

## The index of ideality of correlation: a way to improve predictive potential of QSAR models

### *Supplementary materials*

The general scheme of applications of the CORAL software  
(How to use the CORAL software?)

1. Download CORALSEA folder:

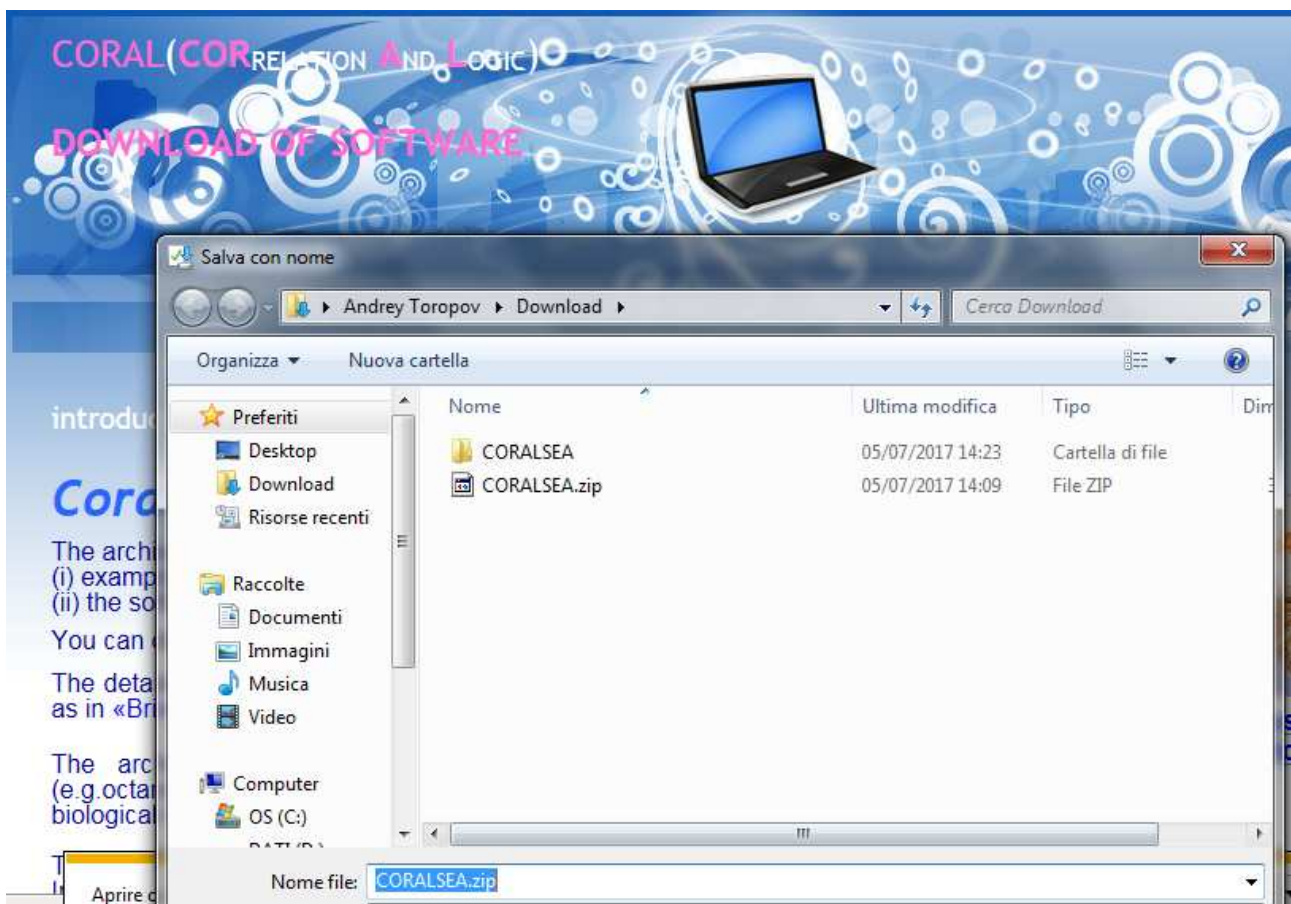
<http://www.insilico.eu/coral/>



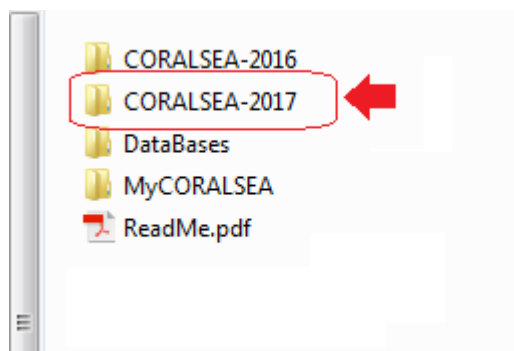
2. Click “DOWNLOAD”



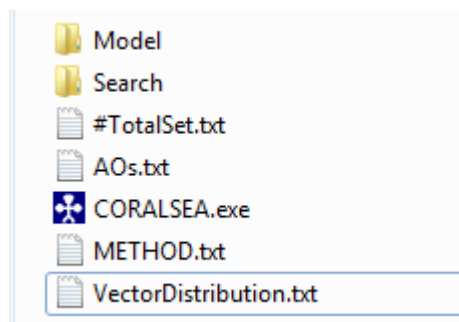
3. Save CORALSEA.zip



4. Unzip CORALSEA.zip



5. open CORALSEA-2017



6. Insert your data in file “#TotalSet.txt” (delete old version):

```

-315 BrC=1C(=O)C(NCCCC[N@@])(C2CCCC2OC)CC=C(Br)C(=O)C=1NCCCC[N@@](C1CCCC1OC)CC 8.575
+200 Br1cc(ccc1)C(=O)N\C(=C/c1cccc1)\C(OCCN(C)C)=O 3.939
#96 Br1cc2onc(c2cc1)CCC1CCN(CC1)C1cccc1 7.301
-201 Br1ccc(cc1)C(=O)N\C(=C/c1cccc1)\C(OCCN(C)C)=O 3.335
+183 Br1ccc(cc1)CN1C(=O)C2(N=C1CCCC)CCCC2 7.155
#199 Br1cccc1C(=O)N\C(=C/c1cccc1)\C(OCCN(C)C)=O 3.161
-314 ClC=1C(=O)C(NCCCC[N@@])(C2CCCC2OC)CC=C(Cl)C(=O)C=1NCCCC[N@@](C1CCCC1OC)CC 8.516
-16 Cl1c2CN3C(=Nc2ccc1)CCCC3 5.770
*13 Cl1c2N=C3N(CCCCC3)C2ccc1 5.820
#37 Cl1c2c(nc3[C@H]4CC[C@H](c3c2N)C4)ccc1 7.370
*222 Cl1c2c(nc3e(CCCC3)c2NCCCC(=O)NCCc2c3e([nH]c2)cccc3)ccc1 9.060
-34 Cl1c2nc3[C@H]4CC[C@H](c3c(N)c2ccc1)C4 7.070
#17 Cl1c2nc3e(CCCC3)c(N)c2ccc1 7.160
-269 Cl1cc(cc(C1)c1)\C=C\1/Oc2c(ccc(OCCCC[N@])(C3cc(OC(=O)NC)ccc3)C)c2/C1=O 7.351
*268 Cl1cc(ccc1Cl)\C=C\1/Oc2c(ccc(OCCCC[N@])(C3cc(OC(=O)NC)ccc3)C)c2/C1=O 8.097
#15 Cl1cc2CN3C(=Nc2ccc1)CCCC3 6.300
+163 Cl1cc2c(cc1)c(NCCCN1CCC(CC1)CC1Cc3cc(OC)c(OC)cc3C1)c1CCCCc1c2 8.975
-173 Cl1cc2c(cc1)c(NCCN)c1CCCCc1c2 7.668
+161 Cl1cc2c(cc1)c(NCCN1CCC(CC1)CC1Cc3cc(OC)c(OC)cc3C1)c1CCCCc1c2 8.585
+19 Cl1cc2c(nc3e(CCCC3)c2N)ccc1 6.260
#10 Cl1cc2c(nc3e([C@H]4CC(=C[C@H](C3)C4)CC)c2N)ccc1 6.372
-579 Cl1cc2nc3e(C4CC(=CC(C4)C3)C)c(NCC[N-][NH+]=N)c2ccc1 8.061
*580 Cl1cc2nc3e(C4CC(=CC(C4)C3)CC)c(NCC[N-][NH+]=N)c2ccc1 7.863
#221 Cl1cc2nc3e(CCCC3)c(NCCCCC(=O)NCCc3c4e([nH]c3)cccc4)c2ccc1 10.000
*345 Cl1cc2nc3e(CCCC3)c(NCCCCCNC(=O)CCc3cc4e(nc(c5CCCOe45)-c4cccc4)cc3)c2ccc1 7.854
-344 Cl1cc2nc3e(CCCC3)c(NCCCCCNC(=O)CCc3cc4e(nc(c5CCCOe45)-c4cccc4)cc3)c2ccc1 7.842
*343 Cl1cc2nc3e(CCCC3)c(NCCCCCNC(=O)CCc3cc4e(nc(c5CCCOe45)-c4cccc4)cc3)c2ccc1 7.955
-342 Cl1cc2nc3e(CCCC3)c(NCCCCCNC(=O)CCc3cc4e(nc(c5CCCOe45)-c4cccc4)cc3)c2ccc1 8.016
+341 Cl1cc2nc3e(CCCC3)c(NCCCCCNC(=O)CCc3cc4e(nc(c5CCCOe45)-c4cccc4)cc3)c2ccc1 7.780
+153 Cl1cc2nc3e(CCCC3)c(NCCNCCCCCOc3c4e5e([nH]c4ccc3)cccc5)c2ccc1 8.590
#154 Cl1cc2nc3e(CCCC3)c(NCCNCCCCCOc3c4e5e([nH]c4ccc3)cccc5)c2ccc1 8.812
+156 Cl1cc2nc3e(CCCC3)c(NCCNCCCCCOc3c4e5e([nH]c4ccc3)cccc5)c2ccc1 8.783
*155 Cl1cc2nc3e(CCCC3)c(NCCNCCCCCOc3c4e5e([nH]c4ccc3)cccc5)c2ccc1 8.668
...
+138 s1c2cc(OC)c(OC)cc2cc1C(=O)CCc1cc[n+](cc1)C1ccc(cc1)C(OC)=O 6.000
-118 s1c2cc(OC)c(OC)cc2cc1C(=O)C=C1CC[N+](CC1)(C1cccc1)C 6.284
#103 s1nc(c2c1cccc2)CCC1CCN(CC1)C1cccc1 7.004

