http://www.insilico.eu/coral/index.html

Correlations
O
R
And
Logic:

SMILES,
Electrons,
Atoms…

Version : January 19, 2016
for Microsoft Windows

Reference Manual

Department of Environmental Health Science
Laboratory of Environmental Chemistry and toxicology
***Head of Laboratory**: Emilio Benfenati, PhD*

Istituto di Ricerche Farmacologiche Mario Negri,
20156, Via La Masa 19, Milano, Italy

Authors:
Andrey A. Toropov, PhD
(development of algorithms for the software CORAL; development a software for the QSPR/QSAR analysis where the molecular structure should be represented by SMILES, and for investigating nanomaterials represented by quasi-SMILES)
Alla P. Toropova, PhD
(development of CORAL web-architecture, the permanent updating of the contain of the web-site CORAL in accordance with new options which are related to the Monte Carlo optimization as well as in accordance with new publications where the CORAL is used as a tool for the QSPR/QSAR analysis)

Contact Us:
andrey.toropov@marionegri.it
alla.toropova@marionegri.it
emilio.benfenati@marionegri.it

Table of Contents

Preface

CORALSEA is software for building up quantitative structure – property / activity relationships (QSPR/QSAR). The building up of QSPR/QSAR is based on the Monte Carlo technique. Molecular structure of each substance involved in the training or test sets should be represented by SMILES.

There are some updates for the software. We hope they can be useful.

Criticism, suggestions, and remarks related to praxis of using CORALSEA will be accepted with gratitude.

We shall do our best in order to answer any questions related to the CORALSEA software.

Authors

January 30, 2015

 *We would like to express our gratitude to experts in the field of the QSPR/QSAR analyses who in different time have helped us in the organization of the software: Prof. J. Leszczynski and Dr. B.F. Rasulev (Interdisciplinary Nanotoxicity Center, Jackson State University, USA), Prof. E.A. Castro and Dr. P.R. Duchowicz (Instituto de Investigaciones Fisicoquímicas Teóricas y Aplicadas, La Plata, Argentina), Prof. K. Roy (Jadavpur University, India), Dr. K. Nesmerak (Charles University in Prague, Czech Republic), Dr. I. Raska Jr (Charles University in Prague, Czech Republic), Dr. J.B. Veselinovic and Dr. A.M. Veselinovic (University of Nis, Serbia), Dr. Xiao-Yun Zhang (Lanzhou University, Republic of China), Dr. V.H. Masand (Department of Chemistry, Vidya Bharati College, Amravati, Maharashtra, India), Dr. A. Worachartcheewan (Mahidol University, Bangkok, Thailand), Dr. K. Ramanathan (VIT University, Vellore, Tamil Nadu, India), Dr. P.G.R. Achary (Siksha 'O' Anusandhan University, Bhubaneswar, India). Also authors express their gratitude to Martyanov S.E. (Teleca, Nizhny Novgorod, Russia) for the developing of algorithm for translation of SMILES into molecular graph.*

How one can use the CORALSEA?

Five steps should be done in order to obtain a QSPR/QSAR model by means of CORALSEA, these steps are the following:

Step 1. Preparation of input data
Step 2. Definition of the method
Step 3. Searching for the best threshold (T*) and the best number of epochs (N*)
Step 4. Checking of the model that is calculated with T* and N*
Step 5. Checking of the approach with a few random splits

We recommend to prepare a copy of MyCORALSEA folder for your experiments.

Step1. Preparation of input data

In order to use the software you must prepare text SMILES-file (i.e. set of strings, each string contains four components) organized as the following:

1.Type of set i.e. '+' sub-training set; '-' calibration set; and '#' test set;
2.Identifier i.e. the number, or CAS number;
3.SMILES;
3.Endpoint value.

Example:

```
#276 ClCC(Cl)Cl 3.09
+31 CCC(Cl)Cl 3.57
+282 ClCC(Cl)CCl 3.72
+297 Clc1ccc(c(c1)Cl)Cl 4.16
#223 [O-][N+](=O)c1cccc(c1Cl)Cl 4.62
#281 Clc1ccccc1Cl 4.81
#287 ClCCCl 2.29
-275 C[C@@H](Cl)CCl 3.34
#288 OCCO 0.48
#177 [O-][N+](=O)c1cc(cc(c1)Cl)Cl 4.46
#300 Clc1cccc(c1)Cl 4.18
+299 ClCCCCl 2.61
-77 [O-][N+](=O)c1cccc(c1)[N+]([O-])=O 3.59
#48 [O-][N+](=O)c1cc(ccc1Cl)Cl 4.26
#228 S=C=Nc1ccc(cc1)N=C=S 6.4
#70 [O-][N+](=O)c1ccc(c(c1)[N+]([O-])=O)Cl 5.4
+44 [O-][N+](=O)c1ccccc1Cl 3.64
#293 CCCCCCCO 3.22
-171 CCCCN=C=S 5.43
-43 Cc1ccccc1[N+]([O-])=O 4.14
#75 Cc1cccc(c1)[N+]([O-])=O 4.04
#219 CCNC(=S)Nc1ccccc1 3.35
+99 CCCS 6.1
```

```
#270 CCCO 0.93
-120 CNC(=O)Oc1ccccc1OC(C)C 4.91
-45 CC[C@@H](C)c1cc(cc(c1O)[N+]([O-])=O)[N+]([O-])=O 6
#294 CCOCCOCCO 1.53
+253 Clc1ccc(c(c1Cl)Cl)c1ccc(c(c1Cl)Cl)Cl 8.78
#250 Clc1ccc(c(c1)c1cc(ccc1Cl)Cl)Cl 6.99
+118 ClCCOCCCl 2.78
+117 OCCNCCO 2.93
-238 Clc1c(c(c(c(c1Cl)Cl)c1ccccc1)Cl)Cl 7.61
#184 Nc1cc(c(c(c1Cl)Cl)Cl)Cl 5.56
```

Component2 is ID for given substance. It can be number 1, 234, 985; It can be CAS number, e.g. 75-07-0, 712-68-5, etc. It can be any other identifier which has no interword space. The number of characters in the ID should be less than 30.

Component3 is simplified molecular input line entry system (SMILES) for given substance;
Component4 is numerical value of endpoint for which QSPR/QSAR model should be built up.

Components 2, 3, and 4 must be separated by ONE (not two or more) interword space , i.e.
Component1Component2[interword space]Component3[interword space]Component4.
Component1 must be connected directly to component2 (without interword).

Having prepared this file you must save it in Folder ‘CORALSEA’ (or better ‘MyCORALSEA’).
The name of the file can be ‘Split.txt’, ‘Split1.txt’, ‘Toxicity.txt’, ‘ld50.txt’, ‘BCF-1.txt’, etc.
The program can work properly if
1. Each string prepared as shown in the above example;
2. No empty or invalid string takes place in the list;
3. The length of SMILES is less than 500;
4. The number of strings is less than 50000.

The file should be prepared by a text editor, e.g. BlockNote:
Word or Excel files cannot be used for CORALSEA.

Examples of situations when the program will be work wrong:

Example 1 // third string is empty

```
#276 ClCC(Cl)Cl 3.09
+31 CCC(Cl)Cl 3.57

+282 ClCC(Cl)CCl 3.72
+297 Clc1ccc(c(c1)Cl)Cl 4.16
...
```

Example 2 // invalid second string: endpoint value is absent

```
#276 ClCC(Cl)Cl 3.09
+31 CCC(Cl)Cl
+282 ClCC(Cl)CCl 3.72
+297 Clc1ccc(c(c1)Cl)Cl 4.16
#223 [O-][N+](=O)c1cccc(c1Cl)Cl 4.62
#281 Clc1ccccc1Cl 4.81
#287 ClCCCl 2.29
```

...
Example 3 // invalid 4-th string: component1 and component2 are absent
```
#276 ClCC(Cl)Cl 3.09
+31 CCC(Cl)Cl 3.57
+282 ClCC(Cl)CCl 3.72
[O-][N+](=O)c1cccc(c1Cl)Cl 4.62
#281 Clc1ccccc1Cl 4.81
```
...
Example 4 // invalid 5-th string: interword between component1 and component2
```
#276 ClCC(Cl)Cl 3.09
+31 CCC(Cl)Cl 3.57
+282 ClCC(Cl)CCl 3.72
+297 Clc1ccc(c(c1)Cl)Cl 4.16
# 223[O-][N+](=O)c1cccc(c1Cl)Cl 4.62
#281 Clc1ccccc1Cl 4.81
#287 ClCCCl 2.29
```

Having correct SMILES-file e.g. ‘MySPLIT1.txt’ in folder ‘MyCORALSEA’ you can start step 2.

Step 2. Definition of the method

1. Run CORALSEA.exe.

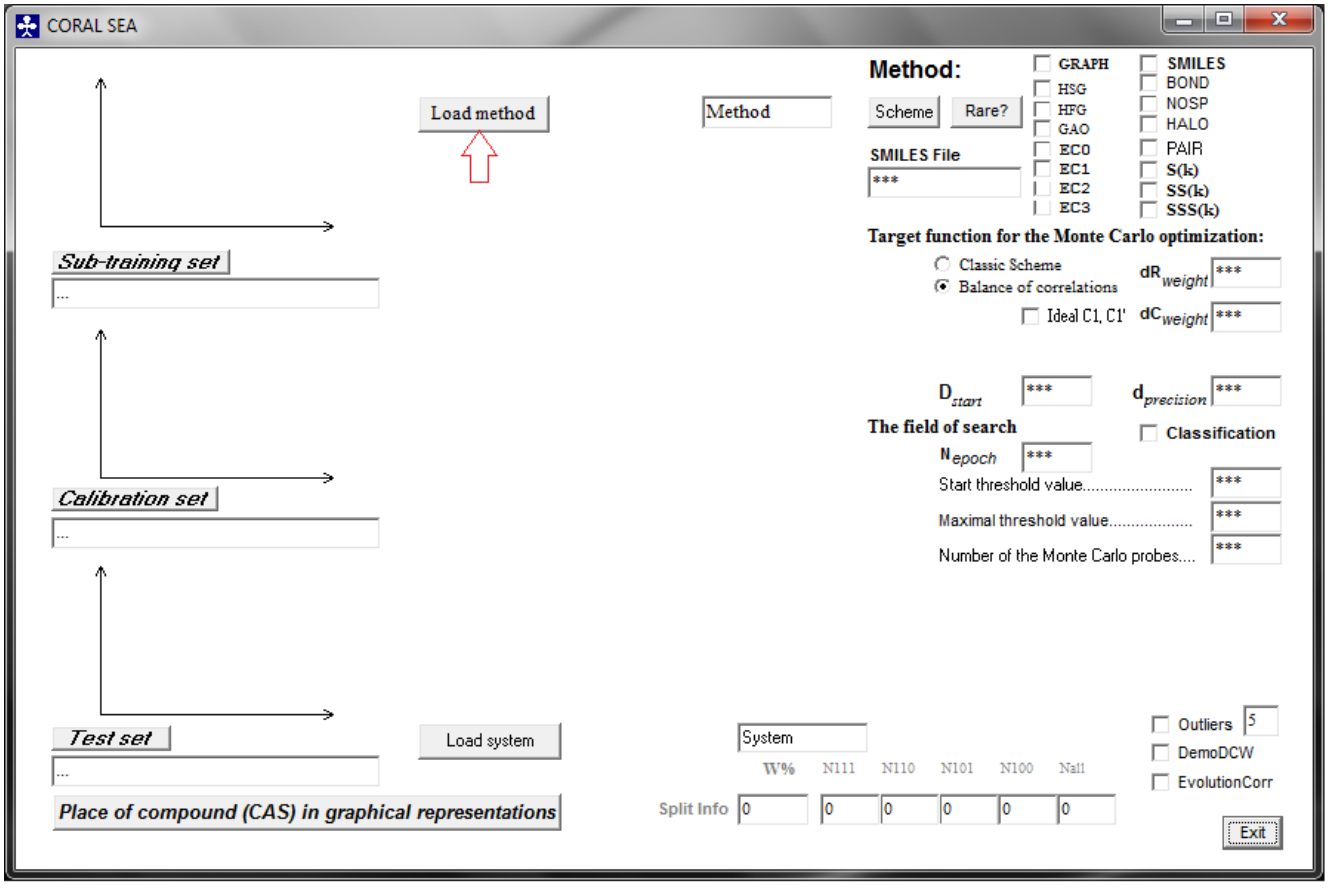
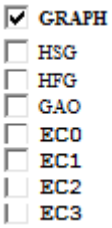


FIGURE 1

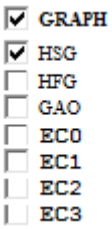
2. Click “Load method” button.

When the method is downloaded, you can correct options according to your task. You can define your method by means of activation / deactivation of available checkboxes. You can also define work parameters (D_{start}, $D_{precision}$, N_{epoch}, dR_{weight}, dC_{weight}).

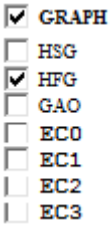


The meaning of options which are related to GRAPH are the following. You can involve the molecular graph in the modeling process by means of selecting box “GRAPH”.
It is necessary to define the kind of the molecular graph. It can be hydrogen suppressed graph (HSG); hydrogen filled graph (HFG); and graph of atomic orbitals (GAO)

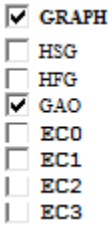
The selection of the kind of the molecular graph can be done as the following:



HSG is selected



HFG is selected



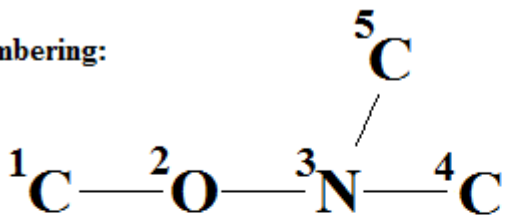
GAO is selected

Also it is necessary to define invariants of the graph which you would like to involve in the modeling process. There are two classes of graph invariants which are available in the CORALSEA: vertices and Morgan vertices' degrees. In the case of HSG and HFG, vertices are representation of the chemical elements, such as carbon, nitrogen, oxygen, etc. In the case of GAO, vertices are representation of electronic structure, i.e. AOs such as $1s^1$, $2s^2$, $2p^5$, $3d^{10}$, etc.

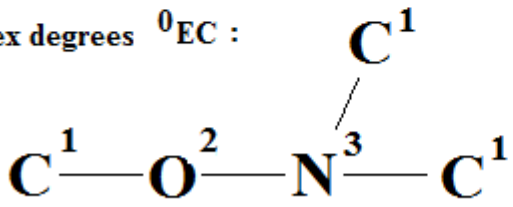
2.1. Example of molecular graphs for Trimethylhydroxylamine (CAS 5669-39-6)

HSG

Numbering:



Vertex degrees ^0EC :

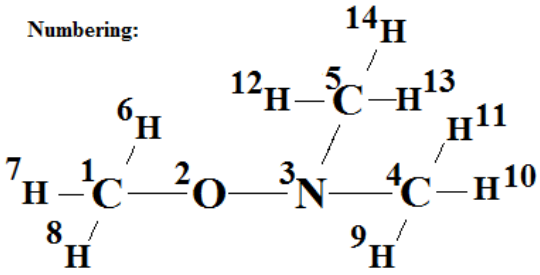


The adjacency matrix

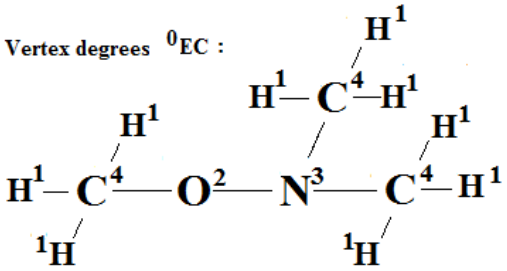
	1	2	3	4	5	^0EC
1	0	1	0	0	0	1
2	1	0	1	0	0	2
3	0	1	0	1	1	3
4	0	0	1	0	0	1
5	0	0	1	0	0	1

HFG

Numbering:



Vertex degrees ^{0}EC :

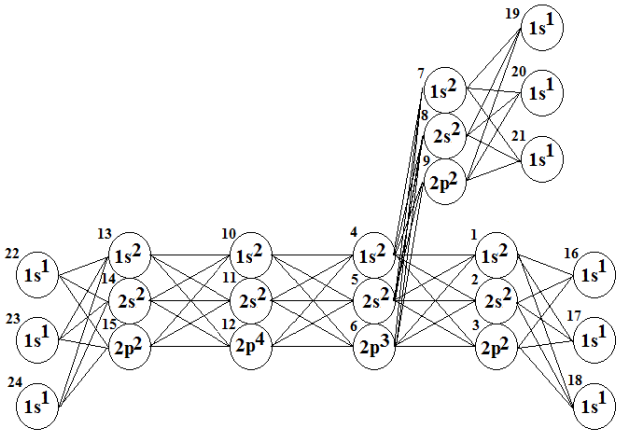


The adjacency matrix

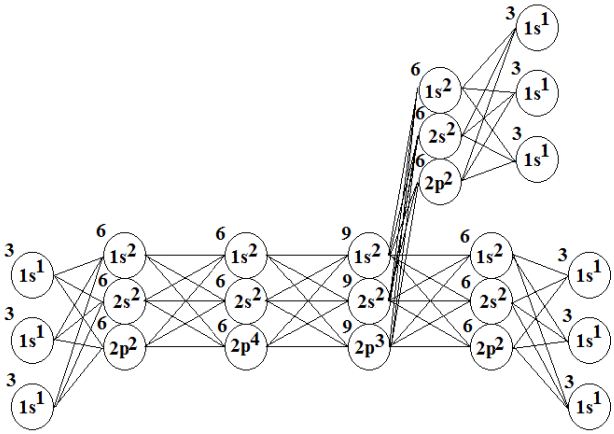
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	^{0}EC
1	0	1	0	0	0	1	1	1	0	0	0	0	0	0	4
2	1	0	1	0	0	0	0	0	0	0	0	0	0	0	2
3	0	1	0	1	1	0	0	0	0	0	0	0	0	0	3
4	0	0	1	0	0	0	0	0	1	1	1	0	0	0	4
5	0	0	1	0	0	0	0	0	0	0	0	1	1	1	4
6	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1
7	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1
8	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1
9	0	0	0	1	0	0	0	0	0	0	0	0	0	0	1
10	0	0	0	1	0	0	0	0	0	0	0	0	0	0	1
11	0	0	0	1	0	0	0	0	0	0	0	0	0	0	1
12	0	0	0	0	1	0	0	0	0	0	0	0	0	0	1
13	0	0	0	0	1	0	0	0	0	0	0	0	0	0	1
14	0	0	0	0	1	0	0	0	0	0	0	0	0	0	1

GAO

Numbering:



Vertex degree ^{0}EC:



