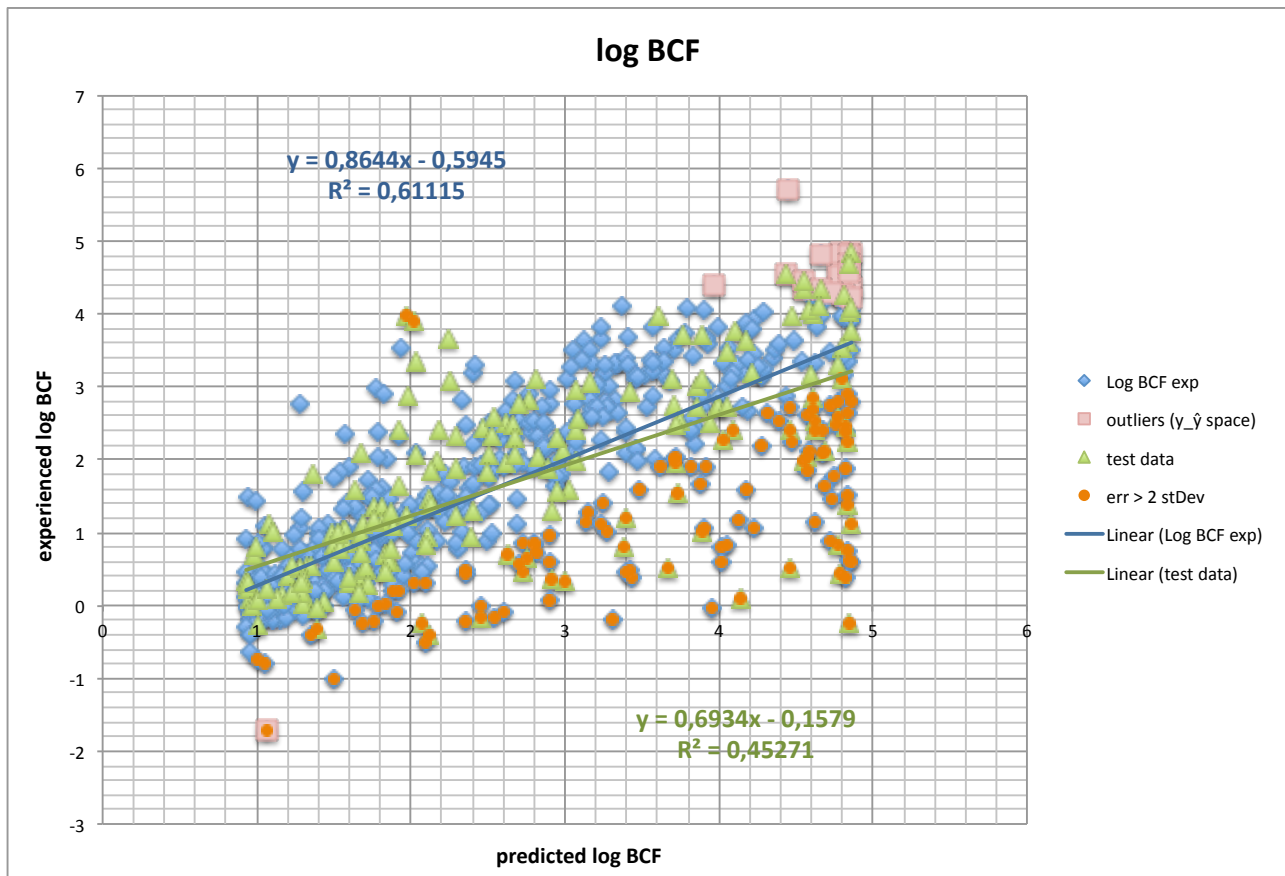
Deviation Slope: **0,24**

Bias: **0,87**

0,24

18. Dimitrov max valid Model

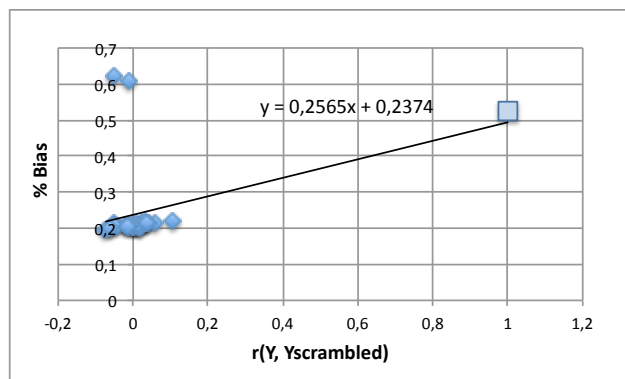
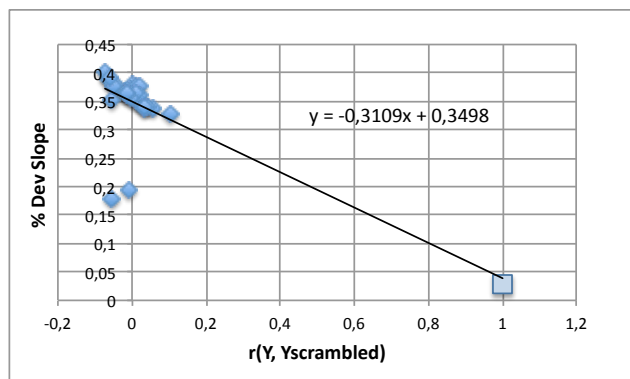
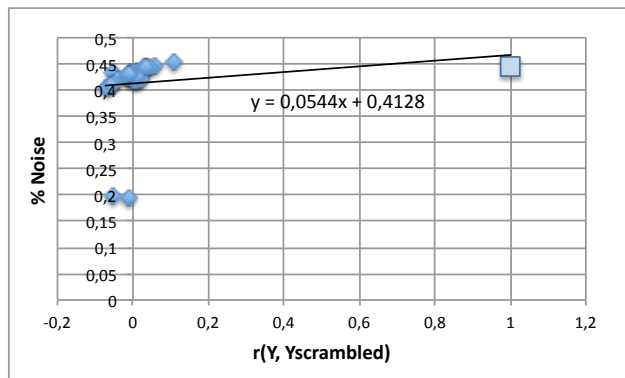
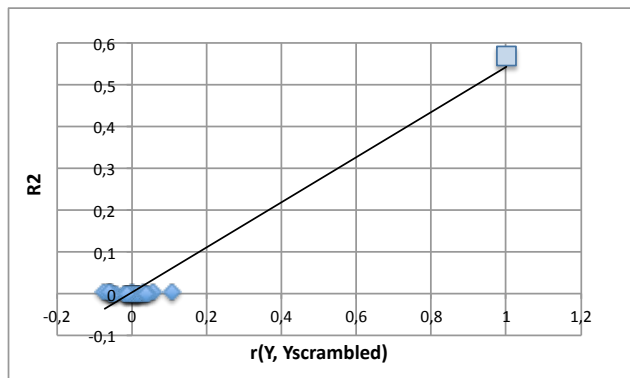
Basic model properties



SCORES

Sample Size: 687	0,98	FN: 5 % FP: 81 %	0,70
Accuracy: 0,57	Slope: 0,78	Intercept: 0,63	0,67
Accuracy Robustness: 0,74	Slope Robustness: 0,80	Intercept Robustness: 0,57	