```

7. Run CORALSEA.exe



CORAL: Loading of method or system

**Method:** Scheme: Additive or Multiplicative

Load method Method.txt

**Training set** → EXPR

**Invisible Training set** → EXPR

**Calibration set** → EXPR

The preparation of split into training and validation sets

Loading of details of built model

Model Details.txt

Place of compound (CAS) in graphical representations

Search for duplicates in SMILES Search for duplicates in CAS (ID)

**SMILES for Training and Calibration sets**

Graph ☐ HSG ☐ HFG ☐ GAO

ec0 ☐ ec1 ☐ pt2 ☐ ec2 ☐ pt3 ☐ vs2 ☐ ec3 ☐ NNC

C3 ☐ C4 ☐ C5 ☐ C6 ☐ C7

Classification model

Concordance Correlation Coefficient

Classic Scheme

Balance of correlations ☒ dR weight \*\*\*

Ideal slopes ☐ dC weight \*\*\*

D\_start \*\*\* d\_limit \*\*\* N\_epoch \*\*\*

Index of Ideality of correlation (IIC) ☐

Start threshold value: \*\*\*

Maximal threshold value: \*\*\*

Number of the Monte Carlo probes: \*\*\*

Depth of Interpretation \*\*\*

Outliers 5

DemoDCW

EvolutionCorr

Split Info

| W% | N111 | N110 | N101 | N100 | NaII | DETECT |
|----|------|------|------|------|------|--------|
| 0  | 0    | 0    | 0    | 0    | 0    | 0      |

EXIT

## 8. Click button

The preparation of split into training and validation sets

CORAL: Loading of method or system

**Method:** Scheme: Additive or Multiplicative

Load method Method.txt

**Training set** → EXPR

**Invisible Training set** → EXPR

**Calibration set** → EXPR

The preparation of split into training and validation sets

Loading of details of built model

Model Details.txt

Place of compound (CAS) in graphical representations

Search for duplicates in SMILES Search for duplicates in CAS (ID)

**SMILES for Training and Calibration sets**

Graph ☐ HSG ☐ HFG ☐ GAO

ec0 ☐ ec1 ☐ pt2 ☐ ec2 ☐ pt3 ☐ vs2 ☐ ec3 ☐ NNC

C3 ☐ C4 ☐ C5 ☐ C6 ☐ C7

Classification model

Concordance Correlation Coefficient

Classic Scheme

Balance of correlations ☒ dR weight \*\*\*

Ideal slopes ☐ dC weight \*\*\*

D\_start \*\*\* d\_limit \*\*\* N\_epoch \*\*\*

Index of Ideality of correlation (IIC) ☐

Start threshold value: \*\*\*

Maximal threshold value: \*\*\*

Number of the Monte Carlo probes: \*\*\*

Depth of Interpretation \*\*\*

Outliers 5

DemoDCW

EvolutionCorr

Split Info

| W% | N111 | N110 | N101 | N100 | NaII | DETECT |
|----|------|------|------|------|------|--------|
| 0  | 0    | 0    | 0    | 0    | 0    | 0      |

EXIT

## 9. Define percentage of available data in the training, invisible training, calibration, and validation sets

**Preparation of split**

The first action is "Load". File "#TotalSet.txt" must exist in your folder.

The second action should be "Do distribution".

The third action should be "Save files".

|    | Input           |      | Output         |
|----|-----------------|------|----------------|
| 1. | #TotalSet.txt   | Load | #TotalSet'.txt |
| 2. | Do distribution |      |                |
| 3. | Save files      |      |                |

Dispersion Limit = 0,01

|                        |   | Planned Distribution | Frequency | Actual Distribution |
|------------------------|---|----------------------|-----------|---------------------|
| #TrainingSet.txt       | + | 0,25                 | 0         | 0                   |
| Invisible training set | - | 0,25                 | 0         | 0                   |
| Calibration set        | # | 0,25                 | 0         | 0                   |
| #ValidationSet.txt     | * | 0,25                 | 0         | 0                   |

**CORAL: Loading of method or system**

**Method:** Scheme: Additive or Multiplicative

**SMILES for Training and Calibration sets**

☐ Graph ☐ HSG ☐ HFG ☐ GAO  
☐ ec0  
☐ pt2  
☐ pt3 vs2  
☐ C4 ☐ C5 ☐ C6 ☐ C7  
☐ NNC  
☐ BOND  
☐ NOSP  
☐ HALO  
☐ HARD  
☐ PAIR

**Classification model**

**Concordance Correlation Coefficient**

**Classic Scheme**

Balance of correlations dR weight \*\*\*  
 Real slopes dC weight \*\*\*  
 \*\*\* d<sub>limit</sub> \*\*\* N epoch \*\*\*

**Index of Ideality of correlation**

Threshold value..... \*\*\*  
 All threshold value..... \*\*\*  
 Number of the Monte Carlo probes..... \*\*\*

**Depth of Interpretation** \*\*\*

☐ Outliers 5  
☐ DemoDCW  
☐ EvolutionCorr

**Place of compound (CAS) in graphical representations**

☐ Search for duplicates in SMILES ☐ Search for duplicates in CAS (ID)

**Split Info**

|   | N100 | Na1 | DEFECT |
|---|------|-----|--------|
| 0 | 0    | 0   | 0      |

10. Save the percentage by click button

Save vector of Distribution

In the future, you can get this distribution by click button

Loading of vector of Distribution

11. Click button



**Preparation of split**

The first action is "Load". File "#TotalSet.txt" must exist in your folder.

The second action should be "Do distribution".

The third action should be "Save files".

|     | Input           | Output |                       |
|-----|-----------------|--------|-----------------------|
| +1. | #TotalSet.txt   | Load   | #TotalSet.txt N = 100 |
| +2. | Do distribution |        |                       |
| +3. | Save files      |        |                       |

Dispersion Limit = 0,01

|                        |   | Planned Distribution | Frequency | Actual Distribution |
|------------------------|---|----------------------|-----------|---------------------|
| #TrainingSet.txt       | + | 0,35                 | 35        | 0,350               |
| Invisible training set | - | 0,35                 | 35        | 0,350               |
| Calibration set        | # | 0,15                 | 14        | 0,140               |
| #ValidationSet.txt     | * | 0,15                 | 16        | 0,160               |

14. Click button

**EXIT**

**CORAL: Loading of method or system**

**Method:** Scheme: Additive or Multiplicative

**SMILES for Training and Calibration sets**

☐ Graph ☐ H3G ☐ HFG ☐ GAO

☐ ec0

☐ pt2 ☐ vs2

☐ pt3 ☐ NNC

☐ C4 ☐ C5 ☐ C6 ☐ C7

**Classification model**

☐ BOND ☐ NOSP ☐ HALO ☐ HARD ☐ PAIR

**Concordance Correlation Coefficient**

**Classic Scheme**

**Balance of correlations** dR weight \*\*\*

**Real slopes** dC weight \*\*\*

\*\*\* d limit \*\*\* N epoch \*\*\*

**Index of Ideality of correlation**

Threshold value..... \*\*\*

al threshold value..... \*\*\*

er of the Monte Carlo probes..... \*\*\*

**Depth of Interpretation** \*\*\*

☐ Outliers 5

☐ DemoDCW

☐ EvolutionCorr

**Preparation of split**

The first action is "Load". File "#TotalSet.txt" must exist in your folder.

The second action should be "Do distribution".

The third action should be "Save files".

|     | Input           | Output |                       |
|-----|-----------------|--------|-----------------------|
| +1. | #TotalSet.txt   | Load   | #TotalSet.txt N = 100 |
| +2. | Do distribution |        |                       |
| +3. | Save files      |        |                       |

Dispersion Limit = 0,01

|                        |   | Planned Distribution | Frequency | Actual Distribution |
|------------------------|---|----------------------|-----------|---------------------|
| #TrainingSet.txt       | + | 0,35                 | 35        | 0,350               |
| Invisible training set | - | 0,35                 | 35        | 0,350               |
| Calibration set        | # | 0,15                 | 14        | 0,140               |
| #ValidationSet.txt     | * | 0,15                 | 16        | 0,160               |

**Training set**

**Invisible Training set**

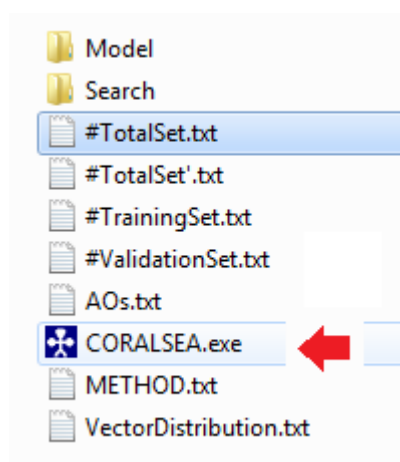
**Calibration set**

**Place of compound (CAS) in graphical representations**

☐ Search for duplicates in SMILES ☐ Search for duplicates in CAS (ID)

