

Dear Colleague,

This letter contains brief description how you can use for your task, an updated version of the CORAL software. The number of modifications are not large one. However, these can be attractive for a user.

It is impossible to answer all possible questions therefore, we provide you email for contacts, if you will have questions:

andrey.toropov@marionegri.it

alla.toropova@marionegri.it



Sincerely,
Alla P. Toropova
Andrey A. Toropov

List of additional options:

- A. Builder of a split into the training (+), invisible training (-), calibration (#) and validation (*) sets.
- B. The checking up SMILES duplicates and / or CAS duplicates, i.e. detecting lines where SMILES or CAS number (or other ID) are identical.
- C. HARD: Super-attribute of SMILES
- D. Depth of interpretation
- E. File CheckUp.txt

Important:

Please read ReadMe.pdf before experiments with the CORALSEA-2016

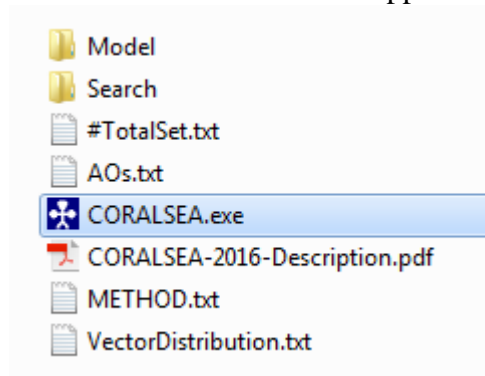
Builder of a split into the training (+), invisible training (-), calibration (#) and validation (*) sets.

In order to use this option you should prepare file #TotalSet.txt. It can be the following (data from *Chemometrics and Intelligent Laboratory Systems* 109 (2011) 94–100):