2.2. The adjacency matrix for graph of atomic orbitals for Trimethylhydroxylamine (CAS 5669-39-6)

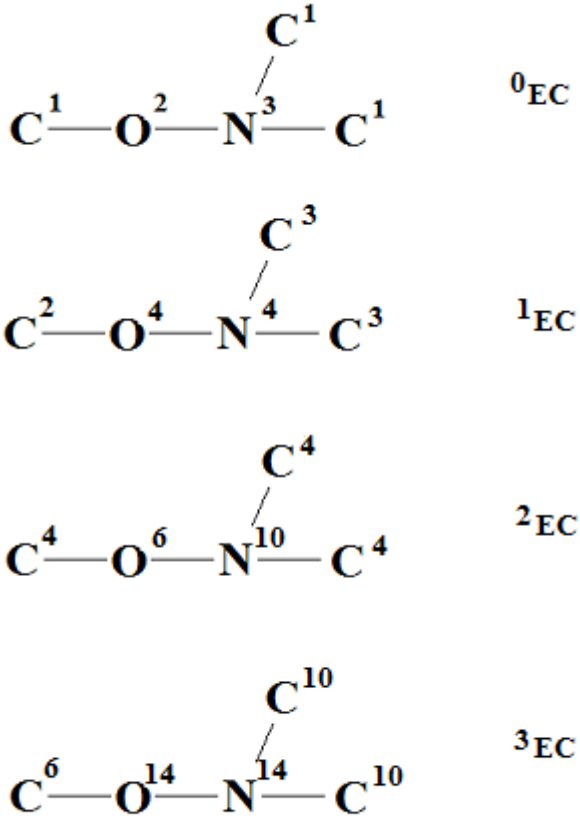
	$1s^2$	$2s^2$	$2p^2$	$1s^2$	$2s^2$	$2p^3$	$1s^2$	$2s^2$	$2p^2$	$1s^2$	$2s^2$	$2p^4$	$1s^2$	$2s^2$	$2p^2$	$1s^1$	$1s^1$	$1s^1$	$1s^1$	$1s^1$	$1s^1$	$1s^1$	$1s^1$	$1s^1$	
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	^{0}EC
$1s^2$ 1	0	0	0	1	1	1	0	0	0	0	0	0	0	0	0	1	1	1	0	0	0	0	0	0	6
$2s^2$ 2	0	0	0	1	1	1	0	0	0	0	0	0	0	0	0	1	1	1	0	0	0	0	0	0	6
$2p^2$ 3	0	0	0	1	1	1	0	0	0	0	0	0	0	0	0	1	1	1	0	0	0	0	0	0	6
$1s^2$ 4	1	1	1	0	0	0	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	9
$2s^2$ 5	1	1	1	0	0	0	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	9
$2p^3$ 6	1	1	1	0	0	0	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	9
$1s^2$ 7	0	0	0	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	0	0	0	6
$2s^2$ 8	0	0	0	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	0	0	0	6
$2p^2$ 9	0	0	0	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	0	0	0	6
$1s^2$ 10	0	0	0	1	1	1	0	0	0	0	0	0	1	1	1	0	0	0	0	0	0	0	0	0	6
$2s^2$ 11	0	0	0	1	1	1	0	0	0	0	0	0	1	1	1	0	0	0	0	0	0	0	0	0	6
$2p^4$ 12	0	0	0	1	1	1	0	0	0	0	0	0	1	1	1	0	0	0	0	0	0	0	0	0	6
$1s^2$ 13	0	0	0	0	0	0	0	0	0	1	1	1	0	0	0	0	0	0	0	0	0	1	1	1	6
$2s^2$ 14	0	0	0	0	0	0	0	0	0	1	1	1	0	0	0	0	0	0	0	0	0	1	1	1	6
$2p^2$ 15	0	0	0	0	0	0	0	0	0	1	1	1	0	0	0	0	0	0	0	0	0	1	1	1	6
$1s^1$ 16	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3
$1s^1$ 17	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3
$1s^1$ 18	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3
$1s^1$ 19	0	0	0	0	0	0	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3
$1s^1$ 20	0	0	0	0	0	0	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3
$1s^1$ 21	0	0	0	0	0	0	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3
$1s^1$ 22	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	0	0	0	0	0	0	0	0	0	3
$1s^1$ 23	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	0	0	0	0	0	0	0	0	0	3
$1s^1$ 24	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	0	0	0	0	0	0	0	0	0	3

Morgan extended connectivity of (k+1)-th order (^kEC_i) for each vertex in a molecular graph is calculated with the extended connectivity of k-th order by equation

$$^{k+1}EC_i = \sum_{a[i,j]\neq 0} {}^{k}EC_j$$

where a[i,j] is element of the adjacency matrix.

For HSG of Trimethylhydroxylamine calculation of the ^{1}EC, 2EC, and 3EC is the following:

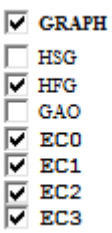


^{0}EC is the number of neighbors for i-th vertex in molecular graph.

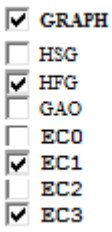
Optimal graph-based descriptor is calculated as the following

$$^{Graph}DCW(Threshold, N_{epoch}) = \sum CW(A_k) + \alpha\sum CW(^0EC_k) + \beta\sum CW(^1EC_k) + \gamma\sum CW(^2EC_k) + \delta\sum CW(^3EC_k) \quad (1)$$

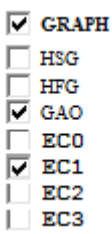
One can use all or some selected extended connectivity values. For example:



^{0}EC, ^{1}EC, ^{2}EC, and ^{3}EC in HSG are involved in the modeling process, i.e. α=1; β=1;γ=1; and δ=1.



^{1}EC and ^{3}EC in HFG are involved in the modeling process, i.e. α=0; β=1;γ=0; and δ=1.



^{1}EC in GAO are involved in the modeling process, i.e. α=0; β=1;γ=0; and δ=0.

IMPORTANT: SMILES are translating into HSG. If HFG is selected, then the HSG is modifying for four chemical elements: Carbon, Nitrogen, Oxygen, and Sulphur. Vertices for listed chemical elements are obtaining addition hydrogen vertices. Other chemical elements are not modifying! In other words, if work set of compounds contains for example Si, it will be better to use HSG, not HFG.

2.3. Example of SMILES attributes

Optimal SMILES-based descriptor is calculated as the following

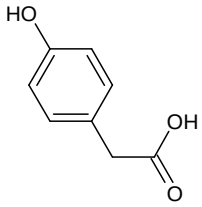
$$^{SMILES}DCW(Threshold, N_{epoch}) = a\sum CW(S_k) + \beta\sum CW(SS_k) + \gamma\sum CW(SSS_k) + \delta \cdot CW(PAIR) + x \cdot CW(NOSP) + y \cdot CW(HALO) + z \cdot CW(BOND) \quad (2)$$

If SMILES=ABCDE, then examples of S_k, SS_k, and SSS_k can be represented as

ABCDE → A + B + C + D + E (S_k)
ABCDE → AB + BC + CD + DE (SS_k)
ABCDE → ABC+BCD+CDE (SSS_k)

More realistic example: if SMILES = Clc1ccccc1
then S_k = (Cl, c,1,c,c,c,c,c,1); SS_k = (Clc,c1,cc,cc,cc,cc,cc,c1); SSS_k = (Clc1,c1c,ccc,ccc,ccc,ccc,cc1).

Finally, an example of the preparation of a list of the attributes S_k,SS_k, SSS_k in CORALSEA format
`SMILES="c1(CC(=O)O)ccc(O)cc1"  CAS= 156-38-7`



S_k zone 1 zone 2 zone 3	SS_k zone 1 zone 2 zone 3	SSS_k zone 1 zone 2 zone 3
c..........		
1..........	c...1.....	
(..........	1...(.....	c...1...(...
C..........	C...(.....	C...(...1...
C..........	C...C.....	C...C...(...
(..........	C...(.....	C...C...(...
=..........	=...(.....	C...(...=...
O..........	O...=.....	O...=...(...
(..........	O...(.....	=...O...(...
O..........	O...(.....	O...(...O...
(..........	O...(.....	(...O...(...
c..........	c...(.....	c...(...O...
c..........	c...c.....	c...c...(...
c..........	c...c.....	c...c...c...
(..........	c...(.....	c...c...(...
O..........	O...(.....	c...(...O...
(..........	O...(.....	(...O...(...
c..........	c...(.....	c...(...O...
c..........	c...c.....	c...c...(...
1..........	c...1.....	c...c...1...

It is to be noted that ‘)’ is changed by ‘(‘, because these symbols are indicators of the same phenomenon (branching). The same situation takes place for ‘[‘ and ‘]’.

Often S_k is sole symbol, but there are exceptions: e.g. chemical elements of two symbols (such as Cl, Br, Na, Cu, etc.); @@ (stereo chemical aspects of the molecular structure); %10, %11 (the number of cycles in molecule more than 9, see http://www.daylight.com/dayhtml/doc/theory/theory.smiles.html).

Important: CORAL software cannot translate SMILES which contain "%" (i.e. %10, %11, etc), '.', and '' into graphs.*

S_k, SS_k, and SSS_k are local SMILES attributes, they are representation of molecular fragments.

PAIR, *NOSP*, *HALO*, and *BOND* are global SMILES attributes which are calculating with SMILES.

Atoms' *PAIRs* are denoted as the following:

	Cl	Br	N	O	S	P	B2	B3
F	++++F---Cl==	++++F---Br==	++++F---N===	++++F---O===	++++F---S===	++++F---P===	++++F---B2==	++++F---B3==
Cl		++++Cl--Br==	++++Cl--N===	++++Cl--O===	++++Cl--S===	++++Cl--P===	++++Cl--B2==	++++Cl--B3==
Br			++++Br--N===	++++Br--O===	++++Br--S===	++++Br--P===	++++Br--B2==	++++Br--B3==
N				++++N---O===	++++N---S===	++++N---P===	++++N---B2==	++++N---B3==
O					++++O---S===	++++O---P===	++++O---B2==	++++O---B3==
S						++++S---P===	++++S---B2==	++++S---B3==
P							++++P---B2==	++++P---B3==
B2								++++B2--B3==

In SMILES the B2 and B3 are indicated by '=' and '#', respectively.

The scheme for calculation of the NOSP. This index related to presence/absence of four chemical elements: nitrogen, oxygen, sulphur, and phosphorus

N	O	S	P	Comments
0	0	0	0	Nitrogen, oxygen, sulphur, and phosphorus are absent
0	0	0	1	Only phosphorus takes place in molecule
0	0	1	0	Only sulphur takes place in molecule
0	0	1	1	Molecule contains sulphur and phosphorus
0	1	0	0	Only oxygen takes place in molecule
0	1	0	1	Molecule contains oxygen and phosphorus
0	1	1	0	Molecule contains oxygen and sulphur
0	1	1	1	Molecule contains oxygen, sulphur, and phosphorus
1	0	0	0	Only nitrogen takes place in molecule
1	0	0	1	Molecule contains nitrogen and phosphorus
1	0	1	0	Molecule contains nitrogen and sulphur
1	0	1	1	Molecule contains nitrogen, sulphur, and phosphorus
1	1	0	0	Molecule contains nitrogen and oxygen
1	1	0	1	Molecule contains nitrogen, oxygen and phosphorus
1	1	1	0	Molecule contains nitrogen, oxygen, and sulphur
1	1	1	1	Molecule contains nitrogen, oxygen, sulphur, and phosphorus

The scheme for calculation of the HALO. This index related to presence/absence of three chemical elements: fluorine, chlorine, and bromine.

F	Cl	Br	Comments

0	0	0	Flourine, chlorine and bromine are absent
0	0	1	Only bromine takes place
0	1	0	Only chlorine takes place
0	1	1	Molecule contains chlorine and bromine
1	0	0	Only fluorine takes place
1	0	1	Molecule contains fluorine and bromine
1	1	0	Molecule contains fluorine and chlorine
1	1	1	Molecule contains fluorine, chlorine, and bromine

The scheme for calculation of the BOND. This index related to presence/absence of three categories of chemical bonds: double, triple, and stereo specific.

=	#	@	**Comments**
0	0	0	Double, triple, and stereo specific bonds are absent
0	0	1	Only stereo specific bonds take place
0	1	0	Only triple bonds take place
0	1	1	Triple and stereo specific bonds take place
1	0	0	Only double bonds take place
1	0	1	Double and stereo specific bonds take place
1	1	0	Double and triple bonds take place
1	1	1	Double, triple, and stereo specific bonds take place

One can select SMILES-based descriptor by the manner similar to the case of the graph-based descriptors. For example,

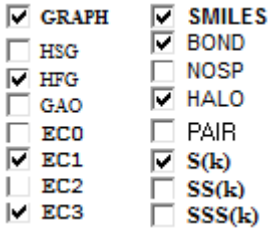
SMILES attributes which are a combination one- and two-elements SMILES attributes are involved in the modeling process, i.e. α=1; β=1; γ=0; δ=0; x=0; y=0; and z=0.

HALO and NOSP indices are involved in modeling process together with one-element attributes, i.e. α=1; β=0; γ=0; δ=0; x=0; y=1; and z=1.

CORALSEA software can be used to build up a hybrid model which is calculated with SMILES-based and GRAPH-based descriptors:

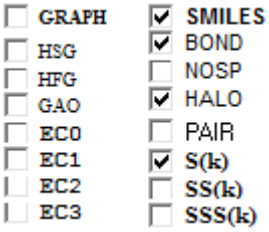
$$^{Hybrid}DCW(Threshold, N_{epoch}) = {}^{SMILES}DCW(Threshold, N_{epoch}) + {}^{Graph}DCW(Threshold, N_{epoch}) \quad (3)$$

For example,

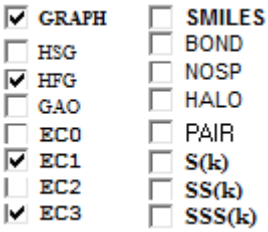


^{1}EC and ^{3}EC in HFG together with HALO, BOND and S_k are involved in the modeling process.

One can use solely SMILES-based descriptors or graph-based descriptors. For example:



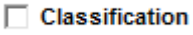
Only SMILES-based descriptors are involved in the modeling process



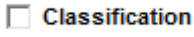
Only Graph-based descriptors are involved in the modeling process

After selection of the options related to SMILES and Graph definitions, one can continue using the CORALSEA software to get a QSPR/QSAR model.

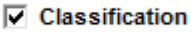
We should comment the following components of FIGURE 1:



Activation of this checkbox leads to preparation of classification model for data of type Yes / No; active / inactive - which are represented by -1 / 1 or 0 / 1. In other words:



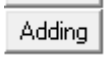
=The building up regression model Y=C0 +C1*DCW



=The building up classification model

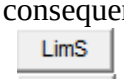
There are two scheme of the calculation with CORAL: the additive scheme (Eq. 1, and Eq. 2) and the multiplicative scheme (Eq. 4 and Eq. 5). By click of this button you can change multiplicative scheme by additive scheme and vice versa.

=The multiplicative scheme

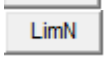


=The additive scheme

This button defines one from two possibilities to detect rare (noise) attributes: the first: the number (LimS) of SMILES in training set which contain the given attribute; the second: the total number (LimN) of attribute in the training set. It is to be noted a SMILES can contains two or more number of the given attribute, consequently, generally speaking LimN ≠ LimS.



=LimS is used as the criterion to detect noise attributes



= LimN is used as the criterion to detect noise attributes

IMPORTANT: in fresh version of the software, automatically the LimS criterion is involved.

$$ {}^{Graph}DCW(Threshold, N_{epoch}) = \prod CW(A_k) \cdot \alpha \prod CW({}^0EC_k) \cdot \beta \prod CW({}^1EC_k) \cdot \gamma \prod CW({}^2EC_k) \cdot \delta \prod CW({}^3EC_k) \quad (4) $$

$$ {}^{SMILES}DCW(Threshold, N_{epoch}) = \cdot \alpha \prod CW(S_k) \cdot \beta \prod CW(SS_k) \cdot \gamma \prod CW(SSS_k) \cdot \delta \cdot CW(PAIR) \cdot x \cdot CW(NOSP) \cdot y \cdot CW(HALO) \cdot z \cdot CW(BOND) \quad (5) $$

In the case of Eq. 4 and Eq. 5, one cannot speak about α=1 or 0, β=1 or 0; but at the level of definition of the DCW(Threshold, Nepoch) the actions are the same as actions which are demonstrated for Eq. 3.

2.4. The Monte Carlo method optimization

The Monte Carlo optimization is some number of epochs of the searching for maximum of a target function. The epoch is a sequence of variation for correlation weight of each molecular attribute (e.g. *Sk, SSk, A_k $EC1_k$,* BOND, HALO, etc.),which leads to increase of target function. FIGURE 2 shows the process for an individual attribute and illustrates the role of D_{start} and role of $D_{precision}$.

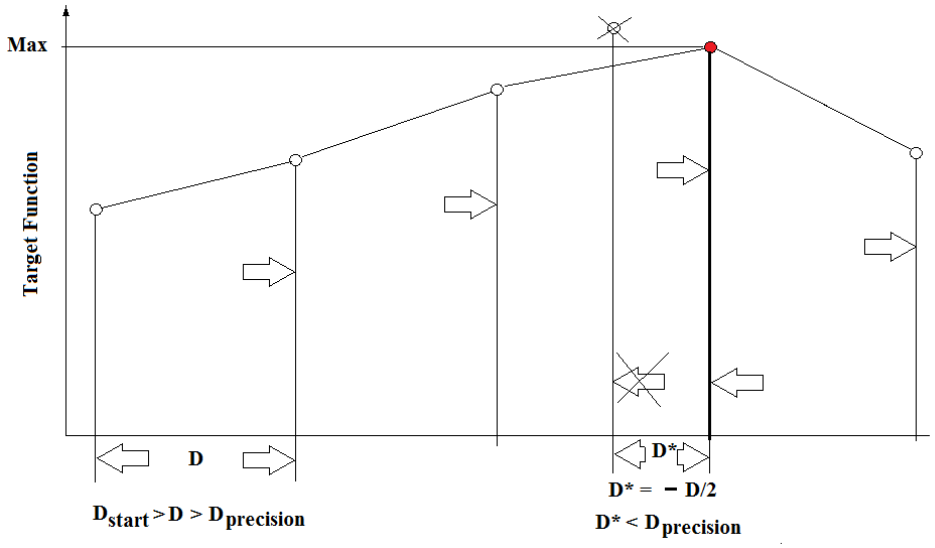


FIGURE 2

Three target functions are available:
(1) The classic scheme, i.e., [Training-Test] system;
(2) Balance of correlations, i.e., '[Sub-training – Calibration – Test] system;
(3) Balance of correlations with ideal slopes.

The first function keep into account only **R**, which is the correlation coefficient between endpoint and optimal descriptor calculated with Eq.1 for the training set. Thus the optimization is the following:
R → maxR

The second function is **BC=R+R' – abs(R-R')*dR_{weight},** balance of correlations: R and R' are correlation coefficient between endpoint and optimal descriptor for sub-training set and calibration set. The role of the calibration set is a preliminary validation of the model. This approach is an attempt to avoid the overtraining. In other words, in the case of balance of correlations, the training set is split into two sets: subtraining and calibration. The **dR_{weight}** is an empirical parameter. This optimization is the following:
BC → maxBC

The third function is **IS=BC – abs(C0+C0'+C1-C1')*dC_{weight},** balance of correlation with ideal slopes: C0 and C0' are intercepts for the sub-training set and calibration set; C1 and C1' are slopes for the sub-training set and calibration set. The balance of correlations can classify as satisfactory the model represented in FIGURE 3. The balance of correlation with ideal slopes is an attempt to avoid the situation. The **dC_{weight}** is an empirical parameter. This optimization is the following:
IS → maxIS

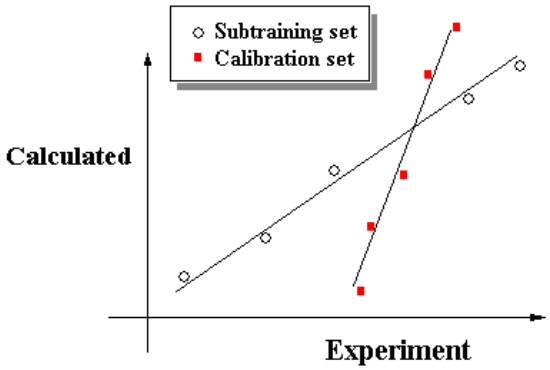


FIGURE 3
FIGURE 4 shows how you can select the target function

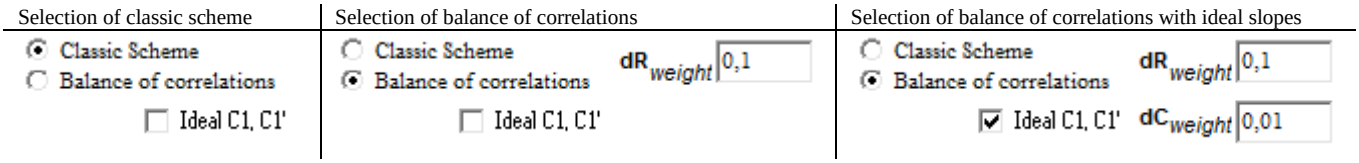


FIGURE 4

2.5. Sketch of theory

Theoretically, the correlation coefficients between experimental and calculated values of the endpoint for sub-training, calibration, and test sets are a mathematical functions of threshold and the number of epochs. FIGURE 5 illustrates this situation.