Error decomposition and Y-scrambling



SCORES

Error Decomposition

Noise: 0,45

Deviation Slope: 0,03

Bias: 0,52

0,63

Y-Scrambling

Accuracy: **0,57**

Noise: **0,06**

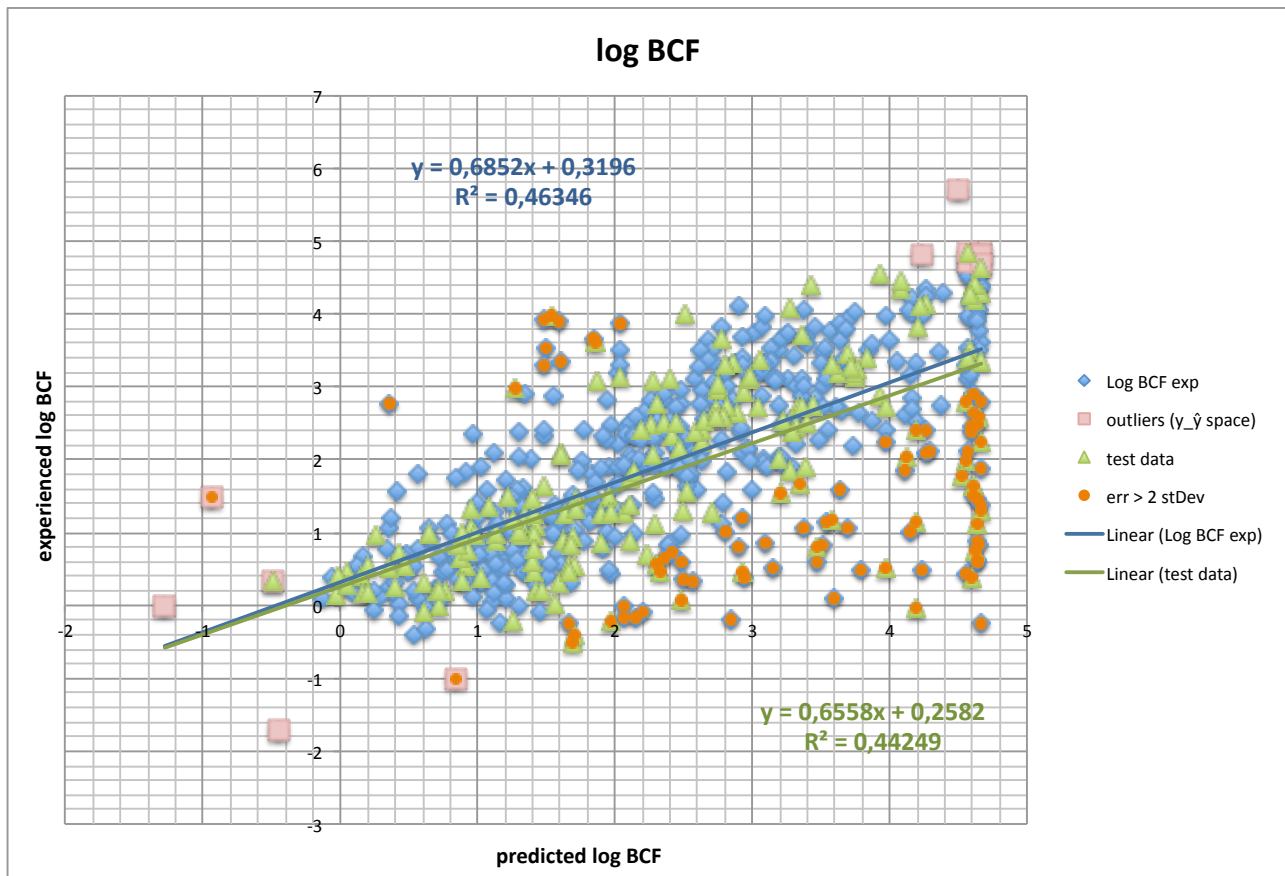
Deviation Slope: **0,64**

Bias: **0,71**

0,45

19. EUSES valid Model

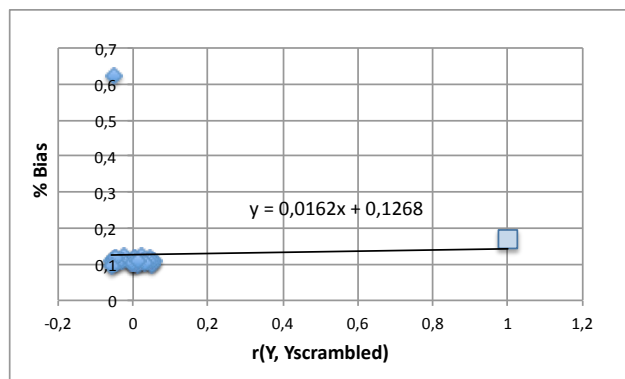