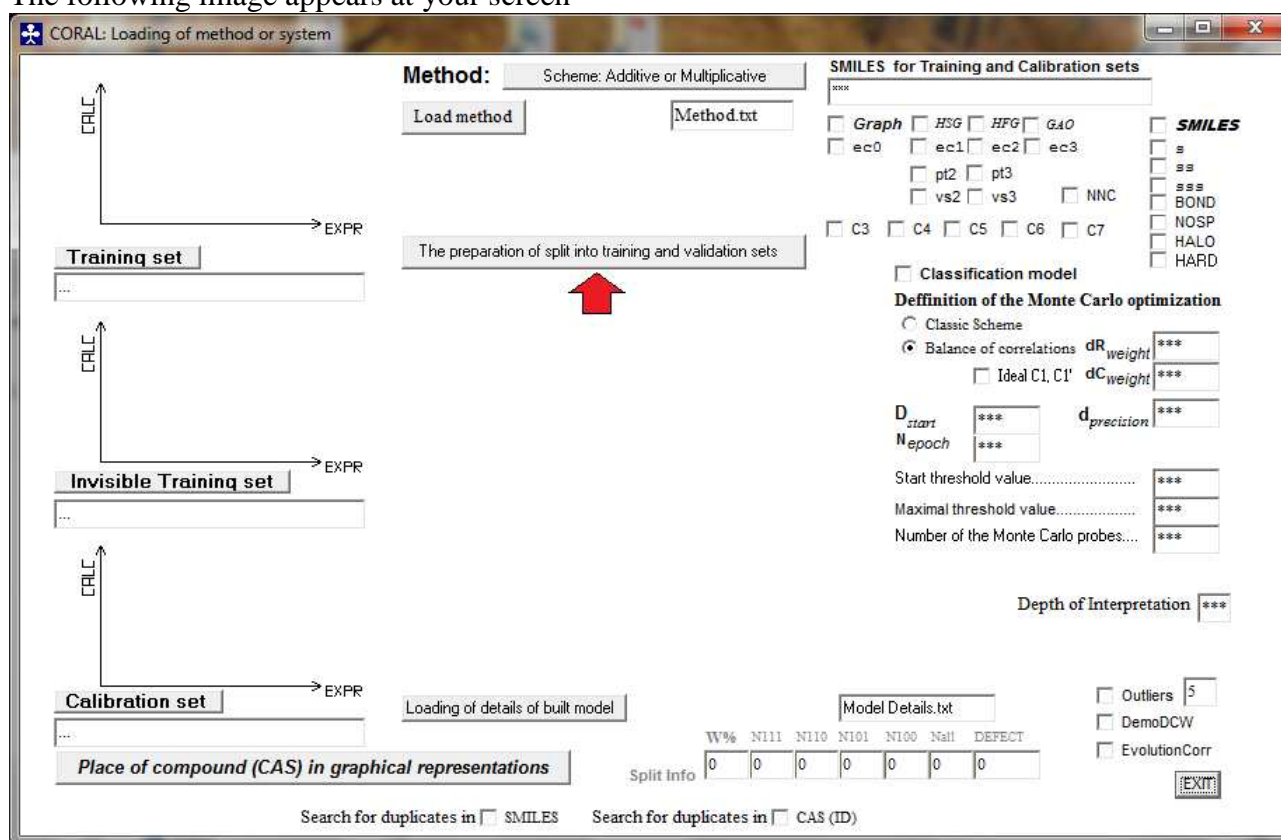
6638-60-4 BrC1ccc2c3ccc(N)cc3Cc2c1 2.62
120-71-8 Nc1cc(C)ccc1OC -2.05
611-34-7 Nc1cccc2ncccc12 -2.00
156-43-4 CCOc1ccc(N)cc1 -2.30
134-32-7 Nc2cccc1cccc12 -0.60
7083-63-8 Nc2cccc3Cc1cccc1c23 1.13
613-13-8 Nc1ccc2cc3cccc3cc2c1 2.62
13177-27-0 Nc4cccc1c4c2cccc3cccc1c23 2.88
578-66-5 Nc1cccc2ccnc12 -1.14
28124-29-0 Nc1ccc2nc3c(nc2c1)cccc3N 0.75
91-59-8 Nc1ccc2cccc2c1 -0.67
17075-03-5 Nc4cc2cccc1ccc3cccc4c3c12 3.16
31835-64-0 Nc1cc(ccc1)c2cc(ccc2)[N+](O-)=O -0.55
137-17-7 Cc1cc(C)c(N)cc1C -1.32
6344-66-7 Nc1cc2c3cccc3Cc2cc1 0.89
91-94-1 Nc1ccc(cc1Cl)c2ccc(N)c(Cl)c2 0.81
95-68-1 Cc1ccc(N)c(C)c1 -2.22
525-64-4 Nc1ccc2c3ccc(N)cc3Cc2c1 0.48
2693-46-1 Nc4ccc2c1cccc1c3cccc4c23 3.31
153-78-6 Nc1ccc2c3cccc3Cc2c1 1.93
6272-52-2 Nc2cccc2c1ccc(cc1)[N+](O-)=O -0.62
92-67-1 Nc1ccc(cc1)c2cccc2 -0.14
16452-01-0 Nc1ccc(C)cc1OC -1.96
4539-51-9 Nc1ccc2c3cccc3Nc2c1 0.60
121-88-0 Nc1ccc(cc1O)[N+](O-)=O -2.52
1454-80-4 Nc2cccc2c1cccc1N -1.52
1953-38-4 Oc1ccc2c3ccc(N)cc3Cc2c1 0.41
4176-53-8 Nc3cccc2c3ccc1cccc12 2.38
95-78-3 Cc1cc(N)c(C)cc1 -2.40
1140-28-9 Nc1ccc(cc1)c2cccc2[N+](O-)=O -0.92
95-84-1 Cc1ccc(O)c(N)c1 -2.10
2876-23-5 Nc1ccc2nc3cccc3nc2c1 0.55
139-65-1 Nc1ccc(cc1)Sc2ccc(N)cc2 0.31
97-02-9 O=[N+](O-)c1cc(ccc1N)[N+](O-)=O -2.00
00-00-01 CC(C)c1ccc(N)cc1N -3.00
367-25-9 Fc1ccc(N)c(F)c1 -2.70
101-77-9 Nc2ccc(Cc1ccc(N)cc1)cc2 -1.60
119-93-7 Cc1cc(ccc1N)c2ccc(N)c(C)c2 0.01
13177-26-9 Nc2cc4cccc3c1cccc1c(c2)c34 3.23
34862-87-8 Nc2cccc2c1cc(ccc1)[N+](O-)=O -0.89
13177-25-8 Nc4ccc3cccc2c1cccc1c4c23 3.35
621-95-4 Nc2ccc(CCc1ccc(N)cc1)cc2 -2.15
106-47-8 Nc1ccc(Cl)cc1 -2.52
3366-65-2 Nc1cc2ccc3cccc3c2cc1 2.46
371-40-4 Fc1ccc(N)cc1 -3.32