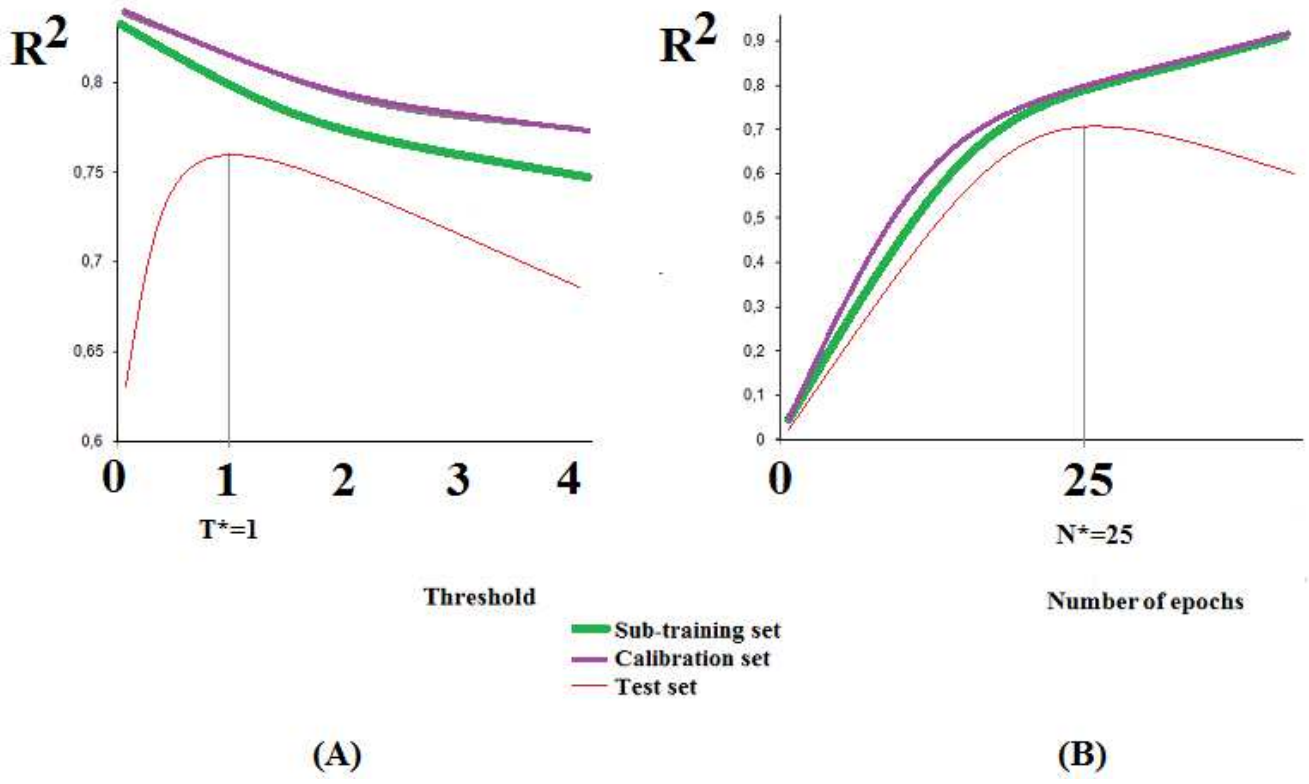


FIGURE 5

It is necessary to choose the Threshold and N_{epoch} which can give satisfactory statistical characteristics *for the test set*. In fact it is the maximum in the surface of R^2_{test} = F(Threshold, N_{epoch}).

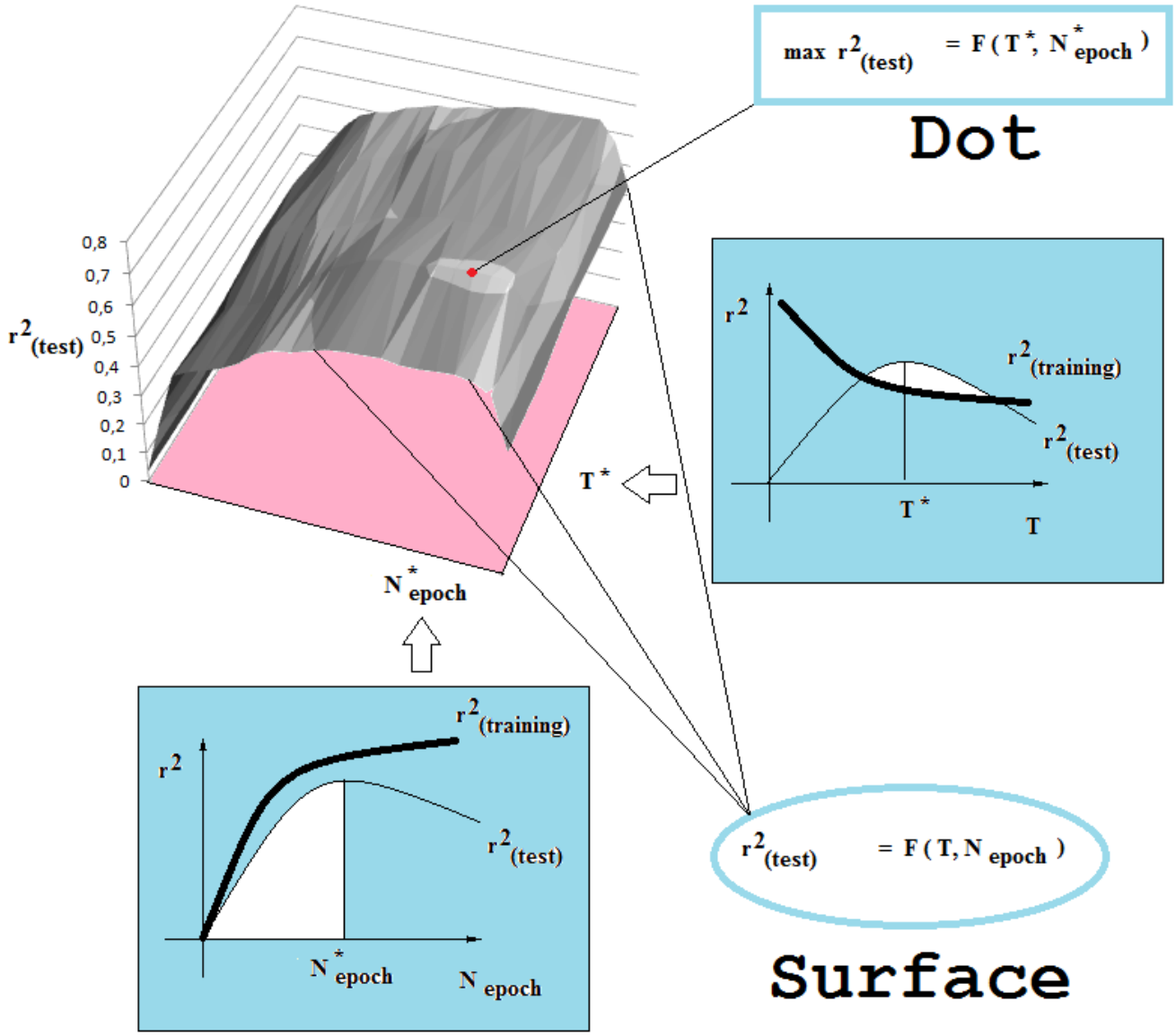


FIGURE 6

Thus, the main aim of the CORALSEA software may be formulated as the search for T* and N* which are producing the maxR^2_{test} (FIGURE 6).

Step 3. Searching for the best threshold (T*) and the best number of epochs (N*). Structure of output data

If you have started CORALSEA.exe from CORALSEA folder (downloaded from our web site) and if you have clicked button “Load method”, then you will see situation shown in FIGURE 7.
After you click “Search for preferable model (T*,N*)”, the program will ask you to confirm that files which take place in Search folder may be deleted (FIGURE 8)
When the calculation is completed, the program displays the message that work is completed and you can start analysis of results (FIGURE 9).

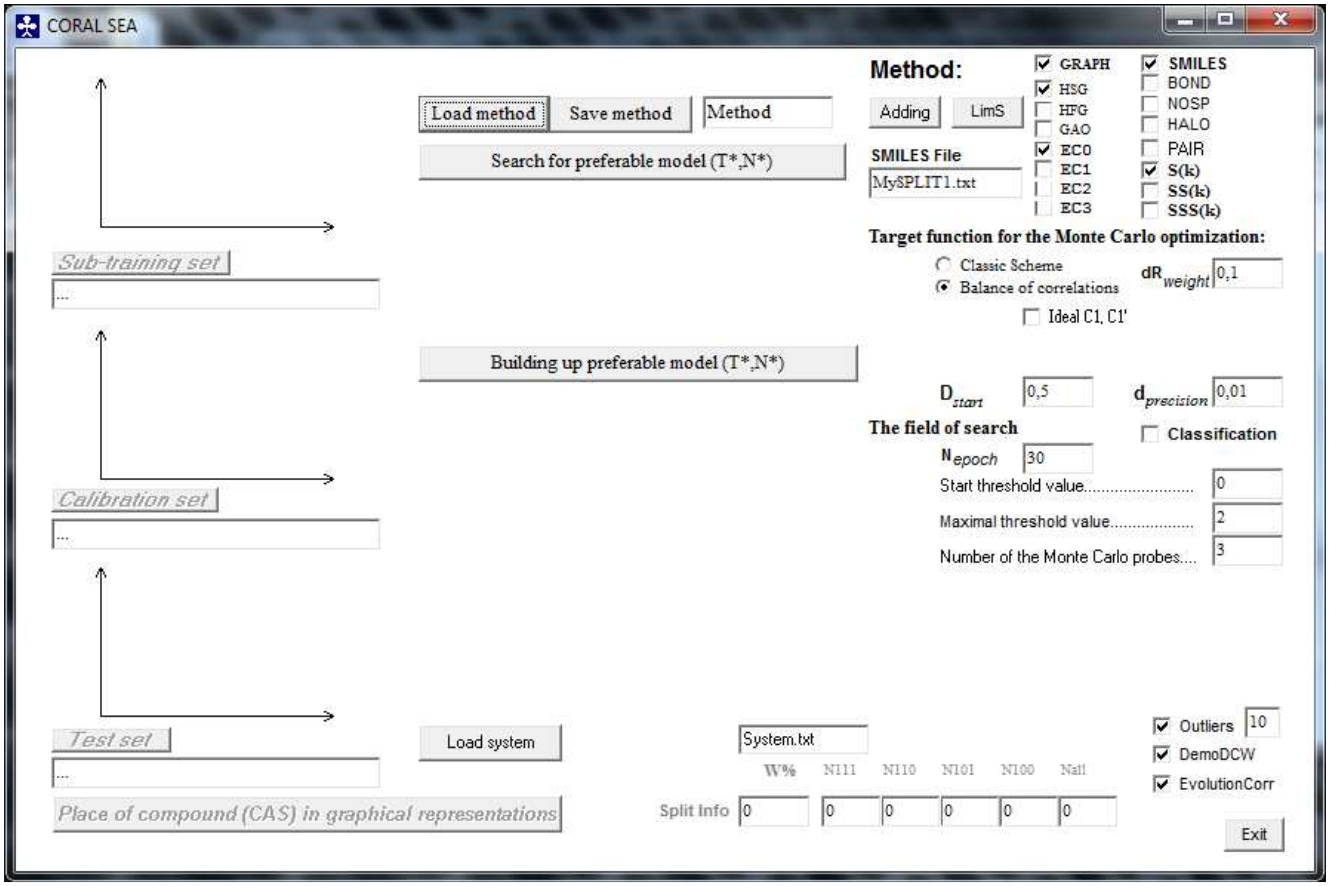


FIGURE 7

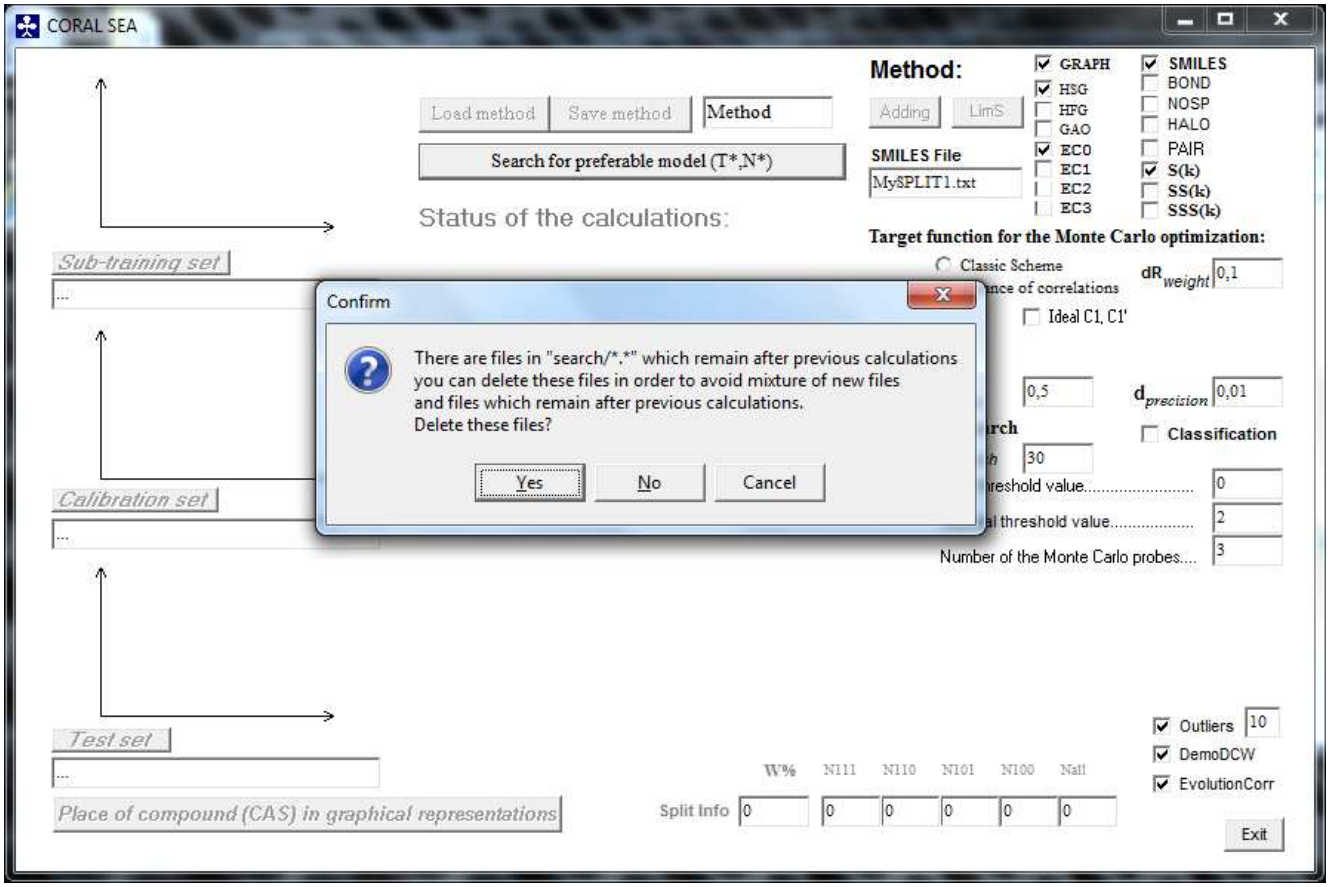


FIGURE 8

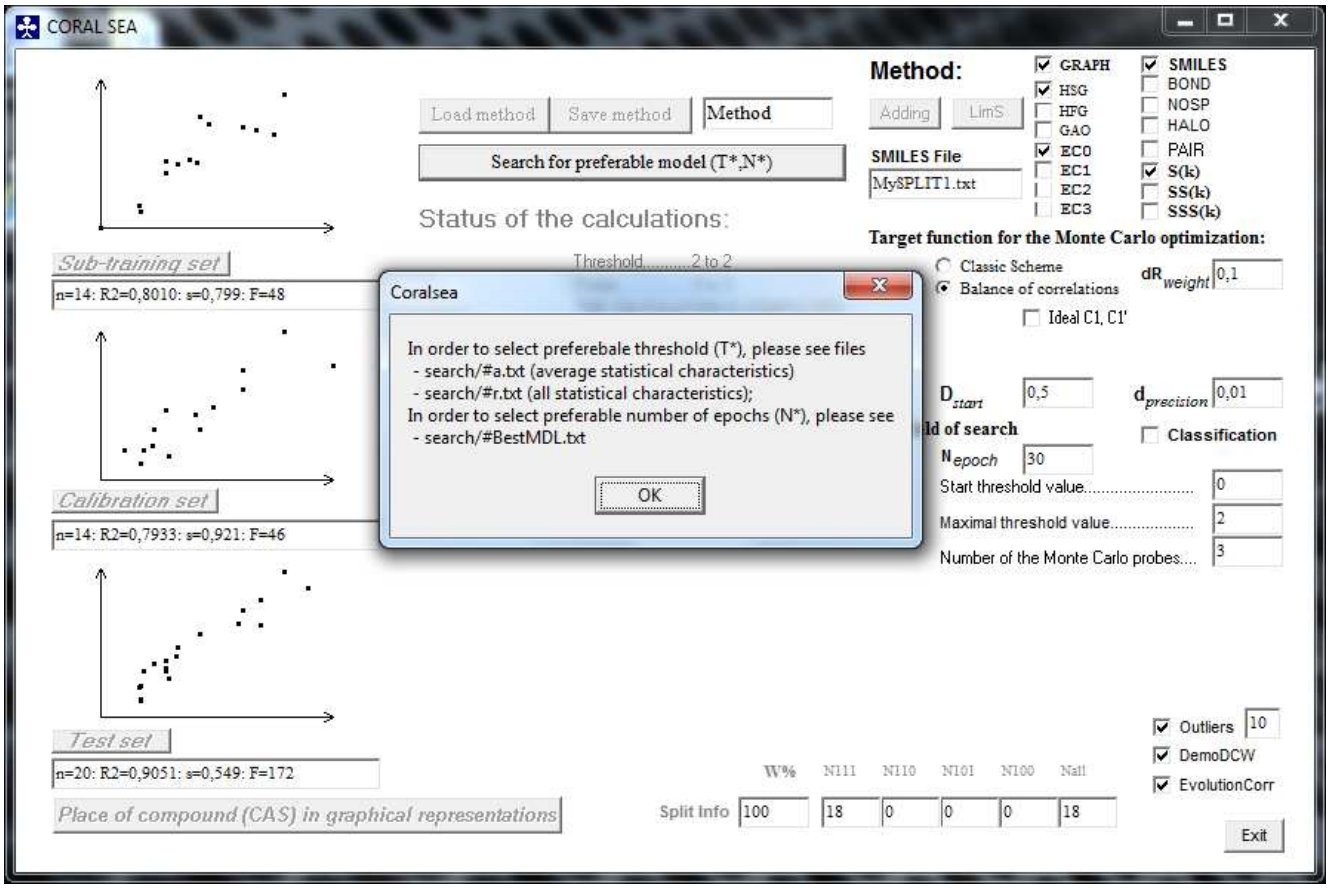


FIGURE 9

3.1. Search/#a.txt

File Search/#a.txt contains the average statistical characteristics of the models for the selected range of threshold and the selected number of epochs:

```
This file contains average values for statistical characteristics which were obtained
in 3 probes of the Monte Carlo optimization
SMILES taken from file MySPLIT1.txt
Method taken from file Method

Hydrogen suppressed graph (HSG) is used in the model
SMILES is used in the model
Threshold from 0 to 2

n is the number of compounds in set;
r is Correlation coefficient;
s is standard error of estimation;
F is Fischer F-ratio.
ns, r2s, ss, and Fs are statistical characteristics of subtraining set.
nc, r2c, sc, and Fc are statistical characteristics of calibration set.
nv, r2v, sv, and Fv are statistical characteristics of validation set.
Rm2av is the average of Rm2 metric, it should be > 0.5 [1]
[1] PK Ojha,I Mitra, RN Das,K Roy,Chemometr Intell Lab 107(2011)194-205
Number of epochs of optimization is 30
Number of probes of optimization is 3
Weight for dr in balance of correlations is 0,1
Start step of the optimization is 0,5*CW(SA)
Precision of the optimization is 0,01*CW(SA)
CW(SA) is weight of SA at the start

N111 is the number of SA which take place in subtraining,
in calibration, and in validation sets
W% is ratio of N111/(number of all SA)
N110 is number of SA which take place in subtraining and calibration
N101 is number of SA which take place in subtraining and validation
N100 is number of SA which take place in subtraining (only)

Model is build up by means of
Balance of correlations
The dr-weight is 0,1

EC0 is involved
SS(k) are off
SSS(k) are off

Trshd:Nact :  ns :    rs2:       ss :    fs :  nc :    rc2:       sc :    fc :  nv :    rv2:       sv :    fv : Rm2av : W% :N111:N110:N101:N100
    0:   20:   14: 0.8658:     0.656:     77:   14: 0.8307:     0.929:     59:   20: 0.8994:     0.597:    161: 0.8551: 100:  20:   0:   0:   0
    1:   20:   14: 0.8642:     0.660:     76:   14: 0.8298:     0.915:     59:   20: 0.9025:     0.578:    167: 0.8618: 100:  20:   0:   0:   0
    2:   18:   14: 0.7983:     0.805:     47:   14: 0.7975:     0.907:     47:   20: 0.9059:     0.546:    173: 0.8357: 100:  18:   0:   0:   0
```