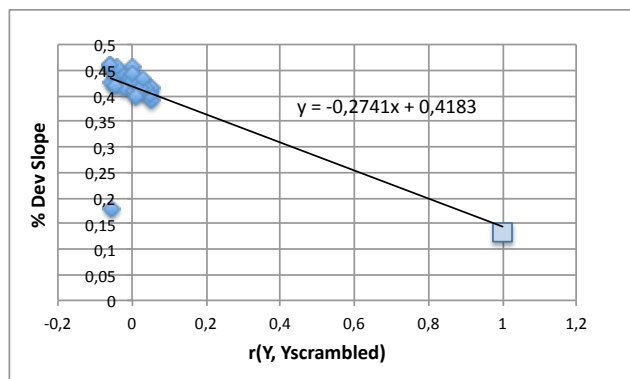
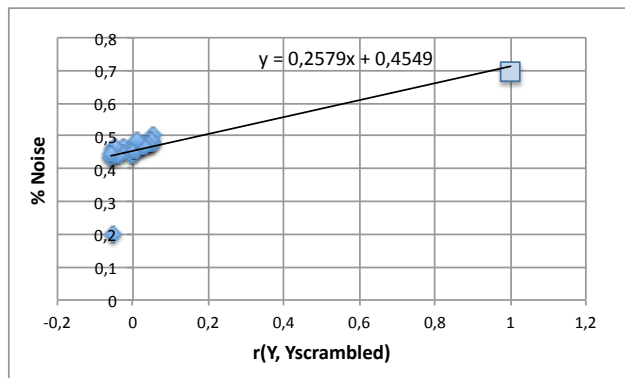
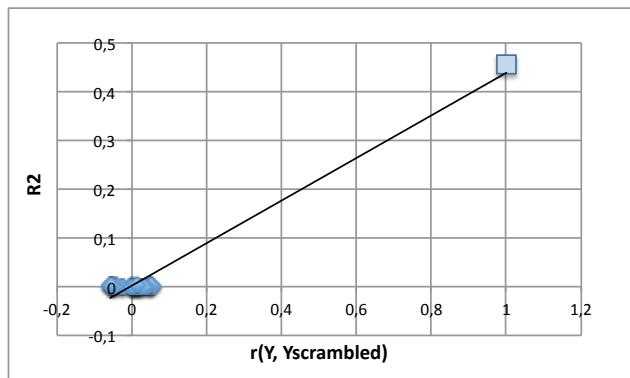
Basic model properties



SCORES

Sample Size: 629	0,90	FN: 21 % FP: 58 %	0,67
Accuracy: 0,46	Slope: 0,67	Intercept: 0,71	0,72
Accuracy Robustness: 0,96	Slope Robustness: 0,97	Intercept Robustness: 0,94	