947-73-9 Nc2cc3ccccc3c1ccccc12 2.98
 2050-89-7 Nc1cc(ccc1)c2cc(N)ccc2 -1.30
 1732-23-6 Nc1cc2ccc3cccc4ccc(c1)c2c34 3.50
 609-20-1 Clc1cc(N)cc(Cl)c1N -0.69
 6957-50-2 CC(=O)Nc1ccc2c(c1)Cc3cc(N)ccc23 1.18
 7704-40-7 Nc1cc2nc3cc(N)ccc3nc2cc1 1.12
 580-15-4 Nc1ccc2ncccc2c1 -2.67
 102-50-1 Nc1ccc(OC)cc1C -3.00
 96187-18-7 Nc1cc(ccc1)c2ccccc2[N+](=[O-])=O -1.30
 492-17-1 Nc2ccccc2c1ccc(N)cc1 -0.92
 16582-03-9 Nc2cccc1nc3c(nc12)cccc3N 0.20
 722-27-0 Nc2ccc(SSc1ccc(N)cc1)cc2 -1.03
 1817-73-8 Nc1c(cc(cc1Br)[N+](=[O-])=O)[N+](=[O-])=O -0.54
 00-00-02 Nc1cc(N)ccc1CCCC -2.70
 101-80-4 Nc1ccc(cc1)Oc2ccc(N)cc2 -1.14
 90-41-5 Nc2ccccc2c1ccccc1 -1.49
 102877-14-5 Nc2cccc1nc3cccc(N)c3nc12 0.04
 6344-63-4 Nc2cccc1c3ccccc3Cc12 0.43
 5869-25-0 Nc3ccc4c2cccc1cccc(c12)c4c3 3.80
 95-51-2 Nc1ccccc1Cl -3.00
 98-16-8 FC(F)(F)c1cc(N)ccc1 -0.80
 606-57-5 [O-][N+](=O)c1c2ccccc2ccc1N -1.17
 53059-29-3 Nc1cc(ccc1)c2ccc(cc2)[N+](=[O-])=O 0.69
 106-40-1 Nc1ccc(Br)cc1 -2.70
 95-85-2 Nc1cc(Cl)ccc1O -3.00
 119-90-4 Nc1ccc(cc1OC)c2ccc(N)c(OC)c2 0.15
 6373-50-8 Nc1ccc(cc1)C2CCCCC2 -1.24
 139-59-3 Nc2ccc(Oc1ccccc1)cc2 0.38
 19900-65-3 CCc1cc(ccc1N)Cc2ccc(N)c(CC)c2 -0.99
 1214-32-0 [O-][N+](=O)c1ccc2c(c1)Cc3cc(N)ccc23 3.00
 92-87-5 Nc1ccc(cc1)c2ccc(N)cc2 -0.39
 776-34-1 [O-][N+](=O)c1ccc(N)c2ccccc12 -1.77
 1141-29-3 Nc1ccc(cc1)c2cc(ccc2)[N+](=[O-])=O 1.02
 1211-40-1 Nc1ccc(cc1)c2ccc(cc2)[N+](=[O-])=O 1.04
 2876-22-4 Nc2cccc1nc3ccccc3nc12 -0.01
 13824-23-2 Fc2cc(Cc1ccc(N)c(F)c1)ccc2N 0.23
 89-63-4 Nc1ccc(Cl)cc1[N+](=[O-])=O -2.22
 580-17-6 Nc1cc2ccccc2nc1 -3.14
 6377-12-4 Nc1cc2c3ccccc3Nc2cc1 -0.48
 95-83-0 Nc1cc(Cl)ccc1N -0.49
 1892-54-2 Nc2ccc3ccc1ccccc1c3c2 3.77
 32316-90-8 Nc1cc(ccc1)c2ccc(N)cc2 0.20
 610-49-1 Nc2cccc1cc3ccccc3cc12 1.18
 18992-86-4 Nc2cccc1c3ccccc3nc12 -1.04
 779-03-3 Nc2c3ccccc3cc1ccccc12 0.87
 18992-64-8 Nc2cccc3Nc1ccccc1c23 -1.42
 2642-98-0 Nc4cc2c(ccc1ccccc12)c3ccccc34 1.83
 1606-67-3 Nc4ccc1ccc2cccc3ccc4c1c23 1.43
 19900-66-4 CC(C)c1cc(ccc1N)Cc2ccc(N)c(c2)C(C)C -1.77
 120209-97-4 Nc1ccc2nc3cc(N)ccc3nc2c1 3.97

Your work folder should be approximately the following. Click CORALSEA.exe



The following image appears at your screen



Click button “The preparation of split...”; Further actions are (i) click “load”; (ii) click “do distribution”; and click “Save files”.