**Split Info** 0 0 0 0 0 0 0 0

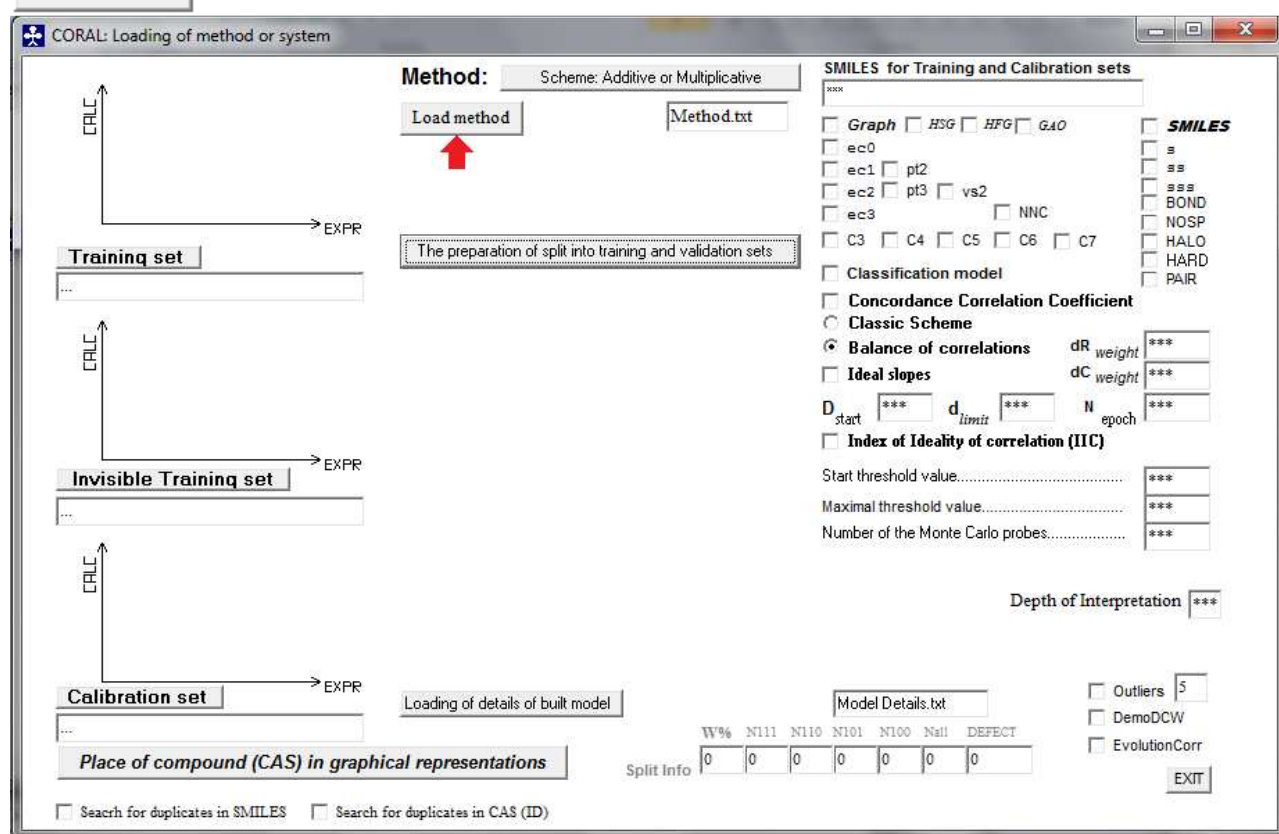
Now you have system to build up a QSPR/QSAR model:



Run CORALSEA.exe

15. Click button

Load method



Now you need define your method. Please read Table “Components of methods”. In order to discuss the applications of the CORAL software the next method (please see next page) will be considered as method selected for building up models.



CORAL: select Phase 1, Phase 2, or change and save method

**Method:** Adding

Load method Save method Method.txt

Phase 1: Search for preferable model (T\*,N\*)

The preparation of split into training and validation sets

Phase 2: Building up preferable model (T\*,N\*)

Loading of details of built model

Place of compound (CAS) in graphical representations

Training set

Invisible Training set

Calibration set

SMILES for Training and Calibration sets

#TrainingSet.txt

☐ Graph ☐ HSG ☐ HFG ☐ GAO

☐ ec0 ☐ ec1 ☐ pt2 ☐ ec2 ☐ pt3 ☐ vs2 ☐ ec3 ☐ NNC

☐ C3 ☐ C4 ☐ C5 ☐ C6 ☐ C7

☐ Classification model

☐ Concordance Correlation Coefficient

☐ Classic Scheme

☒ Balance of correlations dR weight 0,1

☐ Ideal slopes

D\_start 0,5 d\_limit 0,1 N\_epoch 25

☒ Index of Ideality of correlation (IIC) 0,1

Start threshold value..... 1

Maximal threshold value..... 3

Number of the Monte Carlo probes..... 2

Depth of Interpretation 10

☒ Outliers 5

☒ DemoDCW

☒ EvolutionCorr

EXIT

Model Details.txt

| W% | N111 | N110 | N101 | N100 | Nall | DEFECT |
|----|------|------|------|------|------|--------|
| 0  | 0    | 0    | 0    | 0    | 0    | 0      |

Split Info

☒ Search for duplicates in SMILES ☒ Search for duplicates in CAS (ID)

Table “Components of methods”

*The selection of molecular features extracted from graph or SMILES*

- ☐ Graph ☐ HSG ☐ HFG ☐ GAO
- ☐ ec0
- ☐ ec1 ☐ pt2
- ☐ ec2 ☐ pt3 ☐ vs2
- ☐ ec3 ☐ NNC
- ☐ C3 ☐ C4 ☐ C5 ☐ C6 ☐ C7

| No. | Invariants of molecular graphs HSG, HFG, GAO*  | Comment  |
|-----|--|--|
| 1   | $EC0_k = \sum_{(kj)edge} a_{kj}$   | Vertex degree  |
| 2   | $EC1_k = \sum_{(kj)edge} EC0_j$  | Extended connectivity of the first order   |
| 3   | $EC2_k = \sum_{(kj)edge} EC1_j$  | Extended connectivity of the second order  |
| 4   | $PT2_k = \sum_{(kj)edge} \sum_{(jl)edge, l \neq k} a_{kl}$                                     | The number of paths of length 2 which starts from k-vertex. The $a_{kl}$ is element of adjacency matrix. |
| 5   | $PT3_k = \sum_{(kj)edge} \sum_{(jl)edge, l \neq k} \sum_{(lm)edge, m \neq j, m \neq k} a_{km}$ | The number of paths of length 3 which starts from k-vertex   |
| 6   | $S2_k = \sum_{(kj)edge} \sum_{(jl)edge, l \neq k} EC0_l$                                       | Valence shell of the second order  |

|   |  |  |
|---|--|--|
| 7 | $S3_k = \sum_{(kj)edge} \sum_{(jl)edge, l \neq k} \sum_{(lm)edge, m \neq j, m \neq k} ECO_m$ | Valence shell of the third order   |
| 8 | $NNC_k = N_{total} \times 100 + N_{carbon} \times 10 + N_{non-carbon}$                       | The nearest neighboring code, $N_{total}$ is total number of neighbors of k-the vertex; $N_{carbon}$ is number of neighbors vertices represent carbon atoms; and $N_{non-carbon}$ is number of neighbors vertices represent non-carbon atoms |
| 9 | C3, C4, C5, C6, C7   | Specificity of rings (size 3,4,5,6,7): (i) presence/absence of aromaticity; (ii) presence/absence of heteroatoms (atoms $\neq$ carbon)   |

\*) HSG = hydrogen suppressed graph; HFG = hydrogen filled graph; GAO = graph of atomic orbitals

☒ **SMILES**

☒ S

☒ SS

☒ SSS

☐ BOND

☐ NOSP

☐ HALO

☒ HARD

☐ PAIR

| No. | Features extracted from SMILES = ABCDE... | Comment   |
|-----|---|---|
| 1   | S   | A, B, C, D, E   |
| 2   | SS  | AB, BC, CD, DE  |
| 3   | SSS                                       | ABC, BCD, CDE   |
| 4   | BOND                                      | Presence/absence (i) double bonds [denoted '=']; (ii) triple bonds [denoted '#']; (iii) stereo-chemical (3D) bonds [denoted '@' and '@@'] |
| 5   | NOSP                                      | Presence/absence of (i) nitrogen ['N']; (ii) oxygen ['O']; (iii) Sulphur ['S']; and (iv) phosphorus ['P']                                 |
| 6   | HALO                                      | Presence/absence of (i) fluorine ['F']; (ii) chlorine ['Cl']; (iii) bromine ['Br']; and (iv) iodine ['I']                                 |
| 7   | HARD                                      | Association of BOND, NOSP, and HALO   |
| 8   | PAIR                                      | The presence of any pair of molecular features from   |

|  |  |  |
|--|--|--|
|  |  | BOND, NOSP, HALO together, e.g. double bond and chlorine; nitrogen and bromine, triple bond and phosphorus, etc. |
|--|--|--|

*The selection of regression model “endpoint = intercept + slope\*descriptor” (Default)*

☐ Classification model

*The selection of classification model YES/NO or active/inactive*

☒ Classification model

*The selection of criterion to build up a model (CBM):*

*Version 1 = Pearson correlation coefficient (Default)*

☐ Concordance Correlation Coefficient

*Version 2 = concordance correlation coefficient*

☒ Concordance Correlation Coefficient

*Building up a model is optimization of target function. There is a hierarchy of target functions.*

(i)

☐ Classic Scheme

☒ Balance of correlations

dR<sub>weight</sub>

*User should select **Classic Scheme** or **Balance of correlations**.*

*Target function for classic scheme is version 1 or version 2 of CBM. In fact, the Monte Carlo optimization gives maximum of CBM solely for the training set.*

*Balance of correlation is using of target function in form*

$$TF = CBM + CBM' + \text{abs}(CBM - CBM') * dR_{\text{weight}} + \text{abs}(\text{slope} - \text{slope}') * dC_{\text{weight}}$$

(ii)

*Parameters of the Monte Carlo optimization are*

*D<sub>start</sub> is definition how much should be change a correlation weight of molecular feature*

*D<sub>limit</sub> is definition of precession of changes for molecular features*

*Nepoch is number of iterations of the Monte Carlo optimization: one epoch is modification of correlation weights for all molecular features involved in modeling process.*

D<sub>start</sub>  d<sub>limit</sub>  N<sub>epoch</sub>

(iii)

*Index of ideality of correlation can be involved to define improved target function:*

$$TF_m = TF + IIC * IIC_{\text{weight}}$$

☒ Index of Ideality of correlation IIC<sub>weight</sub>

*User can switch off this component*

☐ Index of Ideality of correlation

16. Having defined method, user should save the method by click button

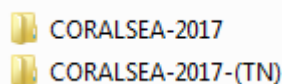
!

When steps 1-16 are completed, user can start building up model for his / her endpoint. The modeling process involves

1. Definition of the method.
2. Definition T\* and N\* which give best statistical quality for the calibration set.
3. Building up the model using T\* and N\* value.
4. Extraction of molecular features, which are promoters of increase or decrease for the endpoint.