Error decomposition and Y-scrambling



SCORES

Error Decomposition

Noise: 0,70

Deviation Slope: 0,13

Bias: 0,17

0,80

Y-Scrambling

Accuracy: 0,46

Noise: 0,48

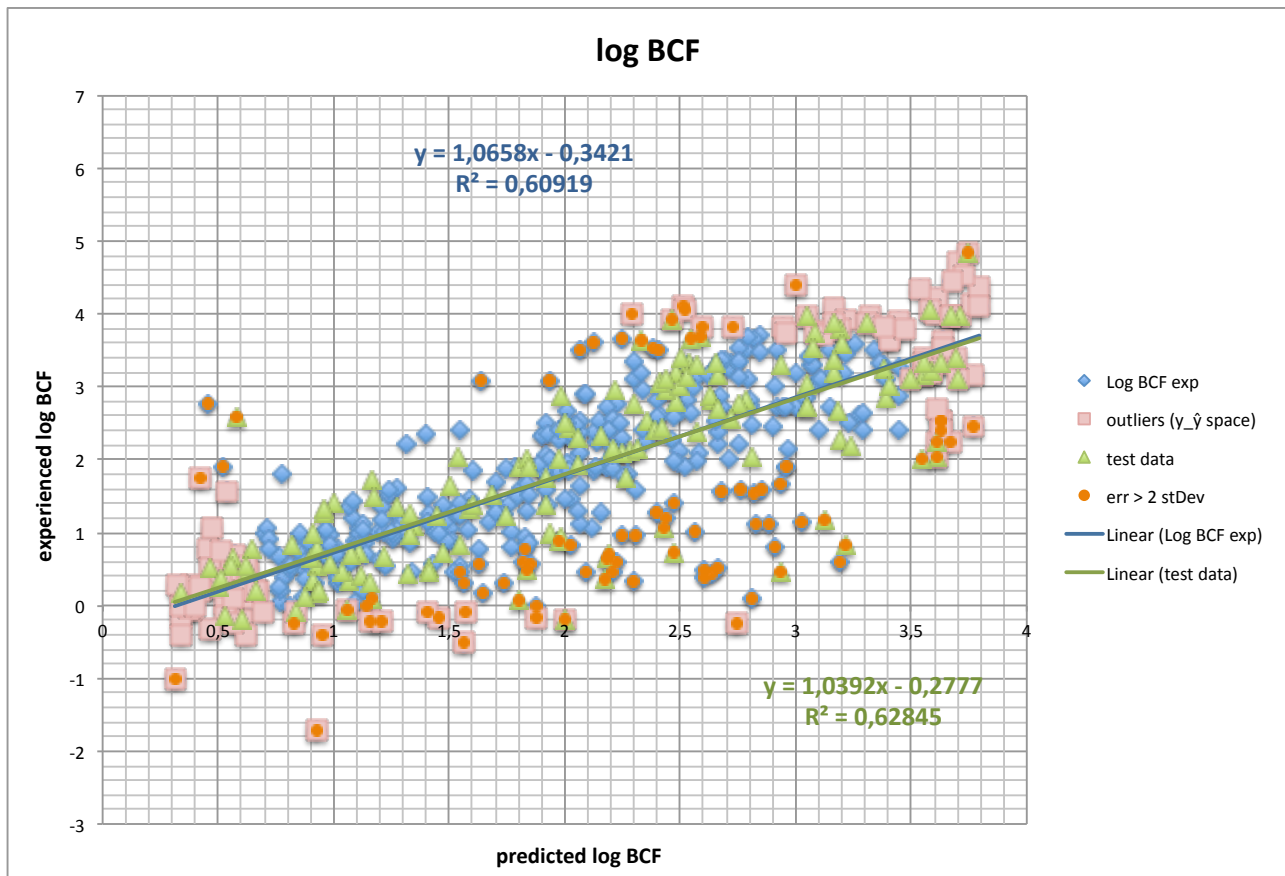
Deviation Slope: 0,57

Bias: 0,96

0,54

20. Veith & Cosian Model

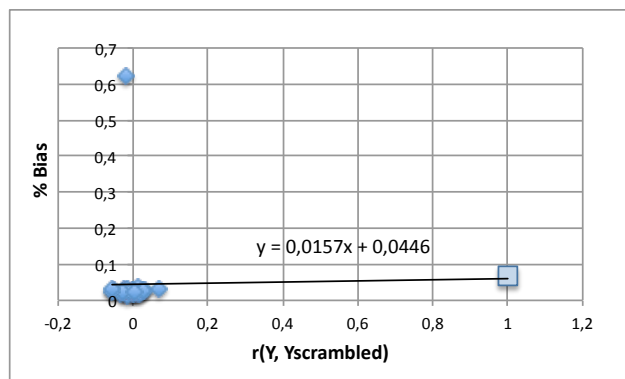
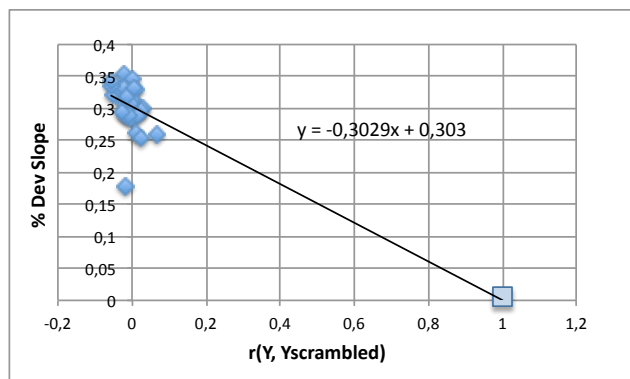
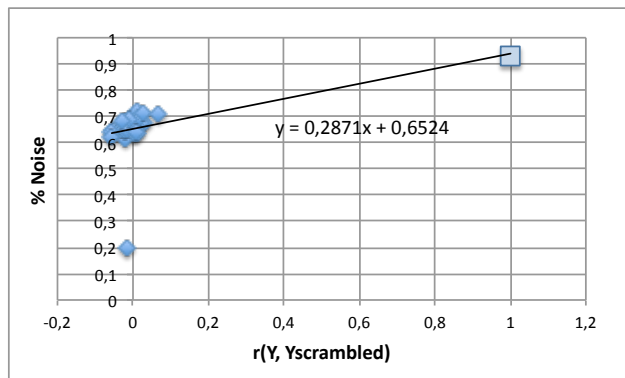
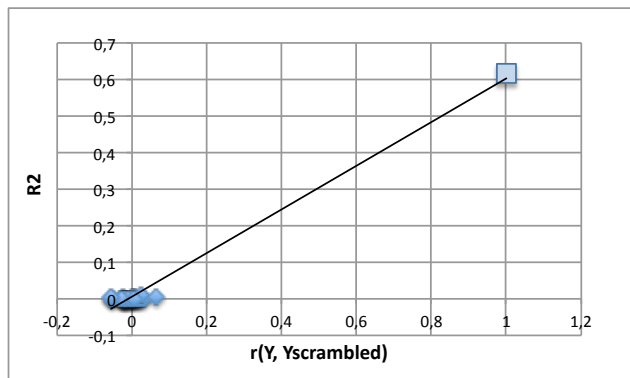
Basic model properties