FIGURE 10

3.2. Search/#r.txt

File Search/#r.txt contains statistical characteristics of the models for each probes:

```
This file contains values for statistical characteristics which were obtained
in 3 probes of the Monte Carlo optimization
SMILES taken from file MySPLIT1.txt
Method taken from file Method

Hydrogen suppressed graph (HSG) is used in the model
SMILES is used in the model

n is the number of compounds in set;
r is Correlation coefficient;
s is standard error of estimation;
F is Fischer F-ratio.
ns, r2s, ss, and Fs are statistical characteristics of subtraining set.
nc, r2c, sc, and Fc are statistical characteristics of calibration set.
nv, r2v, sv, and Fv are statistical characteristics of validation set.
Rm2 metric should be > 0.5 [1]
[1] PK Ojha,I Mitra, RN Das,K Roy,Chemometr Intell Lab 107(2011)194-205]
Number of epochs of optimization is 30
Number of probes of optimization is 3
Threshold from 0 to 2
Start step of the optimization is 0,5*CW(SA)
Precision of the optimization is 0,01*CW(SA)
CW(SA) is weight of SA at the start

Model is build up by means of
EC0 is involved
SS(k) are off
SSS(k) are off
Balance of correlations
The dr-weight is 0,1


Trshd:Nact :Probe:  ns :    rs2:       ss :    fs :  nc :    rc2:       sc :    fc :  nv :    rv2:       sv :    fv :  Rm2
    0:   20:    1:   14: 0.8655:     0.657:     77:   14: 0.8303:     0.929:     59:   20: 0.9022:     0.588:    166: 0.8582
    0:   20:    2:   14: 0.8662:     0.655:     78:   14: 0.8306:     0.933:     59:   20: 0.8961:     0.610:    155: 0.8505
    0:   20:    3:   14: 0.8658:     0.656:     77:   14: 0.8312:     0.925:     59:   20: 0.8999:     0.594:    162: 0.8566
    0:    :     :      : 0.8658:     0.656:     77:      : 0.8307:     0.929:     59:      : 0.8994:     0.597:    161: 0.8551
    1:   20:    1:   14: 0.8627:     0.664:     75:   14: 0.8293:     0.897:     58:   20: 0.9016:     0.569:    165: 0.8652
    1:   20:    2:   14: 0.8639:     0.661:     76:   14: 0.8317:     0.917:     59:   20: 0.9015:     0.583:    165: 0.8617
    1:   20:    3:   14: 0.8659:     0.656:     77:   14: 0.8285:     0.932:     58:   20: 0.9043:     0.581:    170: 0.8585
    1:    :     :      : 0.8642:     0.660:     76:      : 0.8298:     0.915:     59:      : 0.9025:     0.578:    167: 0.8618
    2:   18:    1:   14: 0.7988:     0.804:     48:   14: 0.7975:     0.909:     47:   20: 0.9042:     0.551:    170: 0.8335
    2:   18:    2:   14: 0.7978:     0.806:     47:   14: 0.7979:     0.907:     47:   20: 0.9067:     0.544:    175: 0.8360
    2:   18:    3:   14: 0.7982:     0.805:     47:   14: 0.7970:     0.906:     47:   20: 0.9068:     0.542:    175: 0.8376
    2:    :     :      : 0.7983:     0.805:     47:      : 0.7975:     0.907:     47:      : 0.9059:     0.546:    173: 0.8357
```

FIGURE 11

File Search/#BestMDL.txt contains data on the best models for test set:

```
Data from MySPLIT1.txt
Representation of molecular structure by
* Hydrogen suppresed graph
* SMILES

Correlation coefficients for the test set
 Threshold:Probe  1:Probe  2:Probe  3: Average  :Dispersion
         0:  0.9356:  0.9320:  0.9341:    0.9339:    0.0015
         1:  0.9387:  0.9392:  0.9368:    0.9382:    0.0010
         2:  0.9126:  0.9153:  0.9143:    0.9140:    0.0011

Preferable the number of epochs of the Monte Carlo optimization
 Threshold:Probe  1:Probe  2:Probe  3: Average  :Dispersion
         0:       7:       6:       7:       6.67:      0.47
         1:       6:       7:       7:       6.67:      0.47
         2:       7:       9:       7:       7.67:      0.94
```

FIGURE 12

One can see from the data shown in FIGURE 12 that for given substances and used split (into the sub-training, calibration and test sets) the preferable threshold is T*=1, and the preferable number of epochs is N*=6.67≈7. Thus, for given split and selected method (FIGURE 7) most informative descriptors is

$$^{Hybrid}DCW(1,7)=^{Graph}DCW(1,7)+^{SMILES}DCW(1,7) \quad (6)$$

where

$$^{Graph}DCW(1,7)=\sum CW(A_k)+\sum CW(^{0}EC_k) \quad (7)$$

$$^{SMILES}DCW(1,7)=\sum CW(S_k) \quad (8)$$

The CORALSEA software gives also technical details, which can be used in research work related to QSPR/QSAR analyses. This information is located in group of files which are represented in Table 1.

Table 1.
Groups of files generated by the CORALSEA software

Type	**Description / Format**	**Completed list of the names of files if threshold diapason is 0-2 and the number of probes is 3**
d	These files contain examples of DCW(Threshold,N_{epoch}) calculation **/d(threshold)-(number of probe).txt**	d0-1.txt,d0-2.txt,d0-3.txt, d1-1.txt,d1-2.txt,d1-3.txt, d2-1.txt,d2-2.txt,d2-3.txt
e	These files contains co-evolution of correlations, i.e. correlations coefficient between experimental and calculated values of an endpoint for subtraining (training), calibration, and test sets in series of epochs **/e(threshold)-(number of probe).txt**	e0-1.txt,e0-2.txt,e0-3.txt, e1-1.txt,e1-2.txt,e1-3.txt, e2-1.txt,e2-2.txt,e2-3.txt
i	These files contains a sequence of idealization of the model for test set by means of removing of a sequence of "worst" outliers **/i(threshold)-(number of probe).txt**	i0-1.txt,i0-2.txt,i0-3.txt, i1-1.txt,i1-2.txt,i1-3.txt, i2-1.txt,i2-2.txt,i2-3.txt
m	These files contain examples of the endpoint model **/m(threshold)-(number of probe).txt**	m0-1.txt,m0-2.txt,m0-3.txt, m1-1.txt,m1-2.txt,m1-3.txt, m2-1.txt,m2-2.txt,m2-3.txt
s	These files contains ordered values of correlation weights for	s0.txt, s1.txt,

	all probes of the Monte Carlo optimization **/s(threshold)**	`s2.txt`
w	These files contains correlation weights for given values of the threshold and number of probe **/w(threshold)-(number of probe).txt**	`w0-1.txt,w0-2.txt,w0-3.txt, w1-1.txt,w1-2.txt,w1-3.txt, w2-1.txt,w2-2.txt,w2-3.txt`

The building up of the sub-groups of files which are indicated by blue can be blocked (Table 2).

Table 2
Definition of the list of files for the output

Option	**Operation**	**Option**	**Operation**
☑ Outliers 5	The sequence of five "worst" outliers for each threshold and each probe is saved in i-files (Table 1)	☐ Outliers 5	The building up of i-files is blocked
☑ DemoDCW	The demonstration of DCW-calculations (the first substance in the list) for each threshold and each probe are saved in d-files (Table 1)	☐ DemoDCW	The building up of d-files is blocked
☑ EvolutionCorr	Data on correlation coefficients for sub-training, calibration, and test sets, for each epoch are saved in e-files (Table 1)	☐ EvolutionCorr	The building up of e-files is blocked

3.3. d-Files

FIGURE 13 shows an example of d-file. The adjacency matrix is typed if graph attributes are involved in the modeling process (if not, the matrix is absent).

```
This file contains example of DCW-calculation

SMILES:  [O-][N+](=O)c1ccc2ccccc2c1
Set (#):   Subtraining +; Calibration -; Validation #
Number of structure 1
CAS  14
Number of Monte Carlo optimization probe 1
Threshold 1
Hydrogen suppressed Graph (HSG) is used in the model
SMILES is used in the model

ID is the number of SMILES attribute (SA) in global list of SA
CW(SA) is the correlation weight for SA
NSs is the number of SA in subtraining set
NCs is the number of SA in calibration set
NVs is the number of SA in validation set

 Structural :           :     :     :     :
  attribute :   CW(SA)   : ID  : NSs : NSc : NVs
    (SA)    :            :     :     :     :

EC0-O...1...:      0.6162:   15:   14:   14:   20
EC0-N...3...:      0.6366:   14:   14:   14:   20
EC0-O...1...:      0.6162:   15:   14:   14:   20
EC0-C...3...:      1.9259:   12:   14:   14:   20
EC0-C...2...:      0.3730:   11:   14:   14:   20
EC0-C...2...:      0.3730:   11:   14:   14:   20
EC0-C...3...:      1.9259:   12:   14:   14:   20
EC0-C...2...:      0.3730:   11:   14:   14:   20
EC0-C...2...:      0.3730:   11:   14:   14:   20
EC0-C...2...:      0.3730:   11:   14:   14:   20
EC0-C...2...:      0.3730:   11:   14:   14:   20
EC0-C...3...:      1.9259:   12:   14:   14:   20
EC0-C...2...:      0.3730:   11:   14:   14:   20
[...........:      0.5429:   18:   14:   14:   20
O...........:      0.9501:   17:   14:   14:   20
-...........:     -0.1134:    3:   14:   14:   20
[...........:      0.5429:   18:   14:   14:   20
[...........:      0.5429:   18:   14:   14:   20
N...........:     -0.7071:   16:   14:   14:   20
+...........:     -0.4141:    2:   14:   14:   20
[...........:      0.5429:   18:   14:   14:   20
(...........:     -1.5118:    1:   14:   14:   20
=...........:      1.7149:    8:   14:   14:   20
O...........:      0.9501:   17:   14:   14:   20
(...........:     -1.5118:    1:   14:   14:   20
c...........:     -0.2813:   19:   14:   14:   20
1...........:      9.0888:    4:   14:   14:   20
c...........:     -0.2813:   19:   14:   14:   20
c...........:     -0.2813:   19:   14:   14:   20
c...........:     -0.2813:   19:   14:   14:   20
2...........:     -0.2921:    5:   10:   10:   13
c...........:     -0.2813:   19:   14:   14:   20
c...........:     -0.2813:   19:   14:   14:   20
c...........:     -0.2813:   19:   14:   14:   20
c...........:     -0.2813:   19:   14:   14:   20
c...........:     -0.2813:   19:   14:   14:   20
2...........:     -0.2921:    5:   10:   10:   13
c...........:     -0.2813:   19:   14:   14:   20
1...........:      9.0888:    4:   14:   14:   20

DCW=    26.56700

The Adjacency Matrix of the molecular Graph

      O   N   O   c   c   c   c   c   c   c   c   c   c
  O   0   1   0   0   0   0   0   0   0   0   0   0   0:   1
  N   1   0   2   1   0   0   0   0   0   0   0   0   0:   3
  O   0   2   0   0   0   0   0   0   0   0   0   0   0:   1
  c   0   1   0   0   1   0   0   0   0   0   0   0   1:   3
  c   0   0   0   1   0   1   0   0   0   0   0   0   0:   2
  c   0   0   0   0   1   0   1   0   0   0   0   0   0:   2
  c   0   0   0   0   0   1   0   1   0   0   0   1   0:   3
  c   0   0   0   0   0   0   1   0   1   0   0   0   0:   2
  c   0   0   0   0   0   0   0   1   0   1   0   0   0:   2
  c   0   0   0   0   0   0   0   0   1   0   1   0   0:   2
  c   0   0   0   0   0   0   0   0   0   1   0   1   0:   2
  c   0   0   0   0   0   0   1   0   0   0   1   0   1:   3
  c   0   0   0   1   0   0   0   0   0   0   0   1   0:   2
```

FIGURE 13

3.4. e-Files

FIGURE 14 shows an example of e-file: R2-sub, R2-clb, and R2-tst are squares of correlation coefficients for the sub-training, calibration, and test sets, respectively; s-sub, s-clb, and s-tst are standard error for sub-training, calibration, and test sets, respectively (FIGURE 5B).

The evolution of correlations. The R is correlation coefficicnet; s is standard error

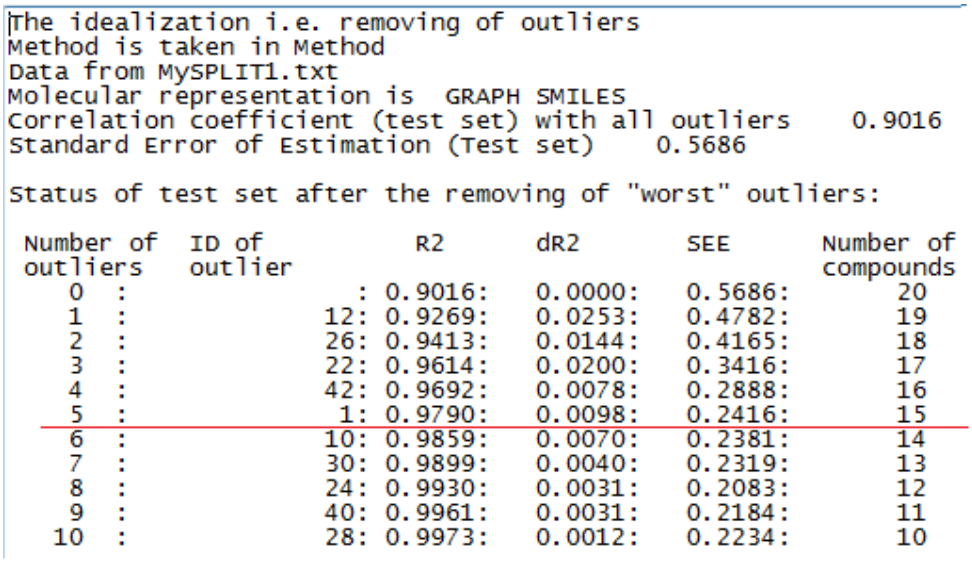
No.	R2-sub	R2-clb	R2-tst	s-sub	s-clb	s-tst
1:	0.7909:	0.7307:	0.8794:	0.819:	0.999:	0.618
2:	0.7871:	0.7575:	0.9012:	0.827:	0.953:	0.565
3:	0.7837:	0.7797:	0.9172:	0.833:	0.915:	0.522
4:	0.7888:	0.7898:	0.9279:	0.823:	0.897:	0.489
5:	0.7958:	0.7974:	0.9341:	0.810:	0.882:	0.467
6:	0.8008:	0.8061:	0.9387:	0.800:	0.870:	0.447
7:	0.8087:	0.8074:	0.9363:	0.784:	0.867:	0.449
8:	0.8147:	0.8110:	0.9338:	0.771:	0.858:	0.453
9:	0.8171:	0.8159:	0.9317:	0.766:	0.857:	0.456
10:	0.8224:	0.8164:	0.9344:	0.755:	0.856:	0.446
11:	0.8243:	0.8214:	0.9285:	0.751:	0.849:	0.463
12:	0.8276:	0.8247:	0.9281:	0.744:	0.850:	0.463
13:	0.8273:	0.8280:	0.9261:	0.745:	0.843:	0.469
14:	0.8291:	0.8286:	0.9260:	0.741:	0.845:	0.469
15:	0.8321:	0.8281:	0.9259:	0.734:	0.850:	0.470
16:	0.8324:	0.8301:	0.9233:	0.733:	0.850:	0.479
17:	0.8381:	0.8285:	0.9226:	0.721:	0.857:	0.481
18:	0.8443:	0.8264:	0.9216:	0.707:	0.866:	0.486
19:	0.8455:	0.8272:	0.9179:	0.704:	0.864:	0.498
20:	0.8461:	0.8287:	0.9157:	0.703:	0.867:	0.507
21:	0.8470:	0.8300:	0.9141:	0.701:	0.867:	0.513
22:	0.8499:	0.8294:	0.9157:	0.694:	0.873:	0.510
23:	0.8517:	0.8297:	0.9144:	0.690:	0.872:	0.515
24:	0.8522:	0.8308:	0.9096:	0.689:	0.874:	0.532
25:	0.8565:	0.8292:	0.9107:	0.679:	0.887:	0.534
26:	0.8560:	0.8311:	0.9065:	0.680:	0.885:	0.548
27:	0.8602:	0.8288:	0.9042:	0.670:	0.892:	0.556
28:	0.8584:	0.8314:	0.9007:	0.674:	0.890:	0.569
29:	0.8627:	0.8288:	0.9020:	0.664:	0.896:	0.567
30:	0.8627:	0.8293:	0.9016:	0.664:	0.897:	0.569

FIGURE 14

3.5. i-Files

FIGURE 15 shows an example of i-file. The number of “worst” outliers (Nw) is limited: the number of structures in test set *must be more 10*. If the number of structures in the test set is less than 10, then these calculations are blocked.

dR2(k+1)=R2(k)-R2(k+1), k=0,Nw-1.
E.g the calculation of dR2(5) from FIGURE 15 is the following 0.9790-0.9692=0.0098



```
The idealization i.e. removing of outliers
Method is taken in Method
Data from MySPLIT1.txt
Molecular representation is  GRAPH SMILES
Correlation coefficient (test set) with all outliers    0.9016
Standard Error of Estimation (Test set)    0.5686

Status of test set after the removing of "worst" outliers:

 Number of  ID of          R2      dR2       SEE       Number of
 outliers   outlier                                    compounds
    0  :              : 0.9016:   0.0000:   0.5686:       20
    1  :             12: 0.9269:   0.0253:   0.4782:       19
    2  :             26: 0.9413:   0.0144:   0.4165:       18
    3  :             22: 0.9614:   0.0200:   0.3416:       17
    4  :             42: 0.9692:   0.0078:   0.2888:       16
    5  :              1: 0.9790:   0.0098:   0.2416:       15
    6  :             10: 0.9859:   0.0070:   0.2381:       14
    7  :             30: 0.9899:   0.0040:   0.2319:       13
    8  :             24: 0.9930:   0.0031:   0.2083:       12
    9  :             40: 0.9961:   0.0031:   0.2184:       11
   10  :             28: 0.9973:   0.0012:   0.2234:       10
```

FIGURE 15

3.6. m-Files

These files contain technical details related to the calculated models for the endpoint.
One can consider three sub-sections in the m-file.