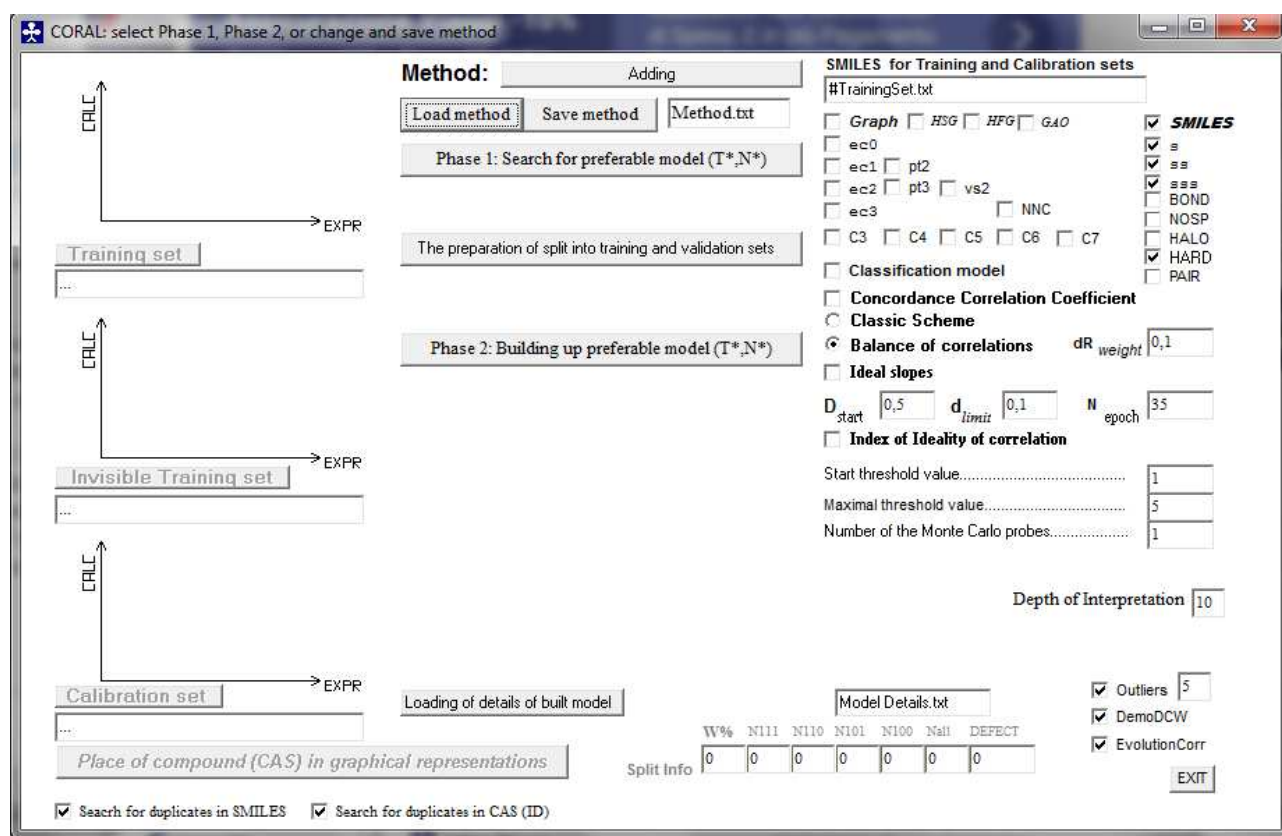
### 1. Definition of the method.

It is a good idea to prepare a copy of CORALSEA-2017 folder, e.g. CORALSEA-2017-(TN)



Run CORALSEA.exe from folder "CORALSEA-2017-(TN)"

The following method has been selected:



### 2. Definition T\* and N\* which give best statistical quality for the calibration set.

The T\* expected from range (1, 2, 3, 4, 5).



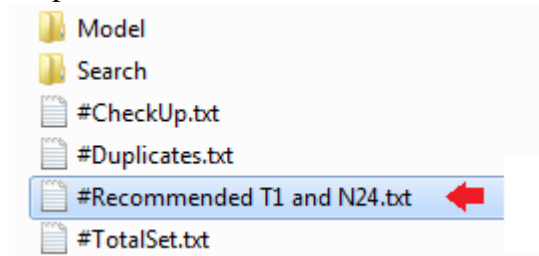
The  $N^*$  expected from range (1, 2, 3, ..., 34, 35).

In order to obtain the  $T^*$  and  $N^*$ , click button

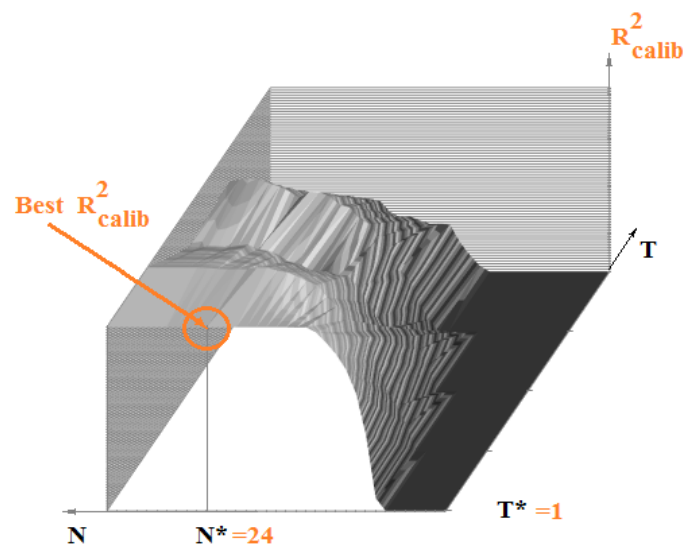
Phase 1: Search for preferable model ( $T^*, N^*$ )

The  $T^*=1$  and  $N^*=24$  were obtained in the described example (in your case it will be other values).

A special file contains this result

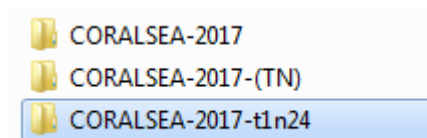


This results can be illustrated as the following



### 3. Building up the model using $T^*$ and $N^*$ value.

Again, it is a good idea to prepare a copy of CORALSEA-2017-(TN) e.g. CORALSEA-t1n24



Run CORALSEA from the folder CORALSEA-2017-t1n24

CORAL: Loading of method or system

**Method:** Scheme: Additive or Multiplicative

**Load method** **Method.txt**

The preparation of split into training and validation sets

**SMILES for Training and Calibration sets**

☐ Graph ☐ H3G ☐ HFG ☐ G4O

☐ ec0 ☐ pt2

☐ ec2 ☐ pt3 ☐ vs2

☐ ec3 ☐ NNC

☐ C3 ☐ C4 ☐ C5 ☐ C6 ☐ C7

☐ Classification model

☐ Concordance Correlation Coefficient

☐ Classic Scheme

☒ Balance of correlations dR weight

☐ Ideal slopes dC weight

D\_start  d\_limit  N\_epoch

☐ Index of Ideality of correlation

Start threshold value.....

Maximal threshold value.....

Number of the Monte Carlo probes.....

Depth of Interpretation

**Training set** **EXPR**

**Invisible Training set** **EXPR**

**Calibration set** **EXPR**

Loading of details of built model

**Model Details.txt**

☐ Outliers

☐ DemoDCW

☐ EvolutionCorr

**Place of compound (CAS) in graphical representations**

Split Info

| W% | N111 | N110 | N101 | N100 | Nall | DETECT |
|----|------|------|------|------|------|--------|
| 0  | 0    | 0    | 0    | 0    | 0    | 0      |

☐ Search for duplicates in SMILES ☐ Search for duplicates in CAS (ID)

**EXIT**

Click

**Load method**

Instead

N\_epoch

One should use

N\_epoch

Click

**Save method**

CORAL: select Phase 1, Phase 2, or change and save method

**Method:** Adding

**Load method** **Save method** **Method.txt**

**Phase 1: Search for preferable model (T\*,N\*)**

The preparation of split into training and validation sets

**Phase 2: Building up preferable model (T\*,N\*)**

**SMILES for Training and Calibration sets**

#TrainingSet.txt

☐ Graph ☐ H3G ☐ HFG ☐ G4O

☒ ec0 ☐ pt2

☒ ec2 ☐ pt3 ☐ vs2

☒ ec3 ☐ NNC

☐ C3 ☐ C4 ☐ C5 ☐ C6 ☐ C7

☐ Classification model

☐ Concordance Correlation Coefficient

☐ Classic Scheme

☒ Balance of correlations dR weight

☐ Ideal slopes

D\_start  d\_limit  N\_epoch

☒ Index of Ideality of correlation IIC\_weight

Start threshold value.....

Maximal threshold value.....

Number of the Monte Carlo probes.....

Depth of Interpretation

**Training set** **EXPR**

**Invisible Training set** **EXPR**

**Calibration set** **EXPR**

Loading of details of built model

**Model Details.txt**

☒ Outliers

☒ DemoDCW

☒ EvolutionCorr

**Place of compound (CAS) in graphical representations**

Split Info

| W% | N111 | N110 | N101 | N100 | Nall | DETECT |
|----|------|------|------|------|------|--------|
| 0  | 0    | 0    | 0    | 0    | 0    | 0      |

☒ Search for duplicates in SMILES ☒ Search for duplicates in CAS (ID)

**EXIT**

Click

Phase 2: Building up preferable model (T\*,N\*)

Click

Define preferable threshold and press Continue  Continue

CORAL: Building up preferable model

Method: Adding

SMILES for Training and Calibration sets

#TrainingSet.txt

☐ Graph ☐ H3G ☐ HFG ☐ GAO

☐ ec0 ☐ pt2

☐ ec1 ☐ pt3 ☐ vs2

☐ ec2 ☐ NNC

☐ ec3 ☐ C3 ☐ C4 ☐ C5 ☐ C6 ☐ C7

☐ Classification model

☐ Concordance Correlation Coefficient

☐ Classic Scheme

☒ Balance of correlations dR\_weight 0,1

☐ Ideal slopes

D\_start 0,5 d\_limit 0,1 N\_epoch 35

☒ Index of Ideality of correlation IIC\_weight 0,1

Start threshold value..... 1

Maximal threshold value..... 5

Number of the Monte Carlo probes..... 1

Depth of Interpretation 10

Outliers 5

DemoDCV

EvolutionCorr

EXIT

Training set

EXPR

Invisible Training set

EXPR

Calibration set

EXPR

Phase 2: Building up preferable model (T\*,N\*)