SCORES

Sample Size: 545	0,78	FN: 27 % FP: 50 %	0,65
Accuracy: 0,61	Slope: 0,95	Intercept: 0,69	0,81
Accuracy Robustness: 0,97	Slope Robustness: 0,97	Intercept Robustness: 0,94	

Error decomposition and Y-scrambling



SCORES

Error Decomposition

Noise: 0,93

Deviation Slope: 0,0

Bias: 0,07

0,95

Y-Scrambling

Accuracy: 0,61

Noise: 0,56

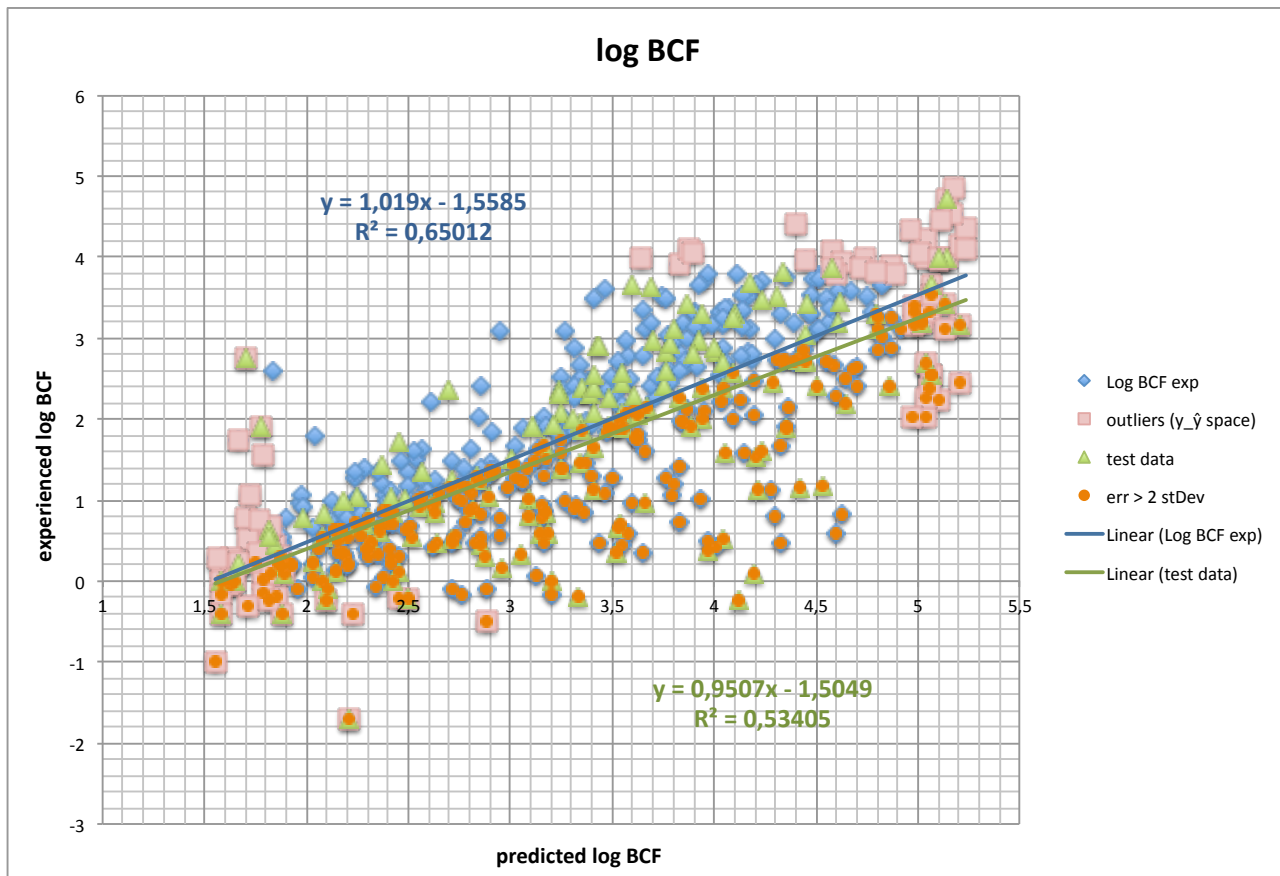
Deviation Slope: 0,60