Sub-section 1

Documentation of used files (Method, MySplit1.txt, etc.); description of the statistical characteristics; and results of the Y-scrambling (FIGURE 16).

```
This file contains experimental and calculated values of the endpoint
Hydrogen suppressed graph (HSG) is used in the model
SMILES is used in the model
Data from SMILES-file (MySPLIT1.txt)
Threshold=1
The number of active SMILES attributes (ASA) =20
IMPORTANT: In the case of classic scheme W%=N101/Nall, otherwise W%=N111/Nall
Percent of ASA with presence in all sets (W%) =100
Intercept (c0) and slope (c1) calculated for each set individually:
Subtraining set: c0=  -6.01661 c1=   0.24249
Calibration set: c0=  -5.92283 c1=   0.22423
Validation set : c0=  -5.75279 c1=   0.22859
Slope and intesept calculated with subtraining set give the model:
Endpoint =  -6.0166000+-   0.2484570   0.2425000+-   0.0076709 * DCW(1,30)
Statistical characteristics of the model:
N is the number of compounds in the set;
R is correlation coefficient;
Q is cross-validated correlation coefficient;
s is standard error of estimation;
MAE is mean absolute error;
F is Fischer F-ratio
Blk is the number of SMILES attributes in given SMILES, which are blocked
Nall is the number of all SMILES attributes in given SMILES string
Y-Scrambling:   14 trails for each average:
The number of trails is equal to number of compounds in
sub-training set
                                : Train : Calib : Test
                                :     14:     14:     20
                                : 0.8627: 0.8293: 0.9016
                               1: 0.4288: 0.3441: 0.0224
                               2: 0.1461: 0.6936: 0.6842
                               3: 0.2197: 0.0192: 0.0906
                               4: 0.1448: 0.0277: 0.1344
                               5: 0.3624: 0.8064: 0.5097
                               6: 0.6528: 0.6466: 0.5852
                               7: 0.2645: 0.6646: 0.2705
                               8: 0.6450: 0.4063: 0.4647
                               9: 0.5858: 0.8279: 0.0890
                              10: 0.0842: 0.6184: 0.1979
Rr2, i.e. average randomized R   : 0.3534: 0.5055: 0.3049
   CRp2=R*sqrt(R2-Rr2) [1]       : 0.6628: 0.5182: 0.7335:
 CRp2 should be greater 0.5 [1]
REFERENCE for Y-scrambling
[1] P.K. Ojha, K. Roy, Comparative QSARs for antimalarial endochins:
    Importance of descriptor-thinning and noise reduction prior to
     feature selection, Chemometr. Intell. Lab. 109 (2011) 146-161
```

FIGURE 16

Sub-section 2

External validation according to criteria from the literature; numerical data on the statistical characteristics of the model (FIGURE 17).

```
External validation characteristics for the model taken from
REFERNCES
[1] Golbraikh A., Tropsha A. J.Mol.Graph.Model. 20(2002)269; // R02, k,kk
[2] Roy P.P., Roy K. Chem. Biol. Drug Des. 73(2009) 442; // Rm2
[3] PK Ojha,I Mitra, RN Das,K Roy,Chemometr Intell Lab 107(2011)194-205
    // Average of Rm2 and absolute difference Rm2(x,y)-Rm2(y,x)
    // x,y are experimental and predicted values of endpoint

The range of endpoint:
Min= -2.1 Max=  4.7 Middle=  1.3

n           =      20
r2          =    0.9016
r02         =    0.8924
rr02        =    0.9000
(r2-r02)/r2 =    0.0103 should be < 0.1 [1]
(r2-rr02)/r2=    0.0018 should be < 0.1 [1]
k           =    1.0052 should be 0.85 <  k < 1.15 [1]
kk          =    0.9205 should be 0.85 < kk < 1.15 [1]
Rm2(test)   =    0.8149 should be > 0.5 [2]

n           =      20
r2          =    0.9016
r02         =    0.9000
rr02        =    0.8924
(r2-r02)/r2 =    0.0018 should be < 0.1 [1]
(r2-rr02)/r2=    0.0103 should be < 0.1 [1]
k           =    0.9205 should be 0.85 <  k < 1.15 [1]
kk          =    1.0052 should be 0.85 < kk < 1.15 [1]
R*m2(test)  =    0.8652 should be > 0.5 [2]

Average Rm2 = 0.8400 should be larger 0.5 [3]
Delta Rm2 = 0.0503 should be lower 0.2 [3]

        :  n :   R2    :  Q2    :      s   :     MAE :  F
SubTrain:  14: 0.8627: 0.8183:    0.664:    0.496:       75
Calibrat:  14: 0.8293: 0.7221:    0.897:    0.737:       58
External:  20: 0.9016: 0.8696:    0.569:    0.380:      165

Subtraining set is indicated by +;
Calibration set is indicated by -;
Validation set is indicated by #

N.B.: If the training-test system is used then calibration set is absent:
      the subtraining set in this case should read as training set.
```

FIGURE 17

Sub-section 3

Lists of substances (their SMILES) involved in the sub-training set (+), in the calibration set (-), and in the test set (#); numerical data on the DCW(threshold, N_{epoch}); experimental and calculated values of the endpoint; the numbers of blocked structural attributes (blk) and total number (all) of structural attributes for each substance (SMILES and/or Graph); and ID for each substance.

:SMILES	: DCW:	Expr:	Calc:	Expr-Calc:	Blk/All:	ID
+:[O-][N+](=O)c1ccc3ccc4c2c(ccc1c23)ccc4[N+]([O-])=O	: 41.55863:	4.090:	4.061:	0.029:	0/ 72:	4
+:[O-][N+](=O)c2ccc3c1ccc(cc1C(=O)c3c2)[N+]([O-])=O	: 33.28100:	2.690:	2.054:	0.636:	0/ 69:	6
+:[O-][N+](=O)c1ccc2cc3ccccc3cc2c1	: 31.63350:	3.050:	1.655:	1.395:	0/ 49:	8
+:[O-][N+](=O)c3ccc4c2ccccc1cccc(c12)c4c3	: 36.60825:	2.600:	2.861:	-0.261:	0/ 57:	11
+:[O-][N+](=O)c1ccc2c3ccccc3Cc2c1	: 29.54375:	1.080:	1.148:	-0.068:	0/ 47:	15
+:[O-][N+](=O)c1ccc2C=Cc3cccc1c23	: 29.16894:	0.970:	1.057:	-0.087:	0/ 46:	18
+:Cc1ccc(cc1)[N+](=O)[O-]	: 18.95562:	-2.100:	-1.420:	-0.680:	0/ 33:	21
+:Cc1c(cccc1[N+](=O)[O-])[N+]([O-])=O	: 23.90600:	-1.340:	-0.219:	-1.121:	0/ 48:	25
+:O=[N+]([O-])c1cc(cc(C)c1[N+]([O-])=O)[N+]([O-])=O	: 25.83275:	0.460:	0.248:	0.212:	0/ 65:	29
+:O=[N+]([O-])c1c(cc(C)cc1[N+]([O-])=O)[N+]([O-])=O	: 25.83275:	1.010:	0.248:	0.762:	0/ 65:	33
+:[O-][N+](=O)c2cc(cc1ccccc12)[N+]([O-])=O	: 28.49375:	0.860:	0.893:	-0.033:	0/ 56:	36
+:[O-][N+](=O)c3cc2c(c1c(cc(cc1C2=O)[N+]([O-])=O)[N+]([O-])=O)c(c3)[N+]([O-])=O	: 37.13450:	2.460:	2.989:	-0.529:	0/103:	39
+:[O-][N+](=O)c1ccc2ncccc2c1	: 18.68800:	-1.050:	-1.485:	0.435:	0/ 39:	43
+:[O-][N+](=O)c4cc2c(ccc1ccccc12)c3ccccc34	: 36.79175:	2.210:	2.905:	-0.695:	0/ 61:	48
-:[O-][N+](=O)c4ccc1ccc2ccc([N+]([O-])=O)c3ccc4c1c23	: 41.55862:	4.740:	4.061:	0.679:	0/ 72:	5
-:[O-][N+](=O)c2cc4cccc3c1ccccc1c(c2)c34	: 36.60825:	3.000:	2.861:	0.139:	0/ 57:	7
-:[O-][N+](=O)c1ccc2c3ccc(cc3Cc2c1)[N+]([O-])=O	: 31.47050:	1.270:	1.615:	-0.345:	0/ 64:	9
-:[O-][N+](=O)c3ccc4c2ccccc1cccc(c12)c4c3	: 36.60825:	2.090:	2.861:	-0.771:	0/ 57:	13
-:[O-][N+](=O)c2ccccc1ccccc12	: 26.56700:	0.280:	0.426:	-0.146:	0/ 39:	17
-:O=[N+]([O-])c1cccc(c1)[N+]([O-])=O	: 25.60512:	-0.510:	0.193:	-0.703:	0/ 46:	19
-:Cc1ccc(cc1[N+](=O)[O-])[N+]([O-])=O	: 23.90600:	-1.290:	-0.219:	-1.071:	0/ 48:	23
-:O=[N+]([O-])c1ccc(C)cc(c1)[N+]([O-])=O	: 20.88237:	-0.720:	-0.953:	0.233:	0/ 50:	27
-:O=[N+]([O-])c1ccc(c(C)cc1[N+]([O-])=O)[N+]([O-])=O	: 25.83275:	1.120:	0.248:	0.872:	0/ 65:	31
-:[O-][N+](=O)c1cc2ccccc2cc1C	: 24.86788:	-0.700:	0.014:	-0.714:	0/ 41:	35
-:[O-][N+](=O)c2cccc1c2cccc1[N+]([O-])=O	: 31.51738:	0.910:	1.626:	-0.716:	0/ 54:	37
-:[O-][N+](=O)c4cc(c1ccc2c(cc([N+]([O-])=O)c3ccc4c1c23)[N+]([O-])=O)[N+]([O-])=O~	: 45.41213:	3.180:	4.996:	-1.816:	0/106:	41
-:[O-][N+](=O)c1cc2c3ccccc3nc2cc1	: 23.66275:	-1.000:	-0.278:	-0.722:	0/ 47:	45
-:[O-][N+](=O)c2c3ccccc3cc1ccccc12	: 31.63350:	0.260:	1.655:	-1.395:	0/ 49:	47
#:[O-][N+](=O)c1ccc2ccccc2c1	: 26.56700:	0.370:	0.426:	-0.056:	0/ 39:	14
#:[O-][N+](=O)c1ccc3ccc4c2c(ccc1c23)c(cc4[N+]([O-])=O)[N+]([O-])=O	: 43.48538:	3.870:	4.529:	-0.659:	0/ 89:	1
#:[O-][N+](=O)c1cc2Cc3cc(cc(c3c2cc1)[N+]([O-])=O)[N+]([O-])=O	: 33.39725:	2.270:	2.082:	0.188:	0/ 81:	2
#:[O-][N+](=O)c1cc([N+]([O-])=O)c4ccc3cccc2ccc1c4c23	: 41.55862:	4.630:	4.061:	0.569:	0/ 72:	3
#:[O-][N+](=O)c4ccc2c1ccccc1c3cccc4c23	: 39.63188:	3.310:	3.594:	-0.284:	0/ 55:	10
#:[O-][N+](=O)c4ccc1ccc2cccc3ccc4c1c23	: 39.63187:	2.170:	3.594:	-1.424:	0/ 55:	12
#:[O-][N+](=O)c1cc2ccc3ccccc3c2cc1	: 31.63350:	1.790:	1.655:	0.135:	0/ 49:	16
#:O=[N+]([O-])c1cc(cc(c1)[N+]([O-])=O)[N+]([O-])=O	: 27.53187:	0.720:	0.660:	0.060:	0/ 63:	20
#:O=[N+]([O-])c1cccc(C)c1[N+]([O-])=O	: 23.90600:	-1.260:	-0.219:	-1.041:	0/ 48:	22
#:Cc1cc(ccc1[N+](=O)[O-])[N+]([O-])=O	: 23.90600:	-0.630:	-0.219:	-0.411:	0/ 48:	24
#:O=[N+]([O-])c1cc(C)ccc1[N+]([O-])=O	: 23.90600:	-1.300:	-0.219:	-1.081:	0/ 48:	26
#:O=[N+]([O-])c1c(c(C)ccc1[N+]([O-])=O)[N+]([O-])=O	: 25.83275:	0.080:	0.248:	-0.168:	0/ 65:	28
#:O=[N+]([O-])c1ccc(c(C)c1[N+]([O-])=O)[N+]([O-])=O	: 25.83275:	0.550:	0.248:	0.302:	0/ 65:	30
#:Cc1c(cc(cc1[N+](=O)[O-])[N+]([O-])=O)[N+]([O-])=O	: 25.83275:	0.160:	0.248:	-0.088:	0/ 65:	32
#:[O-][N+](=O)c2ccc1ccccc1c2C	: 24.86788:	0.080:	0.014:	0.066:	0/ 41:	34
#:[O-][N+](=O)c1cccc2cccc([N+]([O-])=O)c12	: 28.49375:	1.120:	0.893:	0.227:	0/ 56:	38
#:[O-][N+](=O)c1cc2ccc3cccc4ccc(c1)c2c34	: 36.60825:	2.870:	2.861:	0.009:	0/ 57:	40
#:[O-][N+](=O)c1ccccc2ncccc12	: 18.68800:	-0.700:	-1.485:	0.785:	0/ 39:	42
#:[O-][N+](=O)c1ccc2c3ccccc3nc2c1	: 23.66275:	-0.300:	-0.278:	-0.022:	0/ 47:	44
#:[O-][N+](=O)c2cccc3nc1ccccc1c23	: 23.66275:	-0.300:	-0.278:	-0.022:	0/ 47:	46

FIGURE 18

3.7. s-Files

This kind of files (FIGURE 19) contain data on correlation weights for each structural attribute which were obtained in several probes of the Monte Carlo optimization. There are four types of the structural attributes (i) promoters of endpoint increase (correlation weights in all probes are positive); (ii) promoter of endpoint decrease (correlation weights in all probes are negative); (iii) undefined (there are positive and negative correlation weights); and (iv) blocked. These details can be useful in searching for mechanistic interpretations for various endpoints. (It is to be noted ”blocked” structural attributes are absent for model shown in FIGURE 19). s-Files contain also distribution of structural attributes in sub-training (NSs), calibration (NSc), and test sets (NSv).
In the case of multiplicative scheme instead of positive and negative, one should read "larger than unit" (>1) and "smaller than unit" (<1), respectively.

```
This file contains the statistical classification of structural attributes (SA)

Hydrogen suppressed graph (HSG) is used in the model
SMILES is used in the model
The classification is follows:
- if SA has CW(SA)>0 in all probes of the Monte Carlo optimization
then the SA is a promoter of the Endpoint increase (List 1)
- if SA has CW(SA)<0 in all probes of the Monte Carlo optimization
then the SA is a promoter of the Endpoint decrease (List 2)
- if SA has CW(SA)>0 together with CW(SA)<0
then the role of SA is undefined (list 3)
- if SA is blocked, i.e., CW(SA)=0
then the SA without of the model (list 4)

Each list is starting by No.=1, the ID is the numbering in total list of attributes.

NSs, NSc, and NVs are numbers of SMILES which contain SA
in subtraining, calibration, and validation sets, respectively

  No. :  ID : SAk         : CWs Probe 1: CWs Probe 2: CWs Probe 3:        NSs :        NSc :        NSv :
     1:    4:1...........:     9.08875:    13.37800:     6.15225:         14:          14:          20:
     2:    8:=...........:     1.71494:     1.24019:     1.29387:         14:          14:          20:
     3:   12:EC0-C...3...:     1.92588:     1.53325:     1.80369:         14:          14:          20:
     4:   14:EC0-N...3...:     0.63663:     1.05769:     0.21475:         14:          14:          20:   Promoters
     5:   15:EC0-O...1...:     0.61619:     1.17588:     0.47275:         14:          14:          20:   of increase
     6:   17:O...........:     0.95012:     0.33094:     0.10056:         14:          14:          20:
     7:   18:[...........:     0.54288:     0.26862:     0.87400:         14:          14:          20:
     8:    6:3...........:     0.79688:     1.80669:     1.15925:          8:           7:           9:
     9:    7:4...........:     2.35456:     2.85738:     2.36337:          3:           4:           5:
     1:    1:(...........:    -1.51181:    -1.17088:    -1.44913:         14:          14:          20:
     2:   16:N...........:    -0.70713:    -0.08213:    -1.53025:         14:          14:          20:   Promoters
     3:   19:C...........:    -0.28125:    -1.10256:    -0.21594:         14:          14:          20:   of decrease
     4:    9:C...........:    -2.27925:    -2.91788:    -2.33194:          8:           5:           8:
     5:   13:EC0-N...2...:    -3.53525:    -3.98738:    -5.49700:          1:           1:           3:
     6:   20:n...........:    -4.25200:    -2.78906:    -2.56050:          1:           1:           3:
     1:    2:+...........:    -0.41406:    -0.31931:     0.51081:         14:          14:          20:
     2:    3:-...........:    -0.11338:    -0.59656:     0.68250:         14:          14:          20:
     3:   11:EC0-C...2...:     0.37300:     0.72756:    -0.12119:         14:          14:          20:   Undefined
     4:    5:2...........:    -0.29206:     1.00681:     0.19150:         10:          10:          13:
     5:   10:EC0-C...1...:    -0.97275:     0.48156:    -1.36037:          4:           4:           7:
```