Define preferable threshold and press Continue  Continue

C0 = 0 C1 = 1

Loading of details of built model

Model Details.txt

W% N111 N110 N101 N100 Nall DEFECT

Split Info 0 0 0 0 0 0 0

Place of compound (CAS) in graphical representations

☒ Search for duplicates in SMILES ☒ Search for duplicates in CAS (ID)

Click Yes or NO:

CORAL: Building up preferable model

Method: Adding

SMILES for Training and Calibration sets  
#TrainingSet.txt

☐ Graph ☐ HSG ☐ HFG ☐ GAO

☐ ec0 ☐ pt2 ☐ vs2

☐ ec1 ☐ pt3 ☐ NNC

☐ ec2 ☐ C3 ☐ C4 ☐ C5 ☐ C6 ☐ C7

☐ ec3

☐ Classification model

☐ Concordance Correlation Coefficient

☐ Classic Scheme

☒ Balance of correlations dR\_weight 0,1

☐ Ideal slopes

☒ Index of Ideality of correlation IIC\_weight 0,1

D\_start 0,5 d\_limit 0,1 N\_epoch 35

threshold value..... 1

maximal threshold value..... 1

number of the Monte Carlo probes..... 1

Depth of Interpretation 10

☒ Outliers 5

☒ DemoDCW

☒ EvolutionCorr

EXIT

Phase 2: Building up preferable model (T\*,N\*)

Define preferable threshold and press Continue 1 Continue

C0 = 0 C1 = 1

Training set

Invisible Training set

Calibration set

Place of compound (CAS) in graphical representations

Split Info

| W% | N111 | N110 | N101 | N100 | Nall | DEFECT |
|----|------|------|------|------|------|--------|
| 0  | 0    | 0    | 0    | 0    | 0    | 0      |

☒ Search for duplicates in SMILES ☒ Search for duplicates in CAS (ID)

Confirm

There are files in "model/\*.\*" which remain after previous calculations. You can delete these files in order to avoid mixture of new files and files which remain after previous calculations. Delete these files?

Yes No Cancel

Gradually the Monte Carlo optimization is building up a model...

CORAL: Wait please...

Method: Adding

SMILES for Training and Calibration sets  
#TrainingSet.txt

☐ Graph ☐ HSG ☐ HFG ☐ GAO

☐ ec0 ☐ pt2 ☐ vs2

☐ ec1 ☐ pt3 ☐ NNC

☐ ec2 ☐ C3 ☐ C4 ☐ C5 ☐ C6 ☐ C7

☐ ec3

☐ Classification model

☐ Concordance Correlation Coefficient

☐ Classic Scheme

☒ Balance of correlations dR\_weight 0,1

☐ Ideal slopes

☒ Index of Ideality of correlation IIC\_weight 0,1

D\_start 0,5 d\_limit 0,1 N\_epoch 24

Start threshold value..... 1

Maximal threshold value..... 1

Number of the Monte Carlo probes..... 1

Depth of Interpretation 10

☒ Outliers 5

☒ DemoDCW

☒ EvolutionCorr

EXIT

Phase 2: Building up preferable model (T\*,N\*)

Define preferable threshold and press Continue 1 Continue

C0 = 0 C1 = 1

Training set

Invisible Training set

Calibration set

Place of compound (CAS) in graphical representations

Split Info

| W% | N111 | N110 | N101 | N100 | Nall | DEFECT  |
|----|------|------|------|------|------|---------|
| 78 | 376  | 39   | 42   | 28   | 483  | 67,8960 |

☒ Search for duplicates in SMILES ☒ Search for duplicates in CAS (ID)

Selected threshold is 1

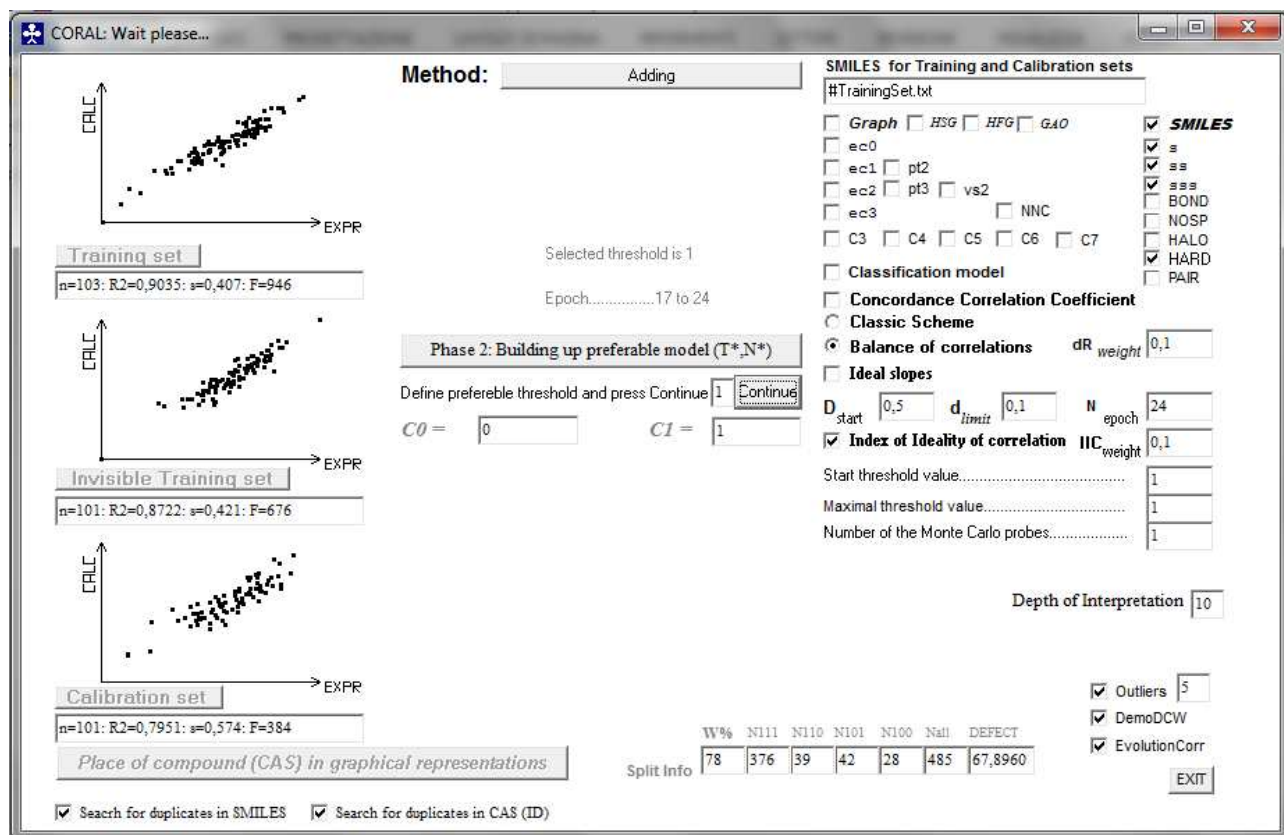
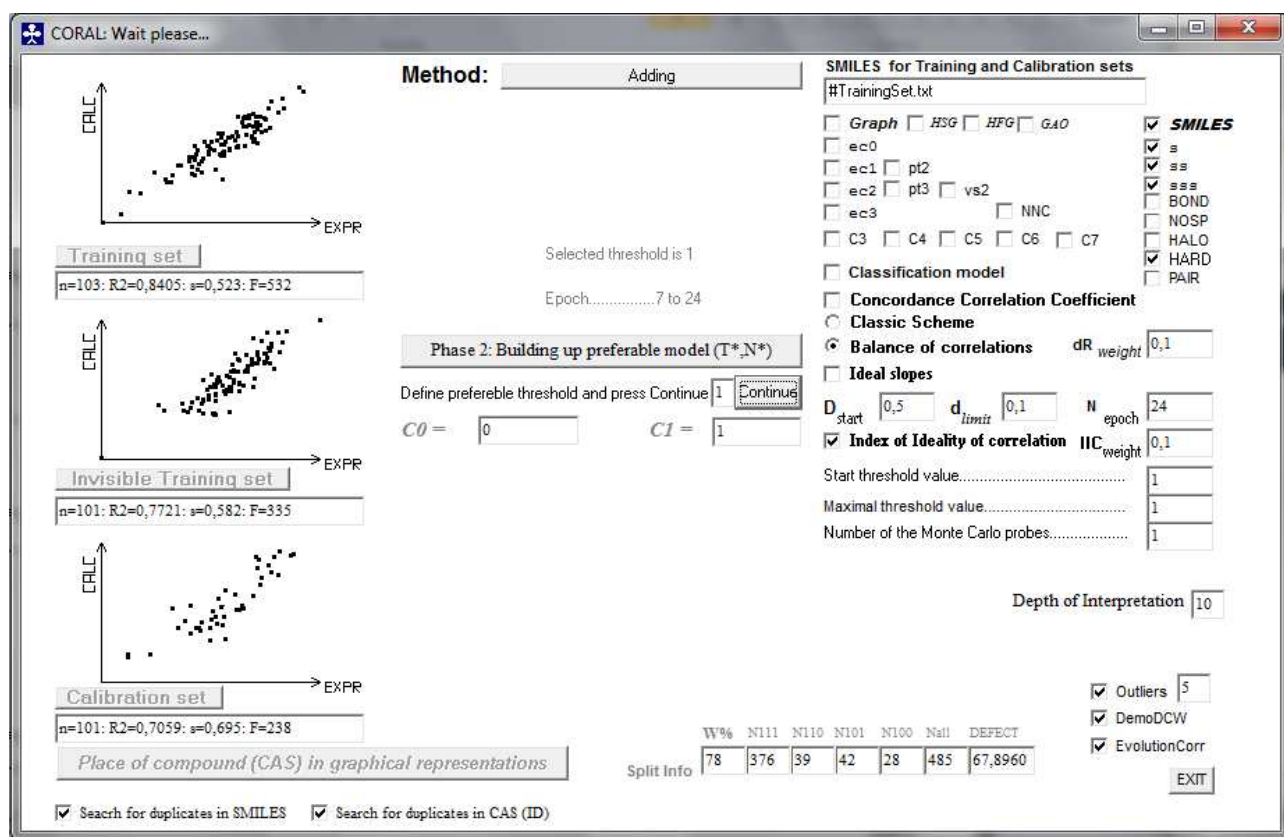
Epoch..... 2 to 24

n=103: R2=0,4706: s=0,953: F=90

n=101: R2=0,4654: s=0,847: F=86

n=101: R2=0,3806: s=1,01: F=61





Finally, the optimization is completed: it is necessary to save the model. Please click