Bias: 0,98

0,62

21. Esculder Gilabert Model

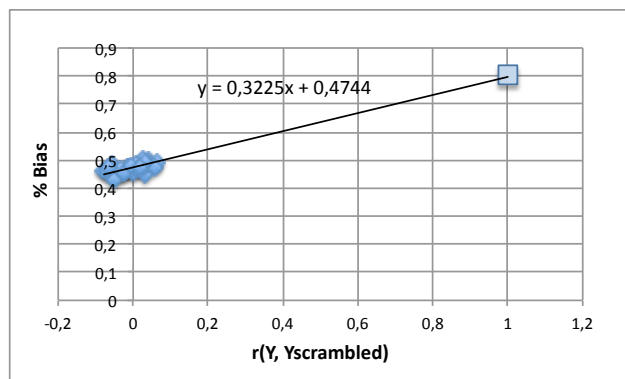
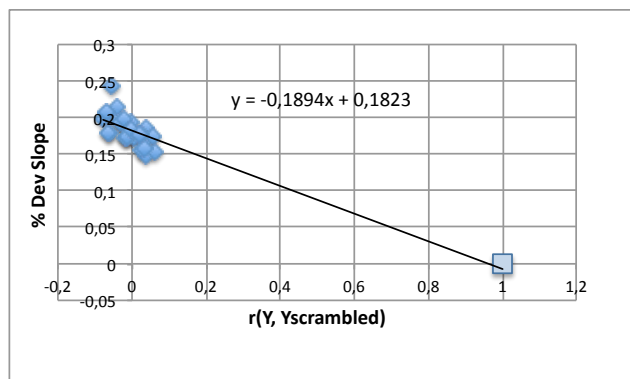
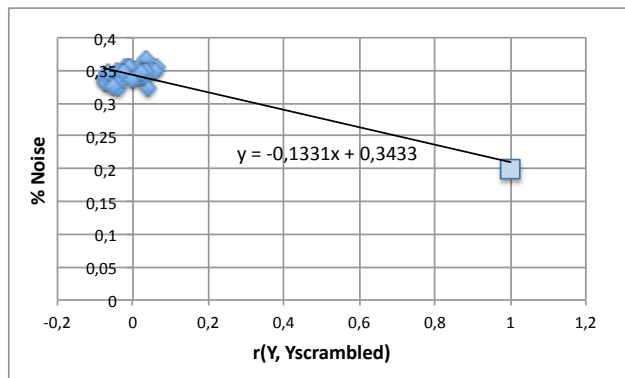
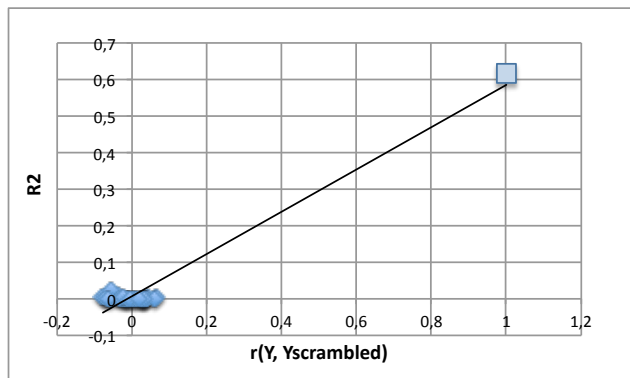
Basic model properties



SCORES

Sample Size: 545	0,78	FN: 0,5 % FP: 96 %	0,68
Accuracy: 0,61	Slope: 0,99	Intercept: 0,0	0,73
Accuracy Robustness: 0,82	Slope Robustness: 0,93	Intercept Robustness: 0,95	