FIGURE 19

3.8. w-Files

The containing of w-file (FIGURE 20) may be separated into two sub-sections: (i) numerical data on the correlation weights of structural attributes together with distribution of the attributes into sub-training (NSs), calibration (NSc), and test (NSv) sets; and (ii) lists of not blocked attributes which are absent (a) in the calibration set; (b) in the test set; and (c) in the calibration and in test sets. If some of these categories are absent, then instead of a list will be print of word “empty”.

```
LIST OF STRUCTURAL ATTRIBUTES (SA) AND THEIR CORRELATION WEIGHTS
Hydrogen suppressed graph (HSG) is used in the model
SMILES is used  in the model
ID is number of SA in the completed list of SAs
NSs is the number of SMILES in subtraining set with SA
NSc is the number of SMILES in calibration set with SA
NVs is the number of SMILES in validation set with SA
| SAk         : CW(SAk)    :         ID  :        NSs :        NSc :        NSv :
(...........:    -0.44050:          1:         14:         14:         20:
+...........:     0.05000:          2:         14:         14:         20:
-...........:    -0.04700:          3:         14:         14:         20:
1...(.......:    -1.00500:          4:          1:          3:          3:
1...........:    -0.99500:          5:         14:         14:         20:
2...(.......:     2.62100:          6:          4:          2:          0:
2...........:     0.33850:          7:         10:         10:         13:
2...1.......:    -0.10200:          8:          3:          3:          2:
3...(.......:    -1.00400:          9:          2:          1:          1:
3...........:     0.58350:         10:          8:          7:          9:
3...2.......:     4.89700:         11:          2:          2:          5:
4...........:     2.65400:         12:          3:          4:          5:
4...3.......:     2.30200:         13:          1:          1:          1:
=...(.......:    -0.80200:         14:         14:         14:         20:
=...........:     1.35000:         15:         14:         14:         20:
=...2.......:     0.79800:         16:          1:          0:          0:
C...(.......:     3.84700:         17:          3:          2:          4:
C...........:    -0.99500:         18:          8:          5:          8:
C...1.......:     1.09900:         19:          2:          1:          0:
C...2.......:     1.00100:         20:          2:          0:          2:
C...3.......:     2.84900:         21:          1:          1:          0:
C...=.......:     1.75100:         22:          1:          0:          0:
EC0-C...1...:    -1.00000:         23:          4:          4:          7:
EC0-C...2...:     0.39500:         24:         14:         14:         20:
EC0-C...3...:     0.49500:         25:         14:         14:         20:
EC0-N...2...:    -1.00300:         26:          1:          1:          3:
EC0-N...3...:     0.30925:         27:         14:         14:         20:
EC0-O...1...:     1.13606:         28:         14:         14:         20:
N...+.......:     0.22075:         29:         14:         14:         20:
N...........:     0.15525:         30:         14:         14:         20:
O...(.......:    -0.10200:         31:         14:         12:         18:
O...-.......:     0.15000:         32:         14:         14:         20:
O...........:     0.48238:         33:         14:         14:         20:
O...=.......:     1.73350:         34:         14:         14:         20:
[...(.......:     0.13138:         35:         14:         14:         20:
[...+.......:    -0.26375:         36:         14:         14:         20:
[...-.......:    -0.56550:         37:         14:         14:         20:
[...........:     0.17813:         38:         14:         14:         20:
[...1.......:     2.60400:         39:          3:          3:          6:
[...4.......:     4.19900:         40:          1:          0:          1:
[...=.......:     4.09800:         41:          2:          3:          5:
[...N.......:    -0.68000:         42:         14:         14:         20:
[...O.......:     0.50737:         43:         14:         14:         20:
[...[.......:     0.49500:         44:         10:         10:         13:
c...(.......:    -0.75300:         45:         14:         14:         20:
c...........:     0.19800:         46:         14:         14:         20:
c...1.......:     1.10000:         47:         14:         14:         20:
c...2.......:     2.16375:         48:         10:         10:         13:
c...3.......:     1.64375:         49:          8:          7:          9:
c...4.......:     2.94900:         50:          3:          4:          5:
c...C.......:    -1.00200:         51:          4:          2:          3:
c...c.......:     0.49700:         52:         14:         14:         20:
n...........:    -0.99600:         53:          1:          1:          3:
n...2.......:    -1.00300:         54:          1:          0:          1:
n...3.......:         0.0:         55:          0:          1:          2:
n...c.......:    -1.00200:         56:          1:          1:          3:
Threshold=1 Number of SMILES Attributes(SA)=56 Number of active SA=55

* List of attributes which are absent in calibration set and not blocked:
   1 =...2.......
   2 C...2.......
   3 C...=.......
   4 [...4.......
   5 n...2.......

* List of attributes which are absent in test set and not blocked:
   1 2...(.......
   2 =...2.......
   3 C...1.......
   4 C...3.......
   5 C...=.......

* List of attributes which are absent in calibration and test sets and not blocked:
   1 =...2.......
   2 C...=.......
```

FIGURE 20

Step 4. Checking of the model that is calculated with T* and N*

If preferable threshold is zero and preferable N_{epoch} is 7 (FIGURE 12) the checking of the model should be done according to the scheme:
1. N_{epoch} should be defined 7: after click of “Load method” (FIGURE 7) define N_{epoch} equal to 7 and click “Save method”. (FIGURE 21)

IMPORTANT:
When you have defined your method YOU MUST SAVE the method by clicking of button ‘Save method’ otherwise, the system will be working according to options of the previous method version.

2. Click button “Building up preferable model (T*,N*) (FIGURE 22)
3. Insert preferable threshold equal to 1; click button “Continue” (FIGURE 23).
4. Confirm that files in folder "Model" can be deleted or click “No” (FIGURE 24).

FIGURE 25 shows the status of the system after these actions.