Save Model Details

CORAL: you must save system now

**Method:** Adding

**SMILES for Training and Calibration sets**  
#TrainingSet.txt

☐ Graph ☐ HSG ☐ HFG ☐ GAO

☐ ec0 ☐ ec1 ☐ pt2 ☐ ec2 ☐ pt3 ☐ vs2 ☐ ec3 ☐ NNC

☐ C3 ☐ C4 ☐ C5 ☐ C6 ☐ C7

☐ Classification model

☐ Concordance Correlation Coefficient

☐ Classic Scheme

☒ Balance of correlations  $dR_{weight}$  0,1

☐ Ideal slopes

$D_{start}$  0,5  $d_{limit}$  0,1  $N_{epoch}$  24

☒ Index of Ideality of correlation  $IIC_{weight}$  0,1

Start threshold value..... 1

Maximal threshold value..... 1

Number of the Monte Carlo probes..... 1

DCW-calculation will be saved in file: DemoDCW.txt

Depth of Interpretation 10

**Training set**  
n=103: R2=0,9116: s=0,389: F=1042

**Invisible Training set**  
n=101: R2=0,8796: s=0,407: F=723

**Calibration set**  
n=101: R2=0,7984: s=0,570: F=392

**Place of compound (CAS) in graphical representations**

Selected threshold is 1

THE CALCULATION IS COMPLETED

**Phase 2: Building up preferable model (T\*,N\*)**

Define preferable threshold and press Continue 1

$C0 =$  4,4891984  $C1 =$  0,0359822

Insert a SMILES for calculation of DCW and EndPoint

Start of DCW' and Endpoint Calculation for inserted SMILES

Start of DCW' and Endpoint calculation for SMILES from file

Model Details.txt

☒ Outliers 5

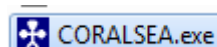
☒ DemoDCW

☒ EvolutionCorr

☒ Search for duplicates in SMILES ☒ Search for duplicates in CAS (ID)

| W% | N111 | N110 | N101 | N100 | Nall | DETECT  |
|----|------|------|------|------|------|---------|
| 78 | 376  | 39   | 42   | 28   | 485  | 67,8960 |

In order to use the model for validation set run CORALSEA.exe



and click Loading of details of built model

CORAL: Loading of method or system

**Method:** Scheme: Additive or Multiplicative

Method.txt

**SMILES for Training and Calibration sets**  
xxx

☐ Graph ☐ HSG ☐ HFG ☐ GAO

☐ ec0 ☐ ec1 ☐ pt2 ☐ ec2 ☐ pt3 ☐ vs2 ☐ ec3 ☐ NNC

☐ C3 ☐ C4 ☐ C5 ☐ C6 ☐ C7

☐ Classification model

☐ Concordance Correlation Coefficient

☐ Classic Scheme

☒ Balance of correlations  $dR_{weight}$  \*\*\*  $dC_{weight}$  \*\*\*

☐ Ideal slopes

$D_{start}$  \*\*\*  $d_{limit}$  \*\*\*  $N_{epoch}$  \*\*\*

☐ Index of Ideality of correlation

Start threshold value..... \*\*\*

Maximal threshold value..... \*\*\*

Number of the Monte Carlo probes..... \*\*\*

Depth of Interpretation \*\*\*

**Training set**  
...

**Invisible Training set**  
...

**Calibration set**  
...

**Place of compound (CAS) in graphical representations**

The preparation of split into training and validation sets

Loading of details of built model

☐ Search for duplicates in SMILES ☐ Search for duplicates in CAS (ID)

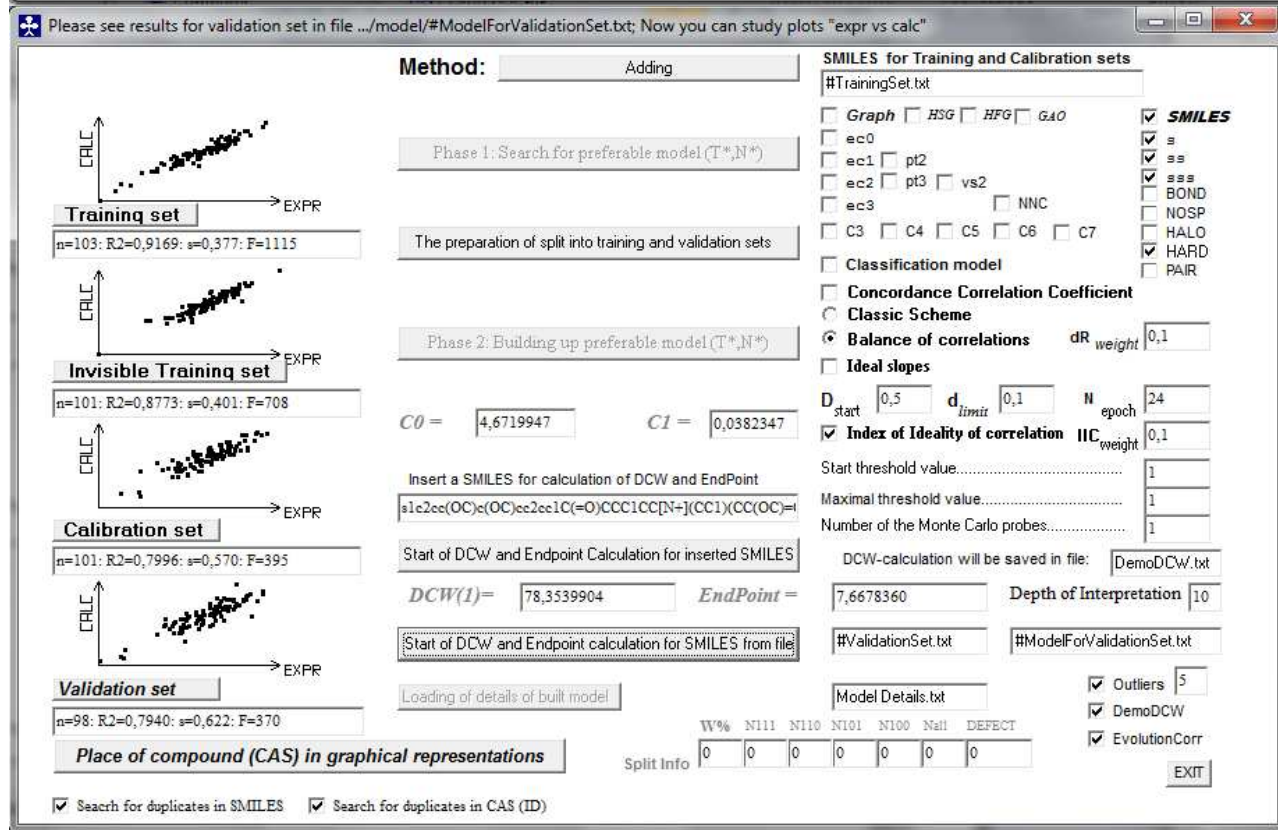
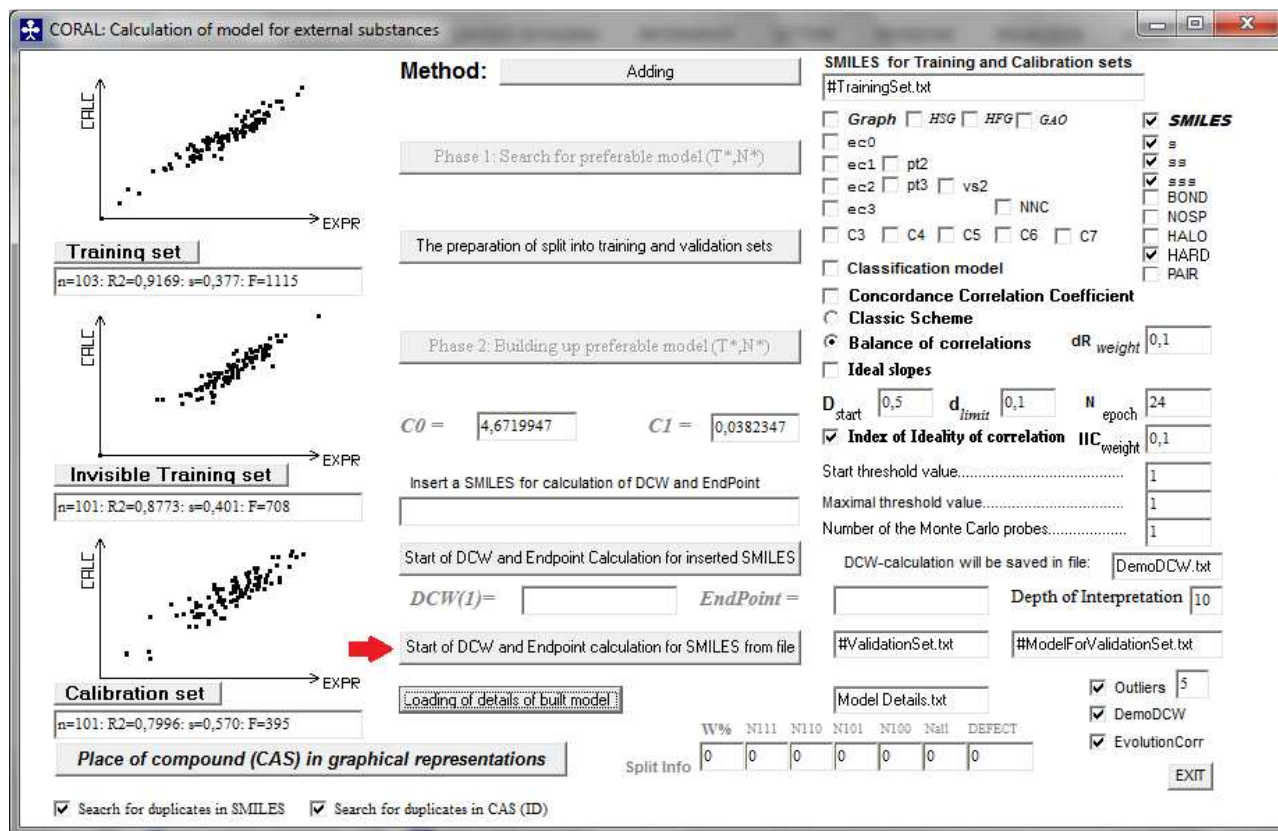
| W% | N111 | N110 | N101 | N100 | Nall | DETECT |
|----|------|------|------|------|------|--------|
| 0  | 0    | 0    | 0    | 0    | 0    | 0      |

Click

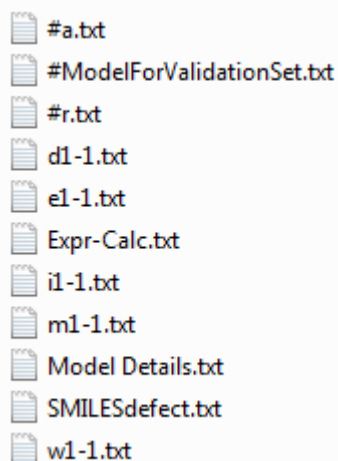
Start of DCW and Endpoint calculation for SMILES from file

#ValidationSet.txt

#ModelForValidationSet.txt



When these actions are completed folder, Model contains the following list of files



#a.txt  
#ModelForValidationSet.txt  
#r.txt  
d1-1.txt  
e1-1.txt  
Expr-Calc.txt  
i1-1.txt  
m1-1.txt  
Model Details.txt  
SMILESdefect.txt  
w1-1.txt

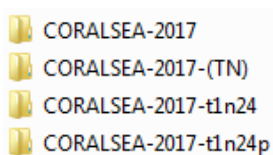
The #ModelForValidationSet.txt contains statistical characteristics of the model together with observed and calculated endpoint values for validation set.