#TrainingSet.txt contains lines started from “+” for the training set; by “-” for the invisible training set; by “#” for the calibration set.

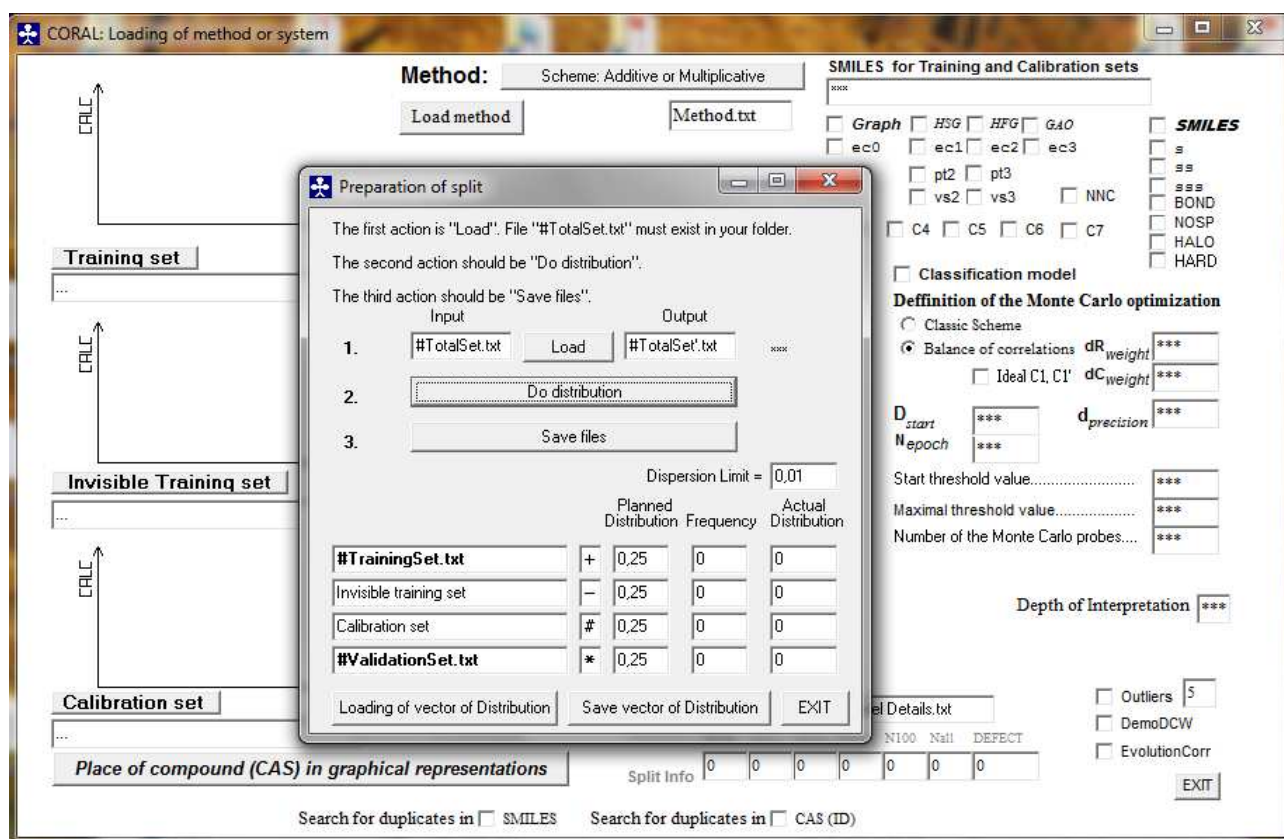
#ValidationSet.txt contains lines started from “*”.

This file utilized to build up a model by traditional scheme, i.e.