Error decomposition and Y-scrambling



SCORES

Error Decomposition

Noise: 0,20

Deviation Slope: 0,0

Bias: 0,80

0,47

Y-Scrambling

Accuracy: **0,61**

Noise: **0,29**

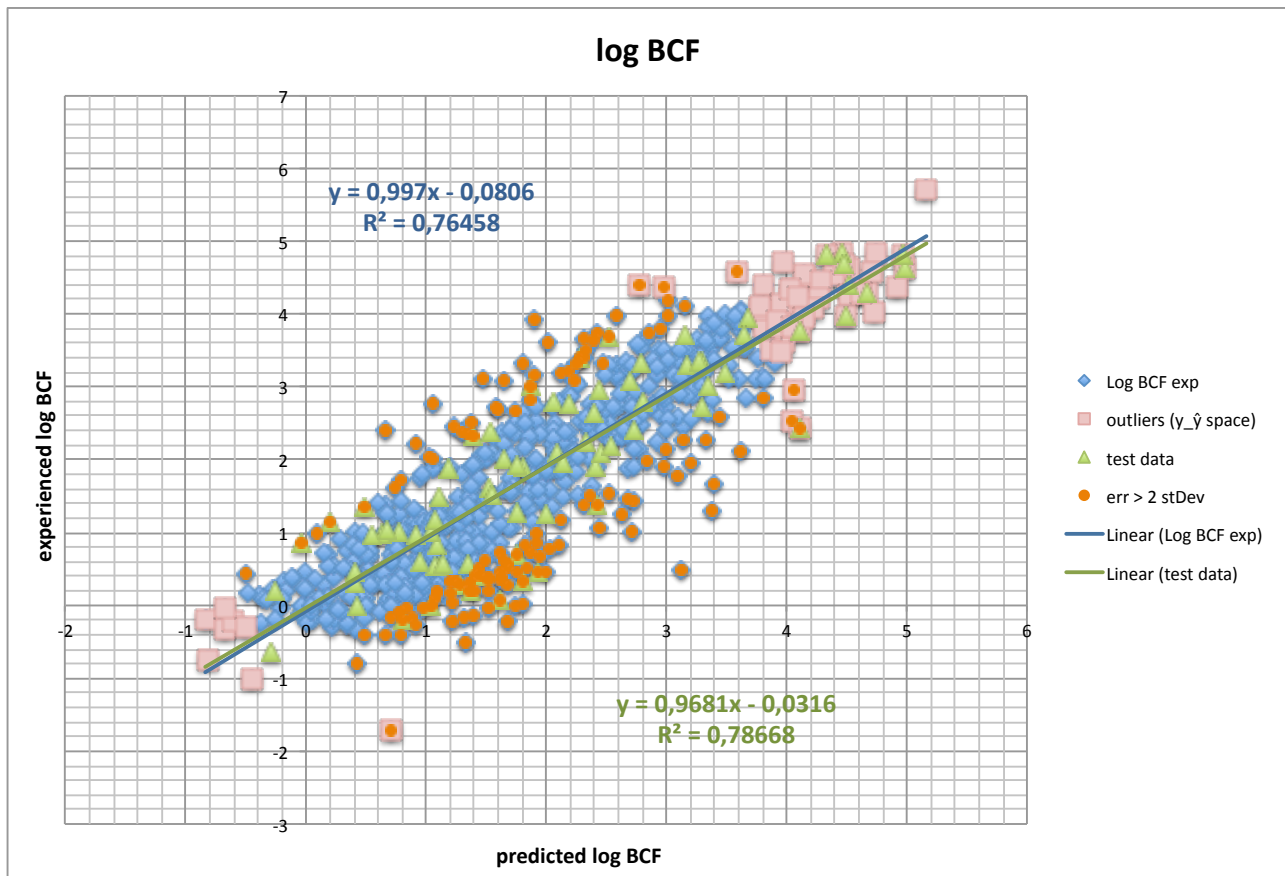
Deviation Slope: **0,37**

Bias: **0,67**

0,44

22. CORAL Model

Basic model properties

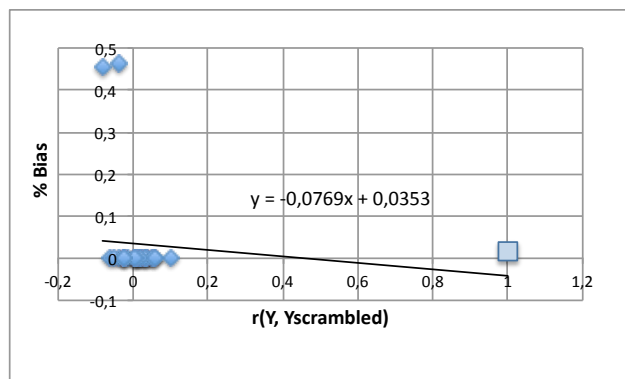
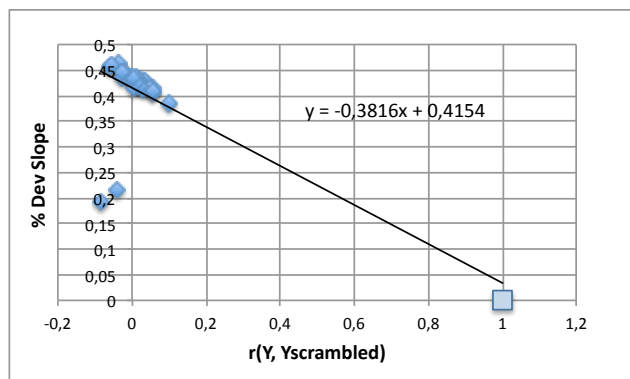
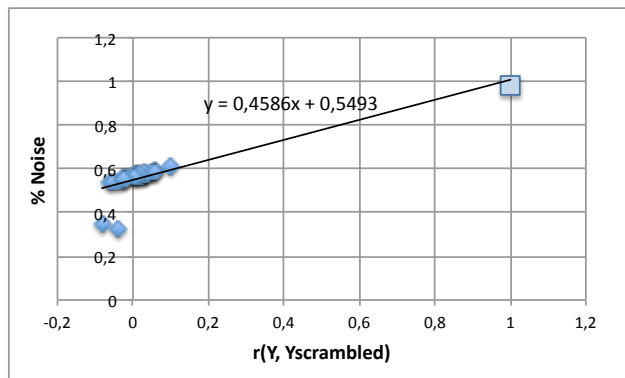
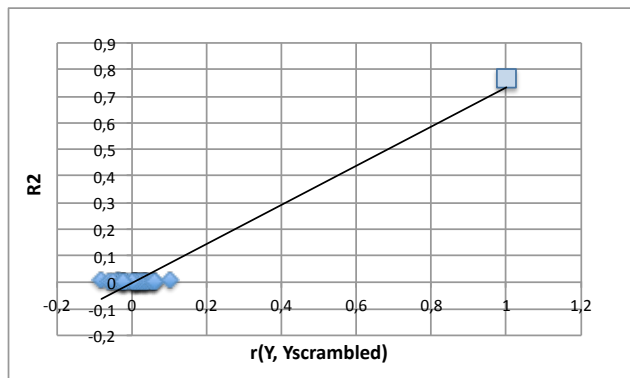