The m1-1.txt contains statistical characteristics of the model together with observed and calculated endpoint values for training set, invisible training set, calibration set.

#### 4. Extraction of molecular features, which are promoters of increase or decrease for the endpoint.

Again, prepare a copy of CORALSEA-2017-t1n24 e.g. CORALSEA-t1n24p

The t1n24p means, “The promoters of increase and decrease for endpoint”, if  $T^*=1$  and  $N^*=24$



CORALSEA-2017  
CORALSEA-2017-(TN)  
CORALSEA-2017-t1n24  
CORALSEA-2017-t1n24p

Run CORALSEA.exe in folder “CORALSEA-2017-t1n24p” and Click “Load method”



CORAL: select Phase 1, Phase 2, or change and save method

**Method:** Adding

Load method Save method Method.txt

Phase 1: Search for preferable model (T\*,N\*)

The preparation of split into training and validation sets

Phase 2: Building up preferable model (T\*,N\*)

Loading of details of built model

Model Details.txt

W% N111 N110 N101 N100 Nall DEFECT

Split Info 0 0 0 0 0 0 0

SMILES for Training and Calibration sets

#TrainingSet.txt

☐ Graph ☐ HSG ☐ HFG ☐ GAO

☐ ec0 ☐ ec1 ☐ pt2 ☐ ec2 ☐ pt3 ☐ vs2 ☐ ec3 ☐ NNC

☐ C3 ☐ C4 ☐ C5 ☐ C6 ☐ C7

☐ Classification model

☐ Concordance Correlation Coefficient

☐ Classic Scheme

☒ Balance of correlations dR weight 0,1

☐ Ideal slopes

D\_start 0,5 d\_limit 0,1 N\_epoch 24

☒ Index of Ideality of correlation IIC\_weight 0,1

Start threshold value..... 1

Maximal threshold value..... 5

Number of the Monte Carlo probes..... 1

Depth of Interpretation 10

☒ Outliers 5

☒ DemoDCW

☒ EvolutionCorr

EXIT

☒ Search for duplicates in SMILES ☒ Search for duplicates in CAS (ID)

Training set

Calc

Expr

Invisible Training set

Calc

Expr

Calibration set

Calc

Expr

Place of compound (CAS) in graphical representations

The task is to carry out three Runs of the Monte Carlo optimization in order to select molecular features which are stable promoters of increase (every time positive correlation weight) and decrease (every time negative correlation weight) for the endpoint. In addition, these molecular features should have significant prevalence in the training set. Thus,  
 Start threshold value is 1; Maximal threshold values is 1;  
 Number of the Monte Carlo probes is 3 (it is possible 4, 5, ... , 10)

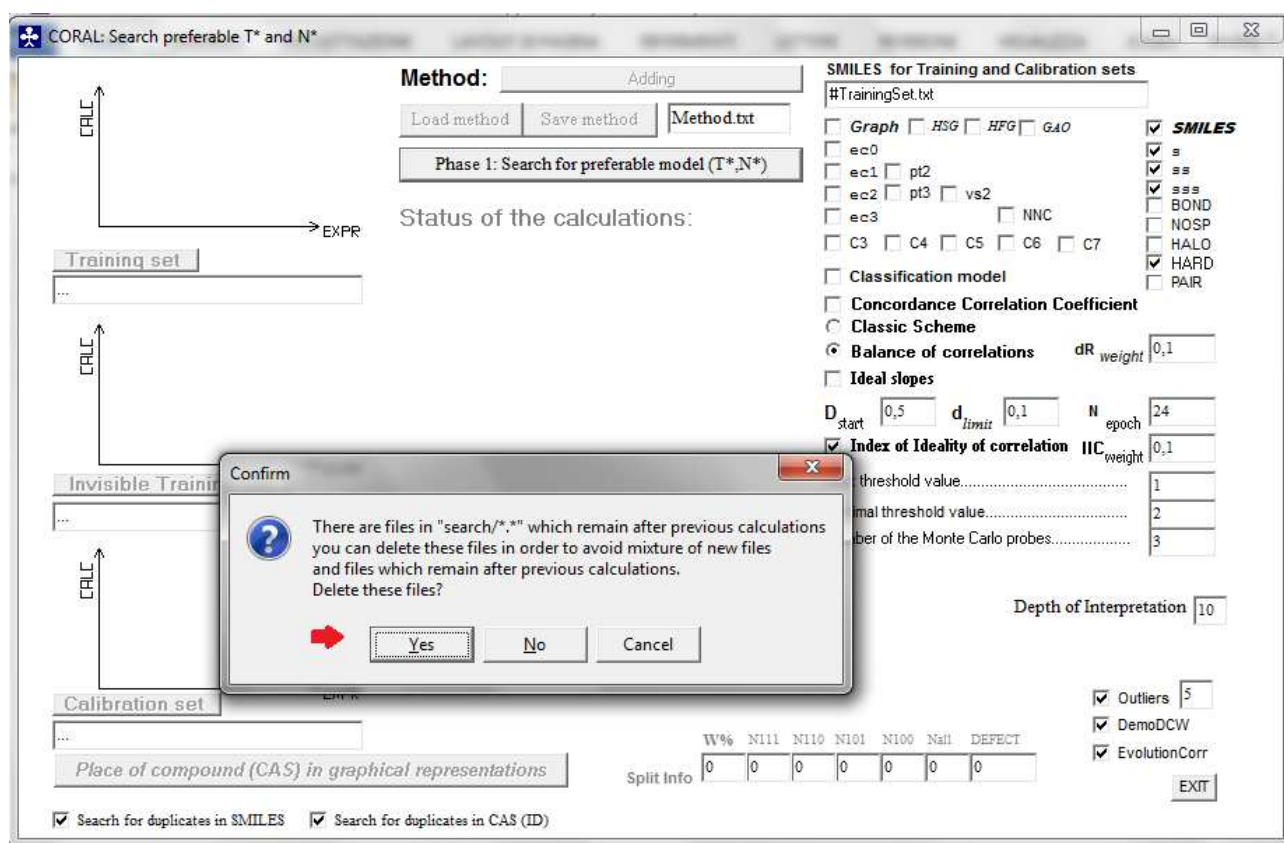
Start threshold value..... 1

Maximal threshold value..... 1

Number of the Monte Carlo probes..... 3

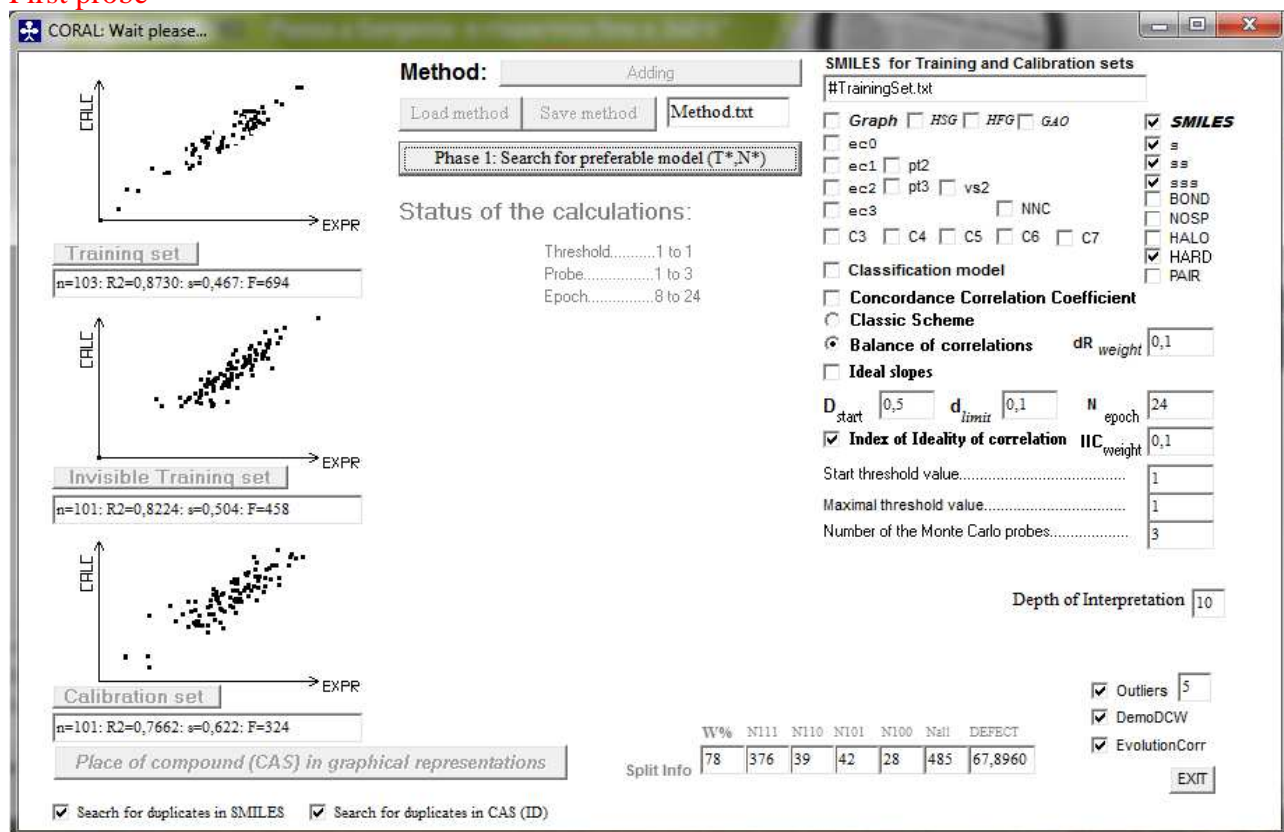
Click button

Phase 1: Search for preferable model (T\*,N\*)



Click Yes or No

First probe



## Second probe

CORAL: Wait please...