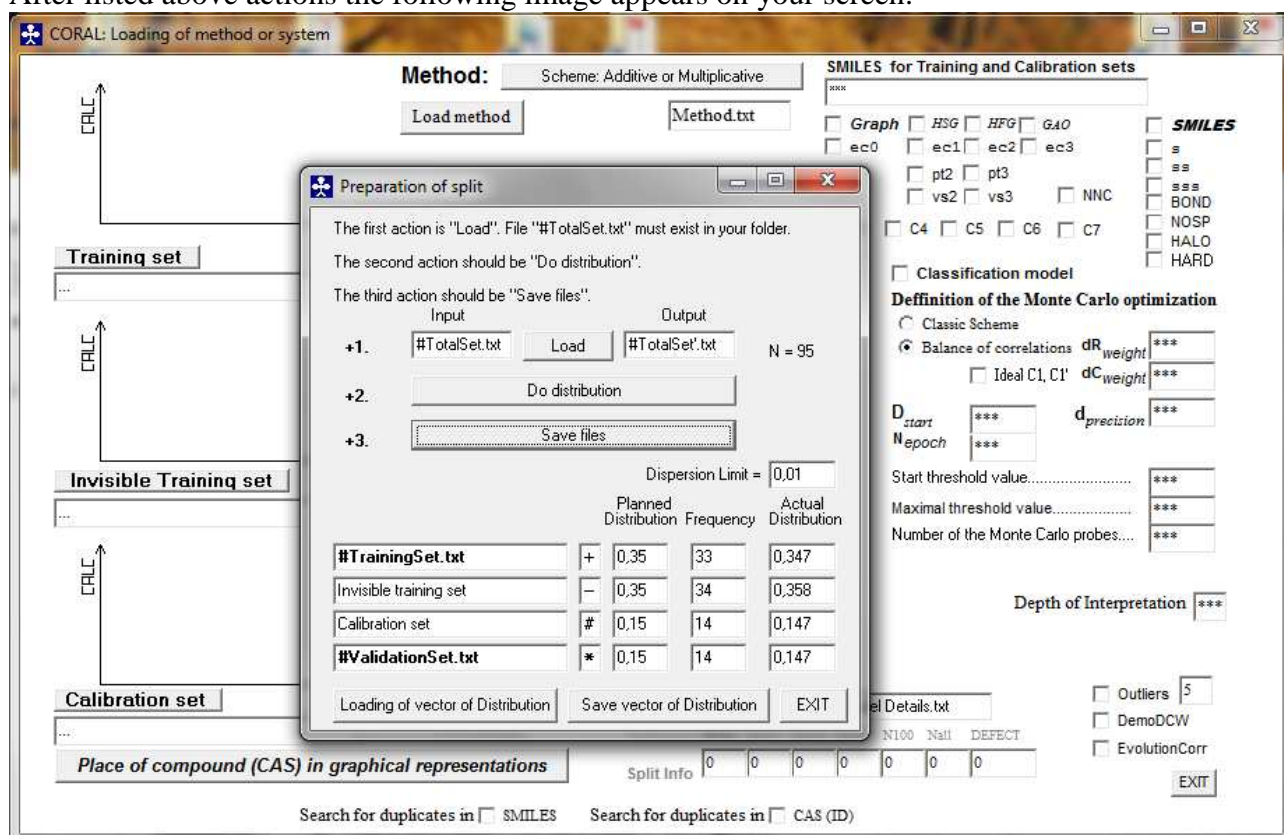
“Training – calibration – validation” System

OR to build up a model by balance of correlations, i.e.

“Training – invisible training - calibration – validation” System



After listed above actions the following image appears on your screen:



Distribution into the training, invisible training, calibration, and validation sets is completed.

Click EXIT for further operations.

You can insert percentage of compounds in the training, invisible training, calibration, and validation sets (Planned distribution). Having inserted these data you can save vector of Distribution, these data will be activated automatically until you will change the vector. You can

change dispersion limit for the percentage, but if you select too small limit the task can become impossible.

If you utilize “Classic Scheme”, then the training and the invisible training will be united in common set. If you utilize balance of correlations the training set and invisible training set have different functions. These functions can be described as the following.

Training set is builder of a model;

Invisible training set is inspector of model during process of building up (blocker for overtraining)

Calibration set is the estimator of the predictive potential (whether take place improvement? If not stop: model is completed).

Validation set is final estimator of predictive potential for external (unknown) compounds.

The checking up SMILES duplicates and / or CAS duplicates, i.e. detecting lines where SMILES or CAS number (or other ID) are identical.