SCORES

Sample Size: 698	1,00	FN: 32 % FP: 42 %	0,65
Accuracy: 0,77	Slope: 0,98	Intercept: 0,95	0,89
Accuracy Robustness: 0,97	Slope Robustness: 0,97	Intercept Robustness: 0,95	

Alternative Non-Testing methods Assessed for REACH Substances

Error decomposition and Y-scrambling



SCORES

Error Decomposition

Noise: 0,98

Deviation Slope: 0,0

Bias: 0,02

0,99

Y-Scrambling

Accuracy: 0,77

Noise: 0,86

Deviation Slope: 0,83

Bias: 0,98

0,83

Model Performance Summary

Table 1 and 2 summarize the resulting scores for the tested models in descending order for not optimized and optimized false negative scores, correspondingly. The final score S is obtained as a weighted mean:

$$S = (0,25S_N + S_{FN} + S_S + S_E + S_P) / 4,25.$$

Table 1: Performance results for tested models in descending order.

Model #	Model Name	S_N	S_{FN}	S_S	S_E	S_P	Model offset	Score S
10	ES arnod-gobas mid	0,99	0,70	0,84	0,97	0,89	0,00	0,858
9	ES arnod-gobas up	0,99	0,70	0,82	0,98	0,89	0,00	0,857
11	ES arnod-gobas low	0,99	0,70	0,83	0,97	0,88	0,00	0,855
22	CORAL	1,00	0,65	0,83	0,99	0,89	0,00	0,849
8	ES regression	1,00	0,65	0,81	0,97	0,87	0,00	0,837
3	Caesar model valid	0,76	0,63	0,88	1,00	0,86	0,00	0,836
2	Caesar model	1,00	0,60	0,84	0,98	0,79	0,00	0,813
4	TEST	0,98	0,63	0,80	0,99	0,75	0,00	0,805
16	Dimitrov valid	0,98	0,62	0,69	0,97	0,81	0,00	0,787
20	Veith & Cosian	0,78	0,65	0,62	0,95	0,81	0,00	0,760
14	Veith valid	0,78	0,66	0,56	0,86	0,72	0,00	0,705
12	ES arnod-gobas up km0	0,99	0,68	0,56	0,71	0,79	0,00	0,703
19	EUSES valid	0,90	0,67	0,54	0,80	0,72	0,00	0,693
13	Mackey valid	0,78	0,63	0,54	0,82	0,74	0,00	0,686

Alternative Non-Testing methods Assessed for REACH Substances

6	Fu base	0,07	0,60	0,50	0,97	0,78	0,00	0,676
7	Fu	0,19	0,65	0,48	0,93	0,73	0,00	0,667
5	Fu acid	0,12	0,67	0,44	0,91	0,72	0,00	0,652
15	Bintein valid	0,73	0,69	0,47	0,82	0,56	0,00	0,641
18	Dimitov max valid	0,98	0,70	0,45	0,63	0,67	0,00	0,633
21	Esculder Gilabert	0,78	0,68	0,44	0,47	0,73	0,00	0,592
17	Nendza max valid	0,76	0,67	0,24	0,48	0,60	0,00	0,510

Table 2: Performance results for tested models with optimized false negative score in descending order.

Model #	Model Name	S_N	S_{FN}	S_S	S_E	S_P	Model offset μ	Score S
10	ES arnod-gobas mid	0,99	0,73	0,84	0,97	0,89	0,20	0,865
9	ES arnod-gobas up	0,99	0,73	0,82	0,98	0,89	0,15	0,863
11	ES arnod-gobas low	0,99	0,73	0,83	0,97	0,88	0,20	0,862
22	CORAL	1,00	0,70	0,83	0,99	0,89	0,55	0,862
3	Caesar model valid	0,76	0,70	0,88	1,00	0,86	0,30	0,853
8	ES regression	1,00	0,71	0,81	0,97	0,87	0,55	0,851
2	Caesar model	1,00	0,68	0,84	0,98	0,79	0,65	0,833
4	TEST	0,98	0,71	0,80	0,99	0,75	0,55	0,823
16	Dimitrov valid	0,98	0,71	0,69	0,97	0,81	0,55	0,806
20	Veith & Cosian	0,78	0,71	0,62	0,95	0,81	0,45	0,775

Alternative Non-Testing methods Assessed for REACH Substances

14	Veith valid	0,78	0,70	0,56	0,86	0,72	0,45	0,714
12	ES arnod- gobas up km0	0,99	0,69	0,56	0,71	0,79	0,20	0,705
6	Fu base	0,07	0,72	0,50	0,97	0,78	0,25	0,704
19	EUSES valid	0,90	0,70	0,54	0,80	0,72	0,30	0,700
13	Mackey valid	0,78	0,68	0,54	0,82	0,74	0,45	0,697
7	Fu	0,19	0,69	0,48	0,93	0,73	0,25	0,678
5	Fu acid	0,12	0,68	0,44	0,91	0,72	0,65	0,654
15	Bintein valid	0,73	0,71	0,47	0,82	0,56	0,30	0,645
18	Dimitov max valid	0,98	0,70	0,45	0,63	0,67	-0,20	0,633
21	Esculder Gilabert	0,78	0,69	0,44	0,47	0,73	-0,50	0,595
17	Nendza max valid	0,76	0,68	0,24	0,48	0,60	-0,50	0,511