**Method:** Adding

Load method Save method Method.txt

**Phase 1: Search for preferable model (T\*,N\*)**

**Status of the calculations:**

Threshold.....1 to 1  
Probe.....2 to 3  
Epoch.....15 to 24

**SMILES for Training and Calibration sets**

#TrainingSet.txt

☐ Graph ☐ HSG ☐ HFG ☐ GAO

☐ ec0 ☐ ec1 ☐ pt2 ☐ ec2 ☐ pt3 ☐ vs2 ☐ ec3 ☐ NNC

☐ C3 ☐ C4 ☐ C5 ☐ C6 ☐ C7

☐ Classification model

☐ Concordance Correlation Coefficient

☐ Classic Scheme

☒ Balance of correlations dR\_weight 0,1

☐ Ideal slopes

D\_start 0,5 d\_limit 0,1 N\_epoch 24

☒ Index of Ideality of correlation IIC\_weight 0,1

Start threshold value.....1

Maximal threshold value.....1

Number of the Monte Carlo probes.....3

Depth of Interpretation 10

☒ Outliers 5

☒ DemoDCW

☒ EvolutionCorr

EXIT

**Training set**  
n=103: R2=0,9008: s=0,412: F=917

**Invisible Training set**  
n=101: R2=0,8626: s=0,437: F=622

**Calibration set**  
n=101: R2=0,7706: s=0,607: F=333

Place of compound (CAS) in graphical representations

☒ Search for duplicates in SMILES ☒ Search for duplicates in CAS (ID)

| W% | N111 | N110 | N101 | N100 | Nall | DEFECT  |
|----|------|------|------|------|------|---------|
| 78 | 376  | 39   | 42   | 28   | 483  | 67,8960 |

Split Info

## Third probe

CORAL: Wait please...

**Method:** Adding

Load method Save method Method.txt

**Phase 1: Search for preferable model (T\*,N\*)**

**Status of the calculations:**

Threshold.....1 to 1  
Probe.....3 to 3  
Epoch.....16 to 24

**SMILES for Training and Calibration sets**

#TrainingSet.txt

☐ Graph ☐ HSG ☐ HFG ☐ GAO

☐ ec0 ☐ ec1 ☐ pt2 ☐ ec2 ☐ pt3 ☐ vs2 ☐ ec3 ☐ NNC

☐ C3 ☐ C4 ☐ C5 ☐ C6 ☐ C7

☐ Classification model

☐ Concordance Correlation Coefficient

☐ Classic Scheme

☒ Balance of correlations dR\_weight 0,1

☐ Ideal slopes

D\_start 0,5 d\_limit 0,1 N\_epoch 24

☒ Index of Ideality of correlation IIC\_weight 0,1

Start threshold value.....1

Maximal threshold value.....1

Number of the Monte Carlo probes.....3

Depth of Interpretation 10

☒ Outliers 5

☒ DemoDCW

☒ EvolutionCorr

EXIT

**Training set**  
n=103: R2=0,8996: s=0,415: F=905

**Invisible Training set**  
n=101: R2=0,8736: s=0,416: F=684

**Calibration set**  
n=101: R2=0,8022: s=0,563: F=402

Place of compound (CAS) in graphical representations

☒ Search for duplicates in SMILES ☒ Search for duplicates in CAS (ID)

| W% | N111 | N110 | N101 | N100 | Nall | DEFECT  |
|----|------|------|------|------|------|---------|
| 78 | 376  | 39   | 42   | 28   | 483  | 67,8960 |

Split Info

CORAL: Wait please...

**Method:** Adding  
 Load method Save method Method.txt

**Phase 1: Search for preferable model (T\*,N\*)**

**Status of the calculations:**  
 Threshold.....1 to 1  
 Probe.....3 to 3  
 THE CALCULATION IS COMPLETED

**Phase 2: Building up preferable model (T\*,N\*)**

**Training set**  
 n=103: R2=0,9032: s=0,408: F=942

**Invisible Training set**  
 n=101: R2=0,8776: s=0,409: F=71

**Calibration set**  
 n=101: R2=0,8055: s=0,559: F=410

**SMILES for Training and Calibration sets**  
 #TrainingSet.txt

☐ Graph ☐ HSG ☐ HFG ☐ GAO ☒ SMILES  
☐ ec0 ☐ pt2 ☒ s  
☐ ec1 ☐ pt3 ☒ ss  
☐ ec2 ☐ vs2 ☒ sss  
☐ ec3 ☐ NNC ☐ BOND  
☐ C3 ☐ C4 ☐ C5 ☐ C6 ☐ C7 ☐ NOSP  
☐ PAIR

☐ Classification model

☐ Concordance Correlation Coefficient

☐ Classic Scheme

☒ Balance of correlations dR\_weight 0,1

☐ Ideal slopes

D\_start 0,5 d\_limit 0,1 N\_epoch 24

☒ Index of Ideality of correlation IIC\_weight 0,1

Start threshold value.....1

Maximal threshold value.....1

Number of the Monte Carlo probes.....3

Depth of Interpretation 10

☒ Outliers 5

☒ DemoDCW

☒ EvolutionCorr

**Place of compound (CAS) in graphical representations**

Split Info

| W% | N111 | N110 | N101 | N100 | Nall | DETECT  |
|----|------|------|------|------|------|---------|
| 78 | 376  | 39   | 42   | 28   | 485  | 67,8960 |

☒ Search for duplicates in SMILES ☒ Search for duplicates in CAS (ID)

OK

In order to select preferable threshold (T\*), please see files  
 - search/#a.txt (average statistical characteristics)  
 - search/#r.txt (all statistical characteristics);  
 In order to select preferable number of epochs (N\*), please see  
 - search/#BestMDL.txt

This message confirm that selection of the promoters of increase and decrease for the endpoint is completed. Results of the calculation are available in file "p1.txt" in folder "Search":

This file contains promoters of increase/decrease for endpoint according to selected "depth of interpretation"  
 NSs, NSc, and NSv are the number of SMILES which contain given attribute (SAk) in training, invisible training, and in calibration sets, respectively  
 Defect[SAk] is the difference of probabilities of SAk in training and calibration sets, divided by sum of total numbers of the SAk in the training and calibration sets  
 If attribute SAk absent in training set then defect=1 (maximum)

| No. : | ID : SAk             | CWs Probe 1: | CWs Probe 2: | CWs Probe 3: | NSs : | NSc : | NSv : | Defect[SAk]: |
|-------|----------------------|--------------|--------------|--------------|-------|-------|-------|--------------|
| 1:    | 269:N.....           | 2.18569:     | 0.68922:     | 0.31048:     | 100:  | 98:   | 100:  | 0.0001       |
| 2:    | 468:c...()           | 0.12065:     | 0.00260:     | 0.12031:     | 97:   | 95:   | 92:   | 0.0002       |
| 3:    | 546:c...c...1...     | 1.75409:     | 0.37794:     | 1.99843:     | 86:   | 84:   | 86:   | 0.0001       |
| 4:    | 74:2.....            | 1.12629:     | 1.49703:     | 0.68858:     | 82:   | 81:   | 88:   | 0.0004       |
| 5:    | 324:O...=.....       | 0.18868:     | 0.37085:     | 0.49539:     | 81:   | 79:   | 80:   | 0.0000       |
| 6:    | 474:c...()C...       | 1.93286:     | 1.87042:     | 3.18874:     | 81:   | 80:   | 75:   | 0.0003       |
| 7:    | 498:c...c...1...c... | 0.25323:     | 0.43543:     | 0.12177:     | 81:   | 77:   | 85:   | 0.0003       |
| 8:    | 263:N...()           | 1.43751:     | 1.87751:     | 1.74599:     | 80:   | 81:   | 83:   | 0.0003       |
| 9:    | 36:(...C...()        | 0.56405:     | 0.74863:     | 0.68467:     | 78:   | 79:   | 69:   | 0.0005       |
| 10:   | 152:C...()=...       | 1.12059:     | 0.80791:     | 0.37896:     | 75:   | 71:   | 76:   | 0.0002       |
| 1:    | 58:1.....            | -0.12827:    | -1.24534:    | -1.25043:    | 103:  | 101:  | 101:  | 0.0000       |
| 2:    | 484:c...()           | -1.06588:    | -0.99722:    | -0.37308:    | 101:  | 99:   | 99:   | 0.0000       |
| 3:    | 545:c...c.....       | -0.81382:    | -0.62617:    | -1.12338:    | 101:  | 99:   | 98:   | 0.0001       |
| 4:    | 316:O...()           | -0.68475:    | -0.43853:    | -0.93696:    | 88:   | 83:   | 86:   | 0.0000       |
| 5:    | 188:C...C...()       | -1.30990:    | -0.87686:    | -1.37116:    | 83:   | 79:   | 85:   | 0.0002       |
| 6:    | 371:[.....           | -0.50173:    | -0.56696:    | -0.12607:    | 79:   | 75:   | 69:   | 0.0006       |
| 7:    | 114:=...()           | -1.00032:    | -0.99808:    | -0.31705:    | 76:   | 76:   | 80:   | 0.0003       |
| 8:    | 552:c...c...c...     | -0.55938:    | -1.00385:    | -0.80764:    | 75:   | 78:   | 77:   | 0.0002       |
| 9:    | 153:C...()C...       | -0.18820:    | -0.50241:    | -0.50102:    | 58:   | 60:   | 57:   | 0.0000       |
| 10:   | 82:2...c...()        | -0.37300:    | -0.74658:    | -2.00191:    | 56:   | 56:   | 59:   | 0.0004       |