Duplicates are lines where identical SMILES notations or identical ID, such as CAS number or some other identifier. In order to detect identical SMILES, one should activate “Search for duplicates in SMILES”. In order to detect identical CAS number, one should activate “Search for duplicated in CAS (ID)”. These options are placed down of screen:

CORAL: Loading of method or system

Method: Scheme: Additive or Multiplicative

Load method Method.txt

SMILES for Training and Calibration sets

☐ Graph ☐ HSG ☐ HFG ☐ GAO ☐ ec0 ☐ ec1 ☐ ec2 ☐ ec3 ☐ pt2 ☐ pt3 ☐ vs2 ☐ vs3 ☐ NNC ☐ C3 ☐ C4 ☐ C5 ☐ C6 ☐ C7

☐ SMILES ☐ s ☐ ss ☐ sss ☐ BOND ☐ NOSP ☐ HALO ☐ HARD

☐ Classification model

Definition of the Monte Carlo optimization

☐ Classic Scheme

☒ Balance of correlations dR_weight dC_weight

☐ Ideal C1, C1'

D_start d_precision

N_epoch

Start threshold value.....

Maximal threshold value.....

Number of the Monte Carlo probes....

Depth of Interpretation

☐ Outliers

☐ DemoDCW

☐ EvolutionCorr

Model Details.txt

W% N111 N110 N101 N100 Nall DEFECT

0 0 0 0 0 0

Split Info

Place of compound (CAS) in graphical representations

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

EXIT

HARD: Super-attribute of SMILES

HARD is a global SMILES attribute. This is a sequence of twelve symbols: “0” or “1”.

The 0 means the absence of an local attribute (chemical element or kind of chemical bond),
The 1 means the presence of an local attribute.

Below, three examples of the HARD are represented.

SMILES		BOND			NOSP				HALO			
		=	#	@	N	O	S	P	F	Cl	Br	I
c1cccc1/C(c2cccn2)=N/Nc(n3)sc(c34)cccc4	\$	1	0	0	1	0	0	0	0	0	0	0
COCCCC\C(c1ccc(cc1)C(F)(F)F)=N/OCCN	\$	1	0	0	1	1	0	0	1	0	0	0
C[C@H](CN(C)C)CN1c2ccccc2Sc2ccc(cc12)C#N	\$	0	1	1	1	0	1	0	0	0	0	0

In fact, HARD is an assembling of BOND, NOSP, and HALO into united descriptor extracted from SMILES. There are probabilities of low prevalence of majority of these descriptors, but rare version can be blocked and removed from building up a model (by means of selection of corresponding threshold).

Depth of interpretation

In main “ReadMe.pdf”, the s-Files (Section 3.7) are described. In the CORALSEA-2016 in addition to s-files, p-files are provided. These files contains the number of promoters for increase and the same number of promoters of decrease of examined endpoint. The number is called “Depth of interpretation”. This option should be selected according to a given task. For instance, the depth = 77 for a task where total number of attributes involved in building up a model is 50 is an absolute nonsense.

The screenshot shows the CORAL software interface with the title bar "CORAL: select Phase 1, Phase 2, or change and save method". The interface is divided into several sections:

- Method:** Includes buttons for "Adding", "Load method", "Save method", and "Method.txt". Below these are two phases: "Phase 1: Search for preferable model (T*,N*)" and "Phase 2: Building up preferable model (T*,N*)".
- SMILES for Training and Calibration sets:** A section for "#TrainingSet.txt" with checkboxes for various SMILES features like Graph, H3G, HFG, G4O, ec0, ec1, ec2, ec3, pt2, pt3, vs2, vs3, NNC, C3, C4, C5, C6, C7, and a list of SMILES types (s, ss, sss, BOND, NOSP, HALO, HARD).
- Definition of the Monte Carlo optimization:** Includes options for "Classic Scheme" and "Balance of correlations" with a "dR_weight" input field. It also has fields for "D_start", "N_epoch", "d_precision", "Start threshold value", "Maximal threshold value", and "Number of the Monte Carlo probes".
- Depth of Interpretation:** A red box highlights the input field for "Depth of Interpretation" which is set to 10.
- Calibration set:** Includes a "Loading of details of built model" button and a "Model Details.txt" button.
- Split Info:** A table showing the distribution of compounds across different sets.
- Search for duplicates:** Checkboxes for "SMILES" and "CAS (ID)".

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

File CheckUp.txt

The file CheckUp.txt contains copy of lines from #TrainingSet.txt. If there are some misprints in these lines the program can work wrong or even halt. Sometimes, the analysis of CheckUp.txt helps to correct the #TrainingSet.txt.