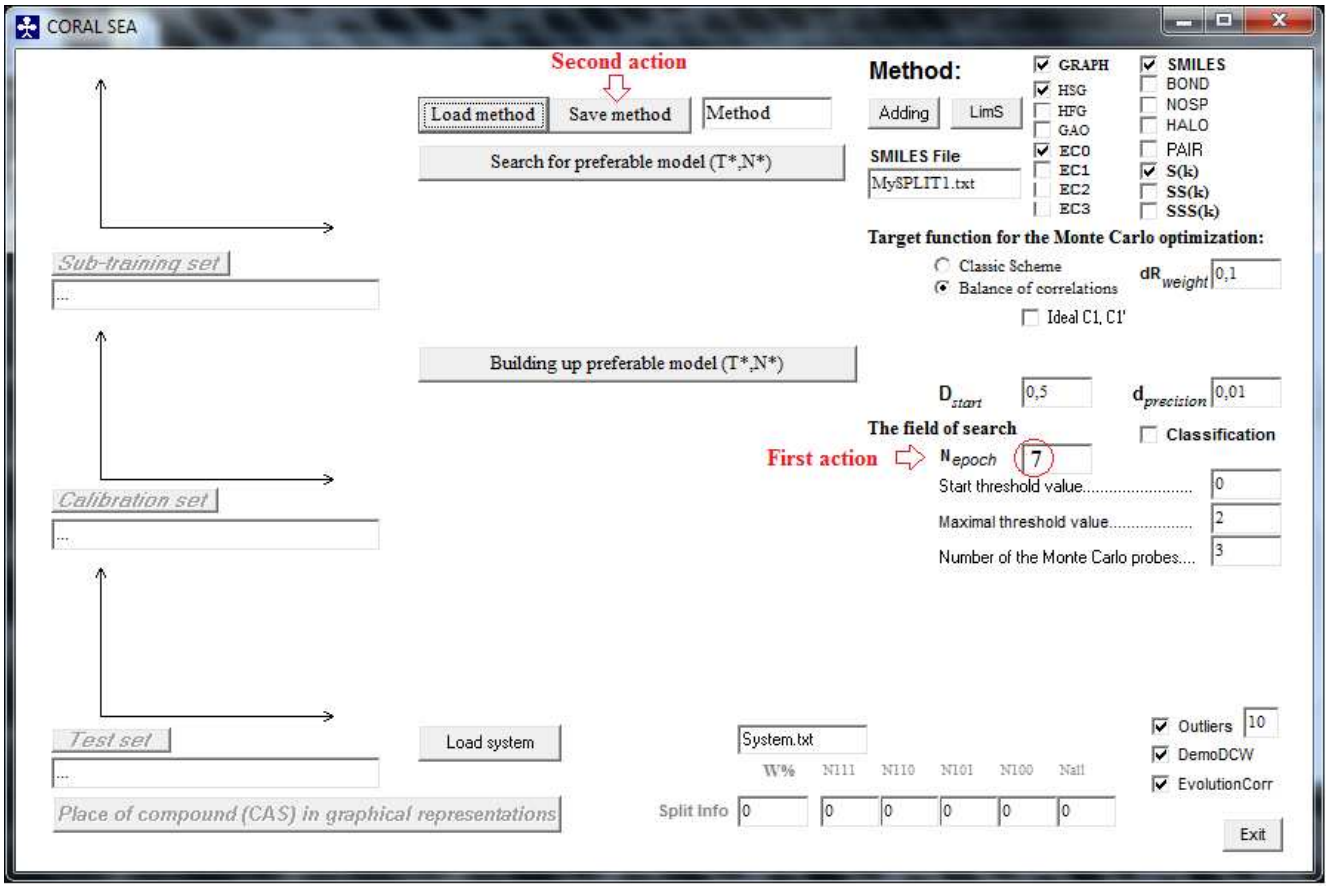


FIGURE 21

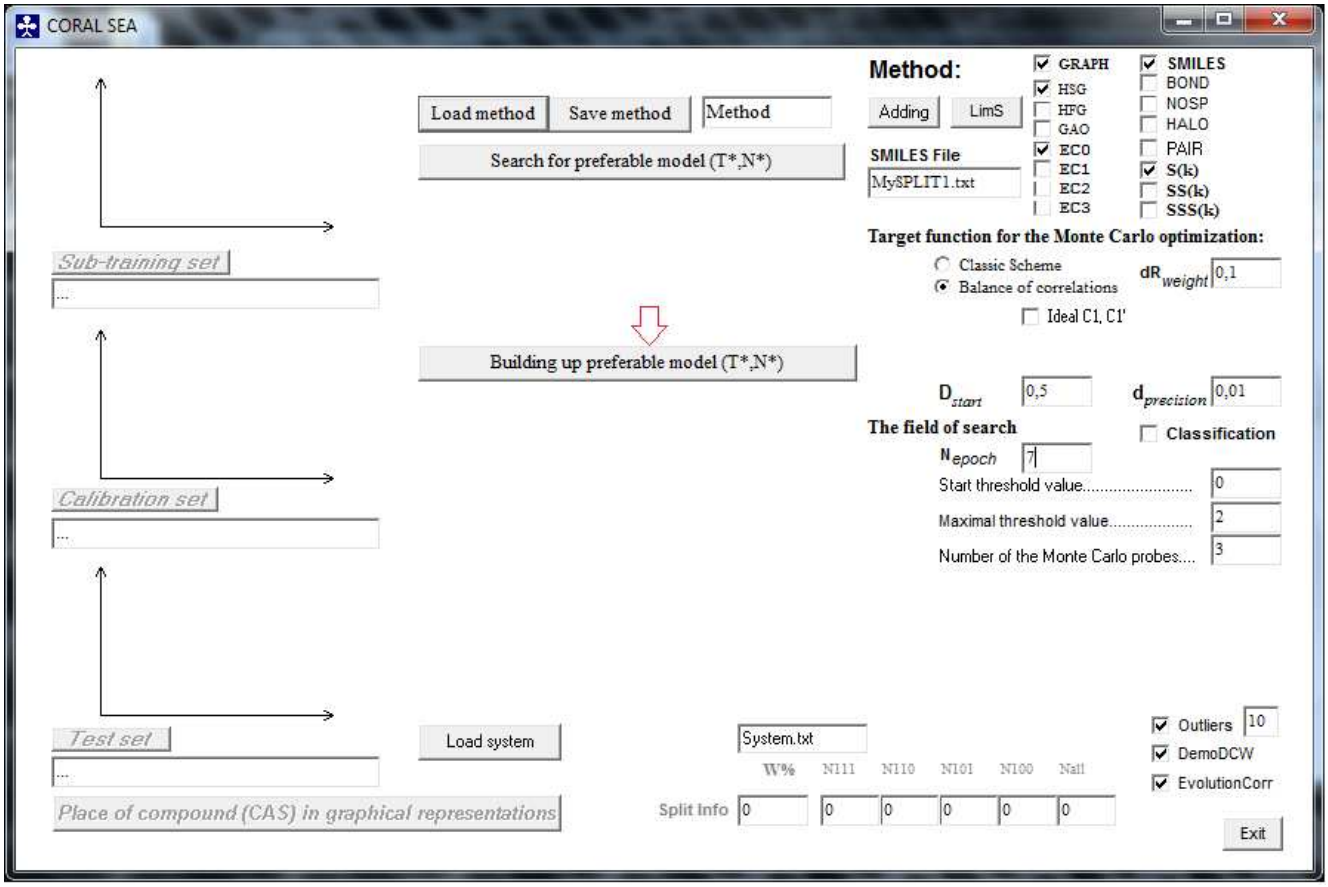


FIGURE 22

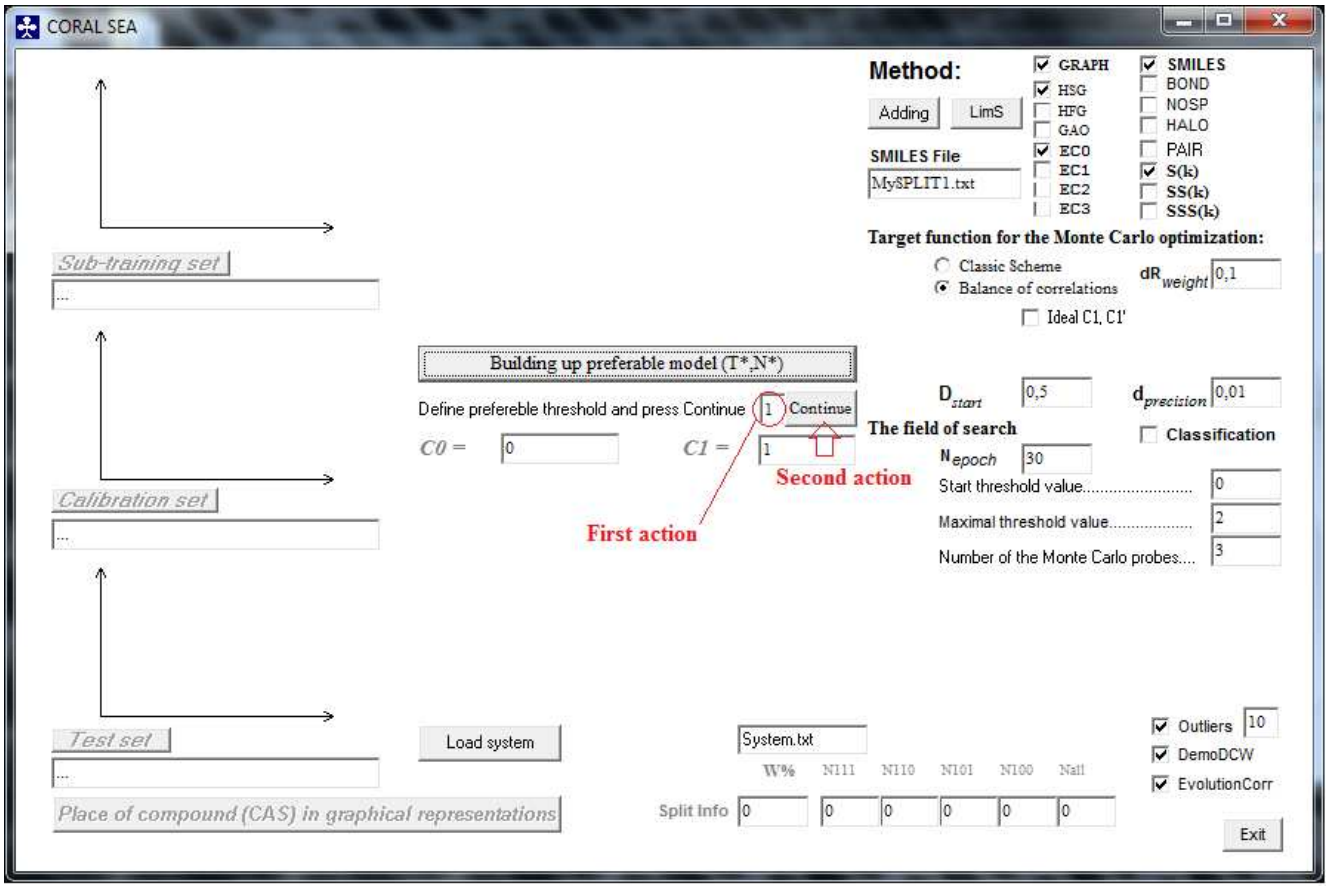


FIGURE 23

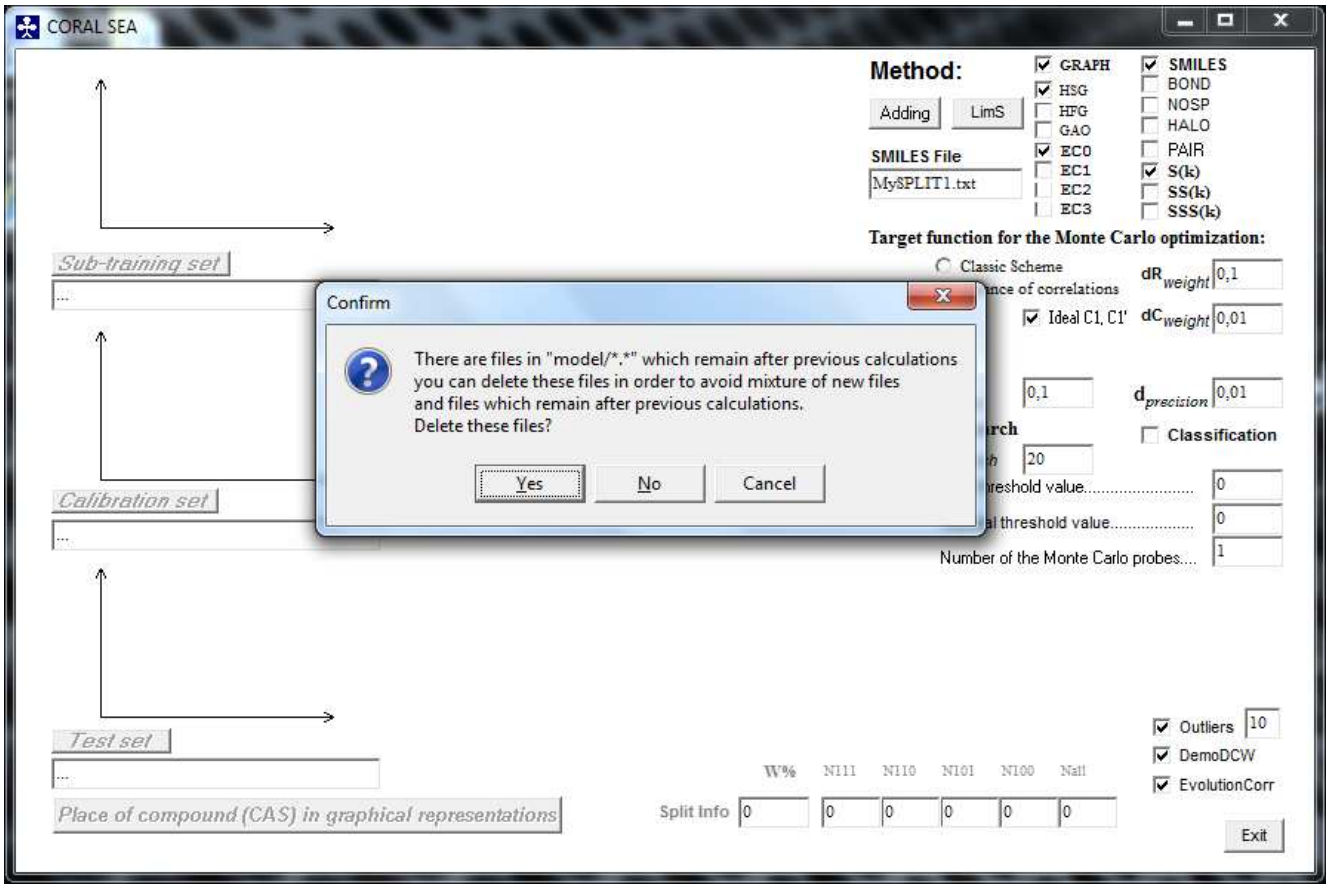


FIGURE 24

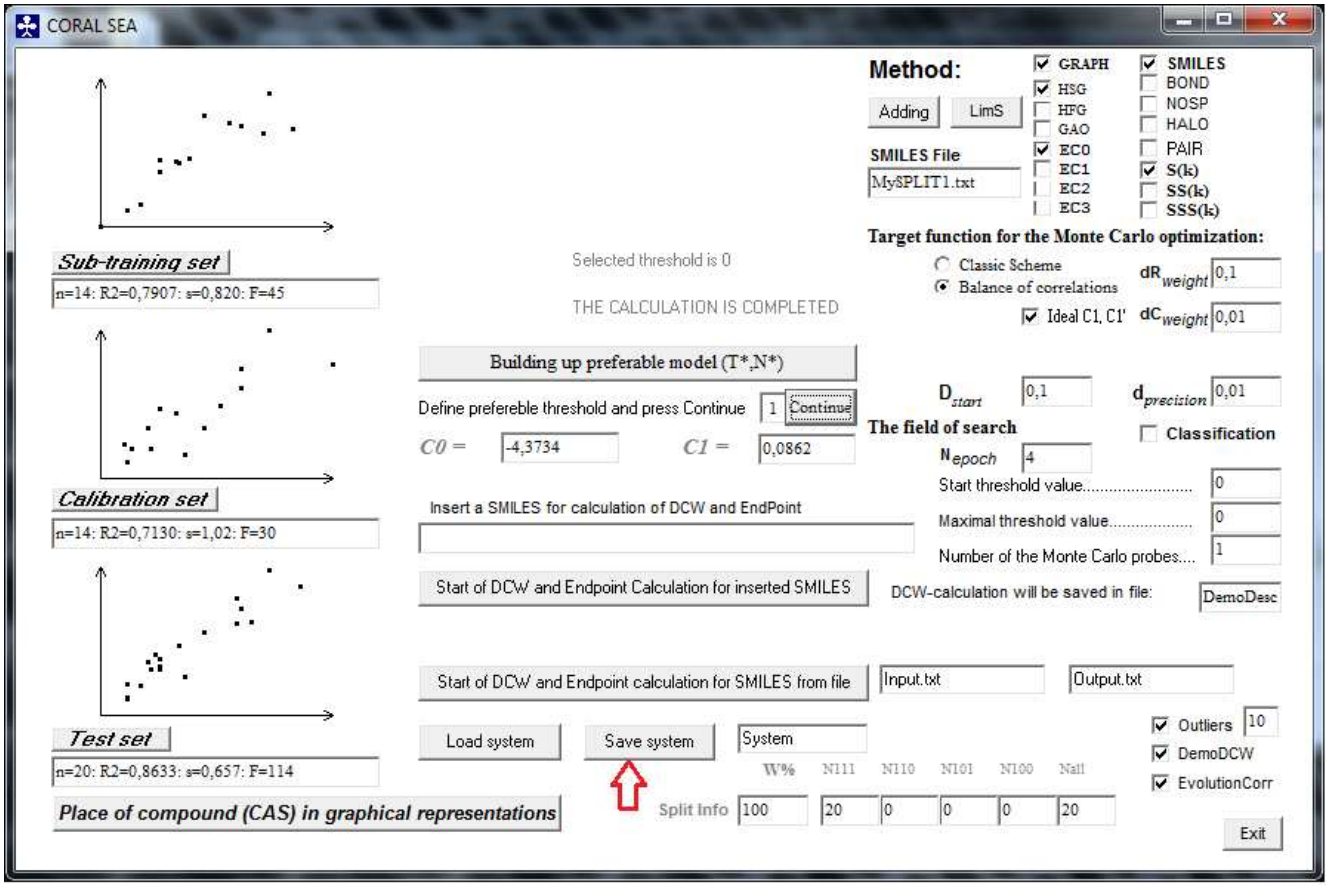


FIGURE 25

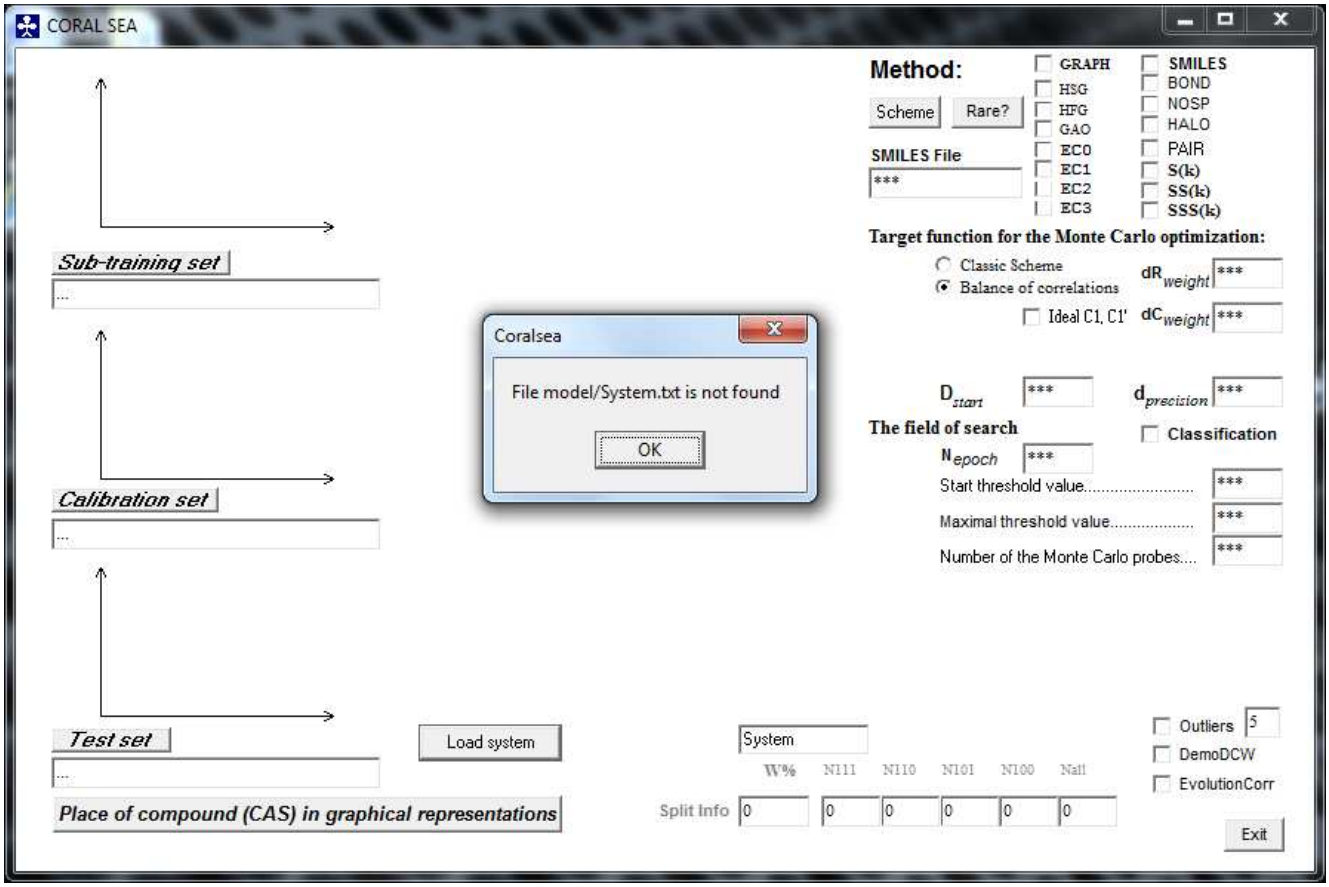
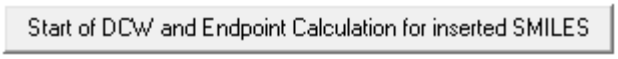


FIGURE 26

4.1. Calculation of the model for sole substance (SMILES)

You can save the model by click of the button “Save system” (FIGURE 25). It gives possibility to use this model in the future. For instance, after future start of the program, you can click of button “Load system” (FIGURE 26) and (if file “system” is not deleted in folder "Model"!) you will see again the picture that is shown in FIGURE 25. If file “system” is deleted you will see picture that is shown in FIGURE 26.

If file “System” available in folder “Model” and you have downloaded the file “System”, you can insert some SMILES into box “Insert a SMILES for calculation of DCW and endpoint” and click button



See FIGURE 27.

Documentation of this calculation one can read in file “DemoDCW.txt” in folder “Model”. Format of the “DemoDCW.txt” is identical to d-File (FIGURE 13).

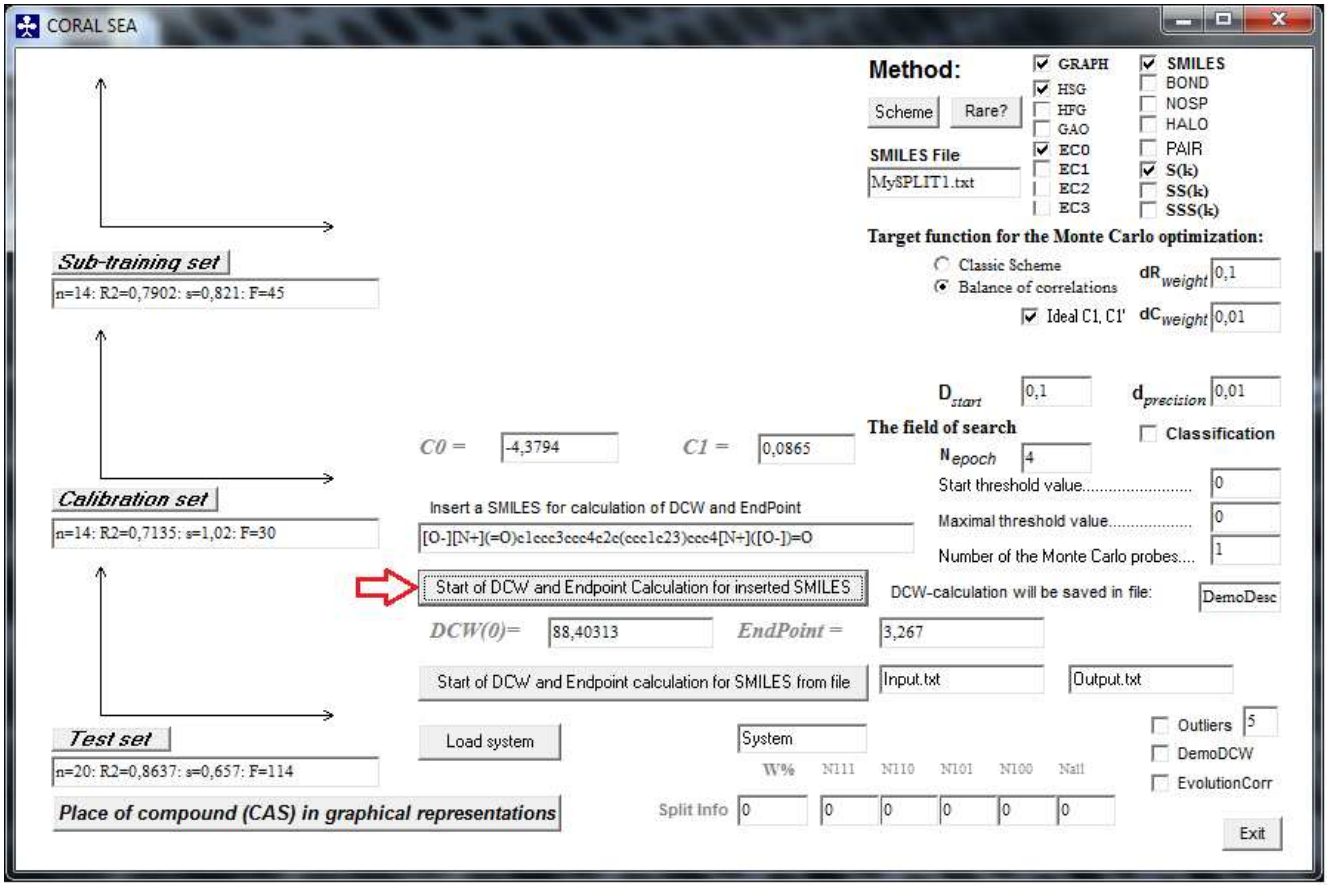


FIGURE 27

4.2. Calculation of the model for a group of substance (SMILES)

You can to carry out calculation of model for a group of substances if prepare input.txt file that is organized as the following:

N // the number of substances (i.e. SMILES strings)
ID1 SMILES1 ENDPOINT1
ID2 SMILES2 ENDPOINT2
...
IDN SMILESN ENDPOINTN

For example, if there are some preliminary data on the endpoint the file " input.txt" can be the following (version 1, regression model)

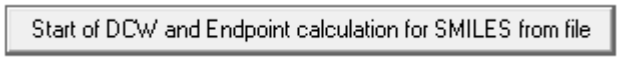
```
5
+4 [O-][N+](=O)c1ccc3ccc4c2c(ccc1c23)ccc4[N+]([O-])=O 4.09
+6 [O-][N+](=O)c2ccc3c1ccc(cc1C(=O)c3c2)[N+]([O-])=O 2.69
+8 [O-][N+](=O)c1ccc2cc3ccccc3cc2c1 3.05
+11 [O-][N+](=O)c3ccc4c2cccc1cccc(c12)c4c3 2.60
+15 [O-][N+](=O)c1ccc2c3ccccc3Cc2c1 1.08
```

If information on the endpoint is not available the input.txt can be the following (version 2, classification model, e.g. -1, 1; also possible 0, 1, i.e. inactive[-1 or 0] /active [1])

```
5
+4 [O-][N+](=O)c1ccc3ccc4c2c(ccc1c23)ccc4[N+]([O-])=O 1.0
+6 [O-][N+](=O)c2ccc3c1ccc(cc1C(=O)c3c2)[N+]([O-])=O 0.0
+8 [O-][N+](=O)c1ccc2cc3ccccc3cc2c1 1.0
+11 [O-][N+](=O)c3ccc4c2cccc1cccc(c12)c4c3 0.0
+15 [O-][N+](=O)c1ccc2c3ccccc3Cc2c1 0.0
```

There is no limitation for the number of substances in the “input.txt” file. But this file should be placed in the folder where CORALSEA.exe is placed.

If the file “input.txt” is available, then click of button (FIGURE 28)



The program will inform you that results of calculations are saved in file “output.txt” (FIGURE 29).

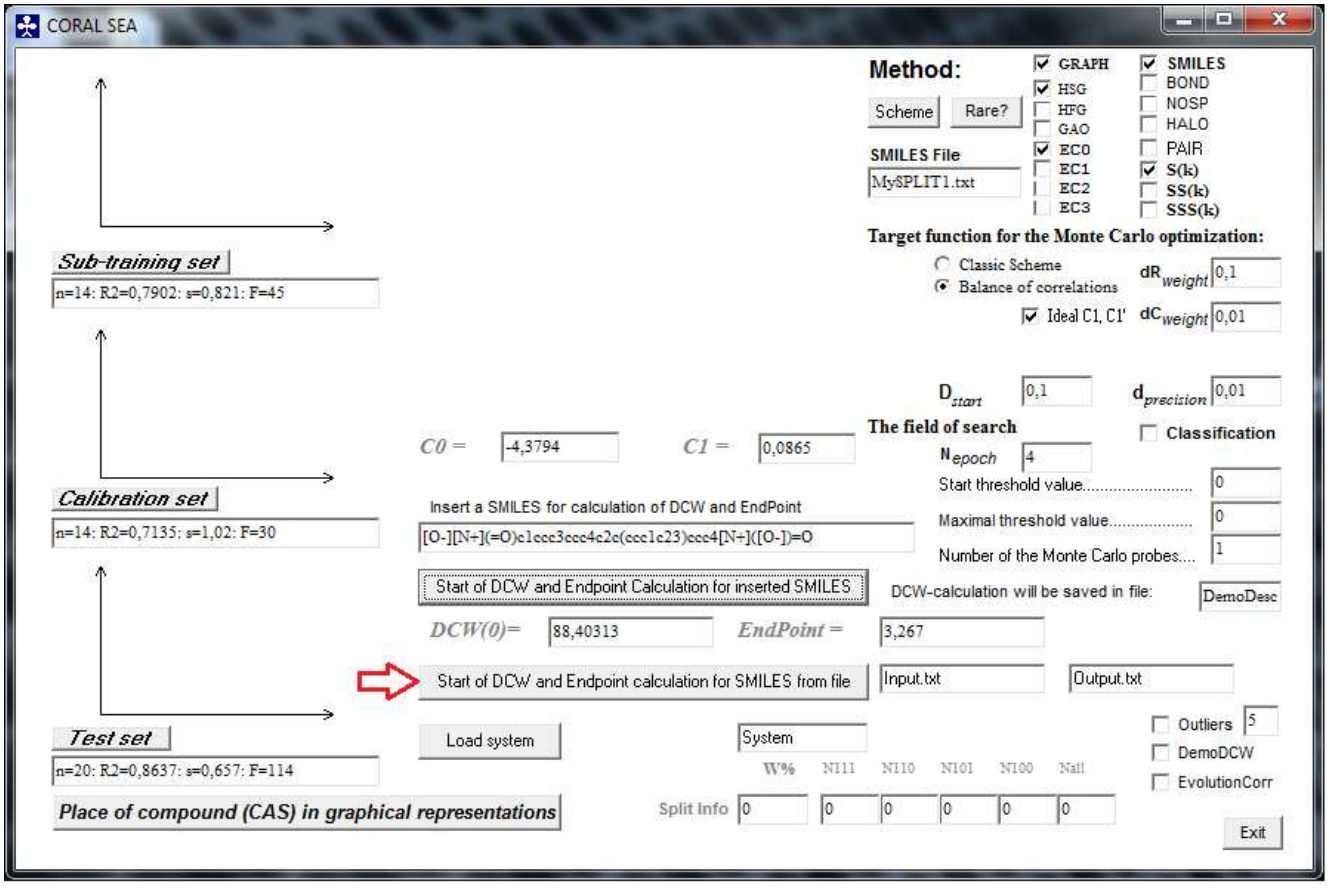


FIGURE 28

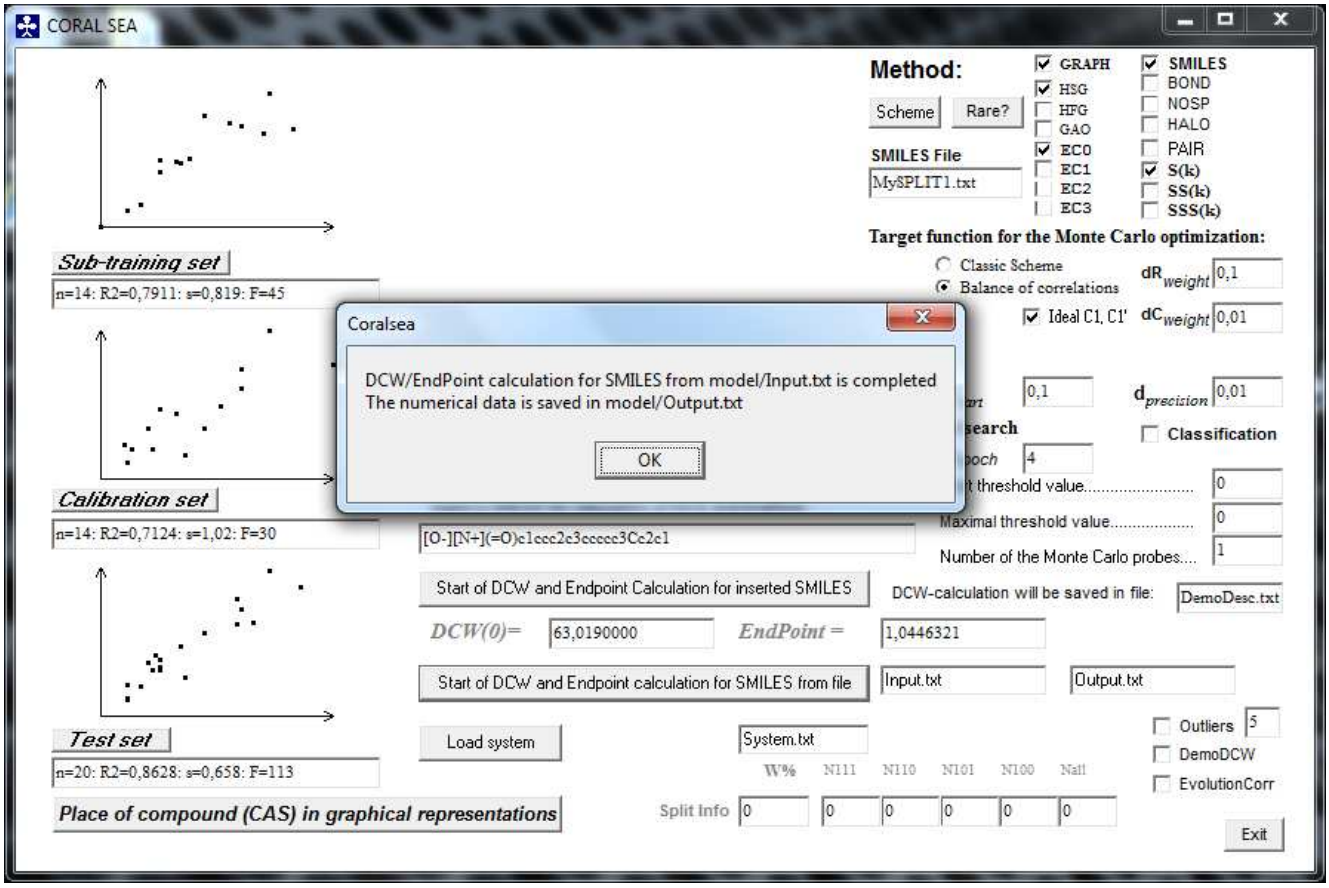


FIGURE 29

The output file that contains data calculated with above file “input.txt” (version 1) may be the following:

ID:SMILES	DCW	Expr	Calc
+4 :[O-][N+](=O)c1ccc3ccc4c2c(ccc1c23)ccc4[N+]([O-])=O	88.81400	4.090	3.260
+6 :[O-][N+](=O)c2ccc3c1ccc(cc1C(=O)c3c2)[N+]([O-])=O	76.21200	2.690	2.178
+8 :[O-][N+](=O)c1ccc2cc3ccccc3cc2c1	67.81200	3.050	1.456
+11 :[O-][N+](=O)c3ccc4c2cccc1cccc(c12)c4c3	79.80600	2.600	2.487
+15 :[O-][N+](=O)c1ccc2c3ccccc3Cc2c1	63.01900	1.080	1.045

The output file that contains data calculated with above file “input.txt” (version 2) may be the following:

ID:SMILES	DCW	Expr	Calc	
+4 :[O-][N+](=O)c1ccc3ccc4c2c(ccc1c23)ccc4[N+]([O-])=O	128.656	1	0.128	FN
+6 :[O-][N+](=O)c2ccc3c1ccc(cc1C(=O)c3c2)[N+]([O-])=O	717.352	0	0.717	FP
+8 :[O-][N+](=O)c1ccc2cc3ccccc3cc2c1	912.121	1	0.912	TP
+11 :[O-][N+](=O)c3ccc4c2cccc1cccc(c12)c4c3	231.766	0	0.231	TN
+15 :[O-][N+](=O)c1ccc2c3ccccc3Cc2c1	111.192	0	0.111	TN

TP true positive; FP false positive; TN true negative; FN false negative

Step 5. Checking of the approach with a few random splits

The statistical characteristics of a CORALSEA model is a mathematical function of many parameters. In particular, the split into sub-training, calibration, and test sets influences the statistical characteristics. Under such circumstances, the analysis of a group of splits becomes important and interesting task.
Since the CORALSEA detects the above-mentioned sets via the first symbols ('+','-','#'), one can prepare a split 2 that is not the same as split 1 by means of the shifting represented in Table 3

Table 3
Possible way to exchange a split 1 by a split 2

Split 1	Split 2
...	...
#276 ClCC(Cl)Cl 3.09	+276 ClCC(Cl)Cl 3.09
+31 CCC(Cl)Cl 3.57	#31 CCC(Cl)Cl 3.57
+282 ClCC(Cl)CCl 3.72	+282 ClCC(Cl)CCl 3.72
+297 Clc1ccc(c(c1)Cl)Cl 4.16	#297 Clc1ccc(c(c1)Cl)Cl 4.16
-223 [O-][N+](=O)c1cccc(c1Cl)Cl 4.62	+223 [O-][N+](=O)c1cccc(c1Cl)Cl 4.62
-281 Clc1ccccc1Cl 4.81	#281 Clc1ccccc1Cl 4.81
#287 ClCCCl 2.29	-287 ClCCCl 2.29
-275 C[C@@H](Cl)CCl 3.34	+275 C[C@@H](Cl)CCl 3.34
#288 OCCO 0.48	+288 OCCO 0.48
-177 [O-][N+](=O)c1cc(cc(c1)Cl)Cl 4.46	-177 [O-][N+](=O)c1cc(cc(c1)Cl)Cl 4.46
+300 Clc1cccc(c1)Cl 4.18	#300 Clc1cccc(c1)Cl 4.18
...	...

Having the split 2 one can repeat the computational experiments in order to answer questions:

- whether the approach is robust for second split?
- whether distributions of structural attributes into sub-training, calibration, and test sets for split 1 and split 2 are equivalent?
- whether there are common outliers for the split 1 and the split 2?

…and maybe for series of other questions.

IMPORTANT
Unfortunately, CORAL can give an unexpected interpretations of a molecular features, e.g. Cs can be recognize as metal cesium or carbon connected to sulphur. In order to avoid such misdetections one should check up lists of SMILES attributes involved in the modeling process. Other possible wrong interpretations can take place for Os, Sn, Co, etc.

Appendix

A1. Places of substances in the diagrams “experiment – calculation”

You can check position of different dots in plot of experiment versus calculated values of an endpoint.
In the case of classic scheme there are two plots (training and test sets).
In the case of balance of correlations there are three plots (sub-training, calibration, and test sets).
When your model is ready, you can select one of the above plots by means of click of a button , e.g. you can select test set (FIGURE 30):

Sub-training set is selected
Calibration set is selected
Test set is selected

Having selected a set (test set), you can click of the following button (FIGURE 31)

After these actions you can check position of different substances in the diagram of experiment versus calculated values of the endpoint, by clicking “yes”, if the sequence of substances is OK or “No” if you would like to change the direction: one possibility from smaller to larger (increase), i.e. #1, #2, #3, …, #7; other possibility from larger to smaller (decrease), i.e. #7, #6, #5, etc. The current substance is indicated by red (FIGURE 32).

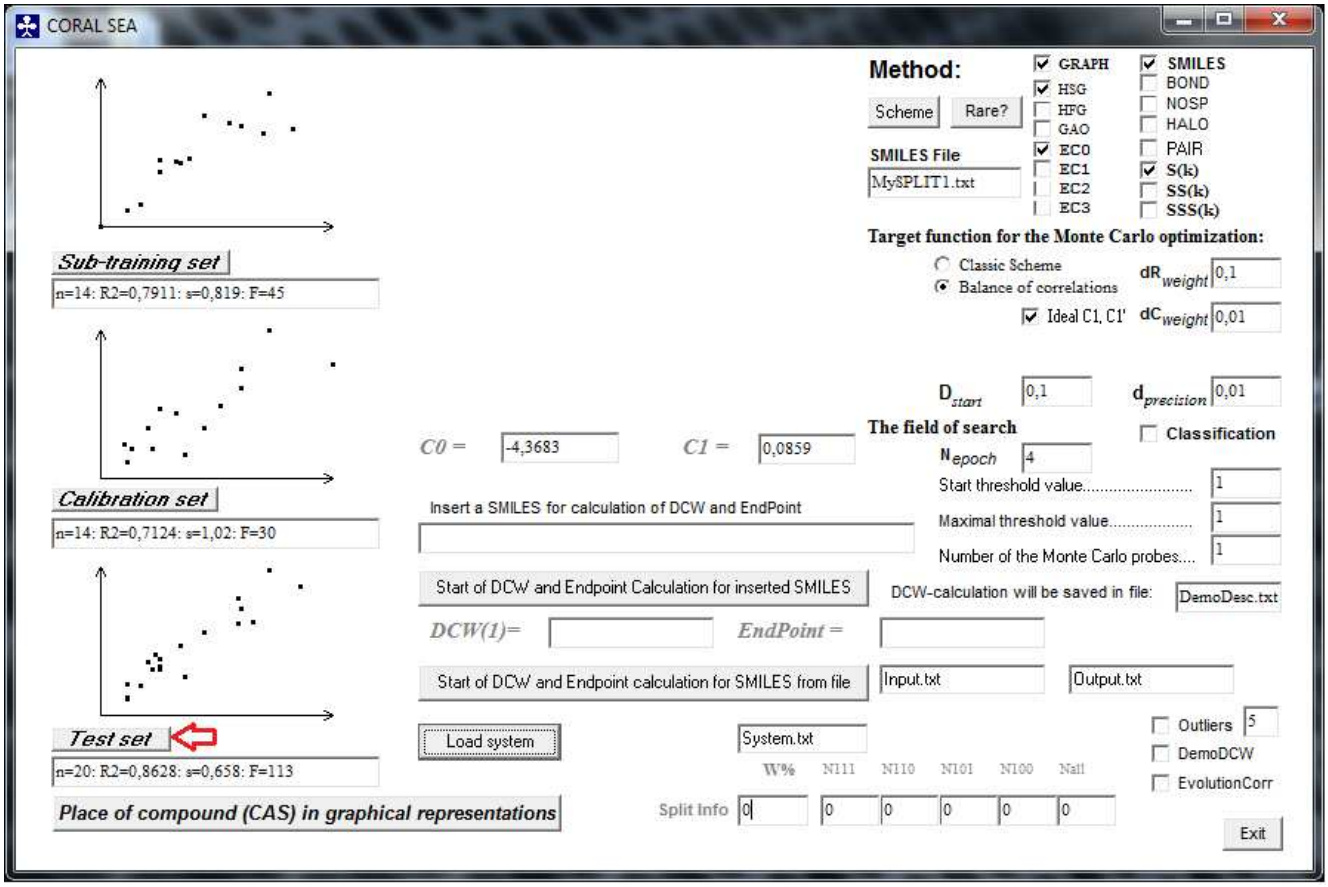


FIGURE 30

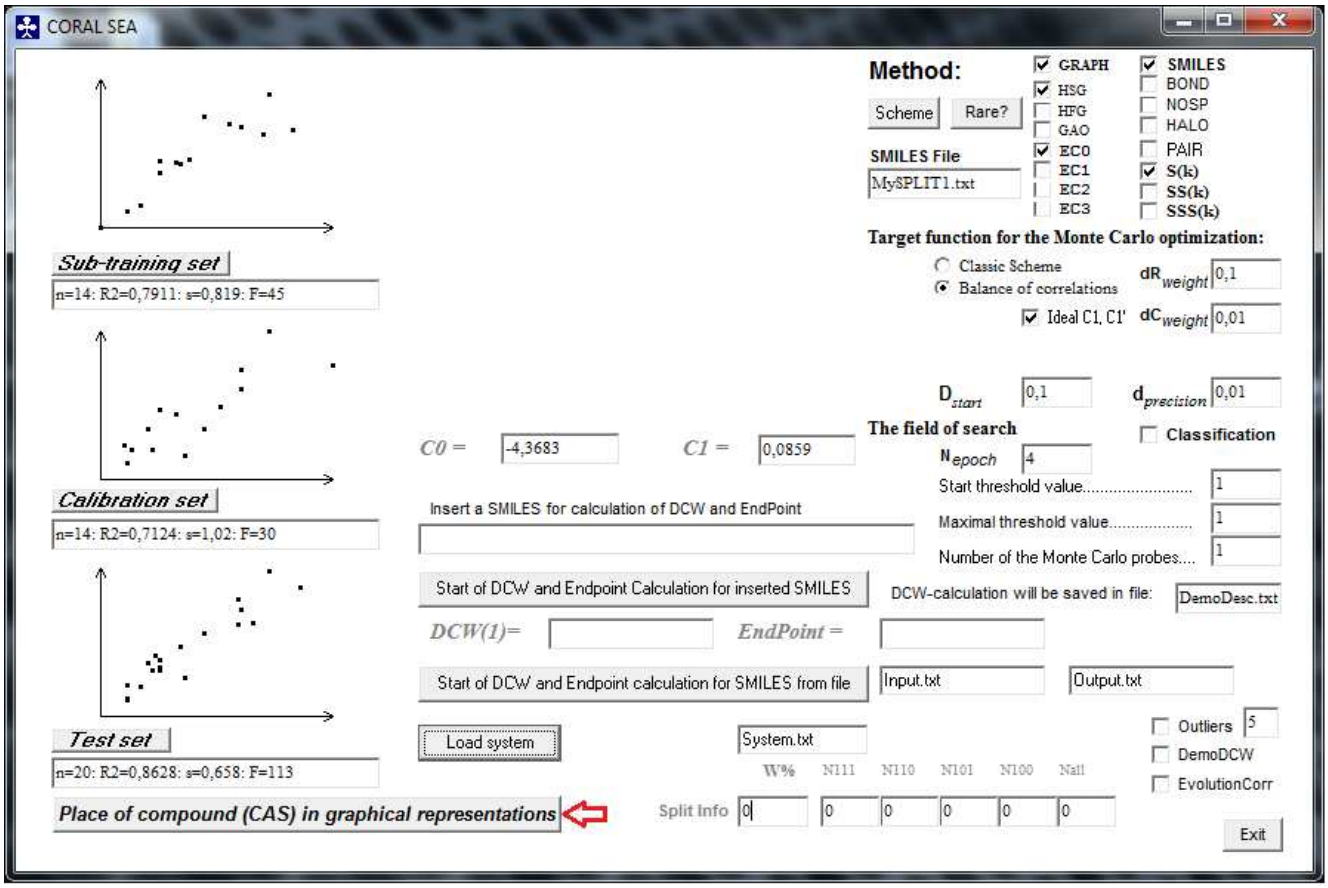


FIGURE 31

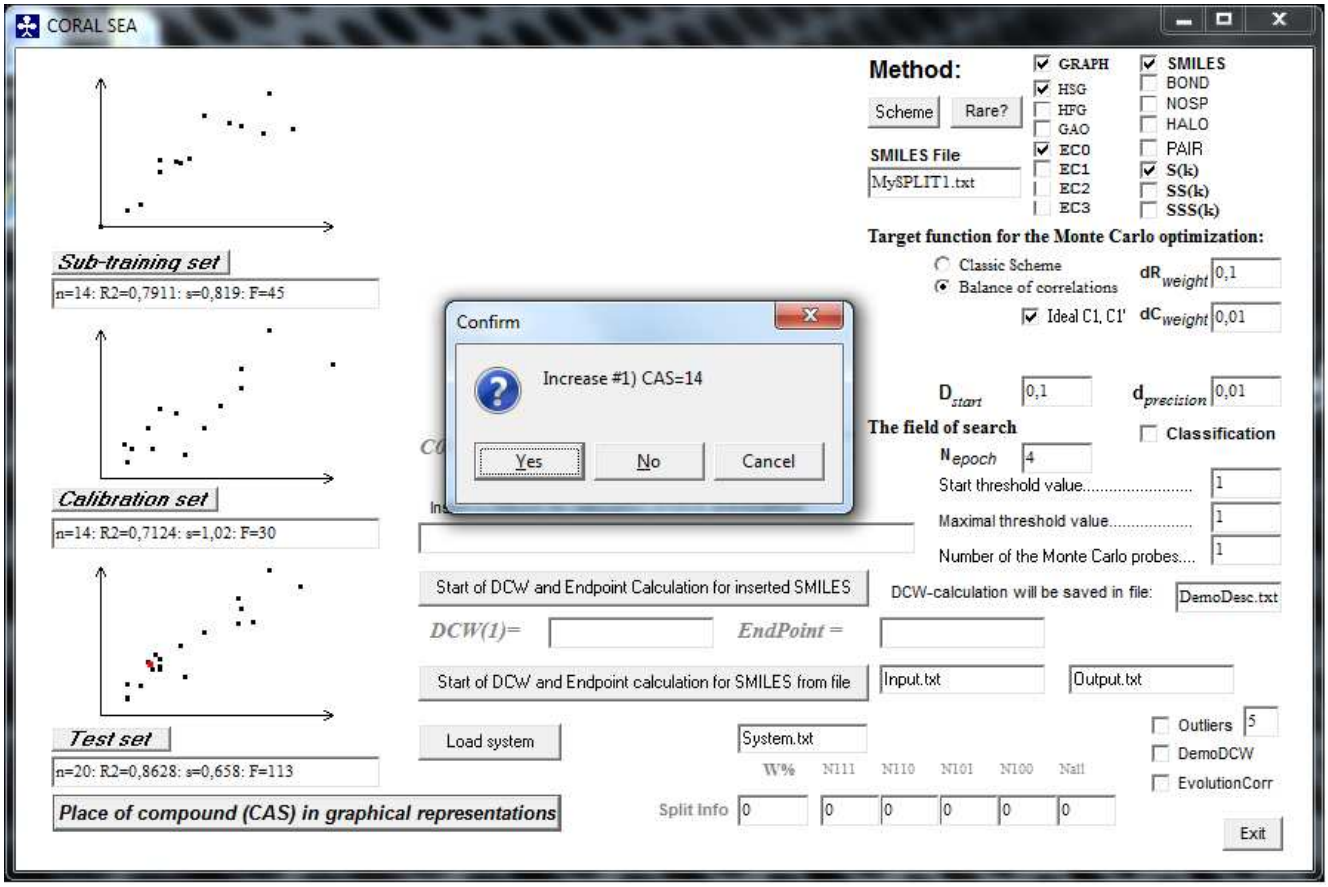

CORAL SEA
Method:
GRAPH
HSG
HFG
GAO
SMILES
BOND
NOSP
HALO
PAIR
S(k)
SS(k)
SSS(k)
Scheme
Rare?
SMILES File
MySPLIT1.txt
EC0
EC1
EC2
EC3
Target function for the Monte Carlo optimization:
Classic Scheme
Balance of correlations
Ideal C1, C1'
dR weight 0,1
dC weight 0,01
D start 0,1
d precision 0,01
The field of search
Classification
N epoch 4
Start threshold value..... 1
Maximal threshold value..... 1
Number of the Monte Carlo probes.... 1
Sub-training set
n=14: R2=0,7911: s=0,819: F=45
Calibration set
n=14: R2=0,7124: s=1,02: F=30
Test set
n=20: R2=0,8628: s=0,658: F=113
Confirm
Increase #1) CAS=14
Yes
No
Cancel
Start of DCW and Endpoint Calculation for inserted SMILES
DCW-calculation will be saved in file:
DemoDesc.txt
DCW(1)=
EndPoint =
Start of DCW and Endpoint calculation for SMILES from file
Input.txt
Output.txt
Load system
System.txt
Outliers
5
DemoDCW
EvolutionCorr
W%
N111
N110
N101
N100
Nall
Split Info
0
Place of compound (CAS) in graphical representations
Exit

FIGURE 32

A2. Classification model

If there are data on some activity in qualitative form, i.e. as active /inactive data, these data can be expressed as -1/1 data (or 0/1).

For this situation the CORALSEA provides a model that can be represented graphically as

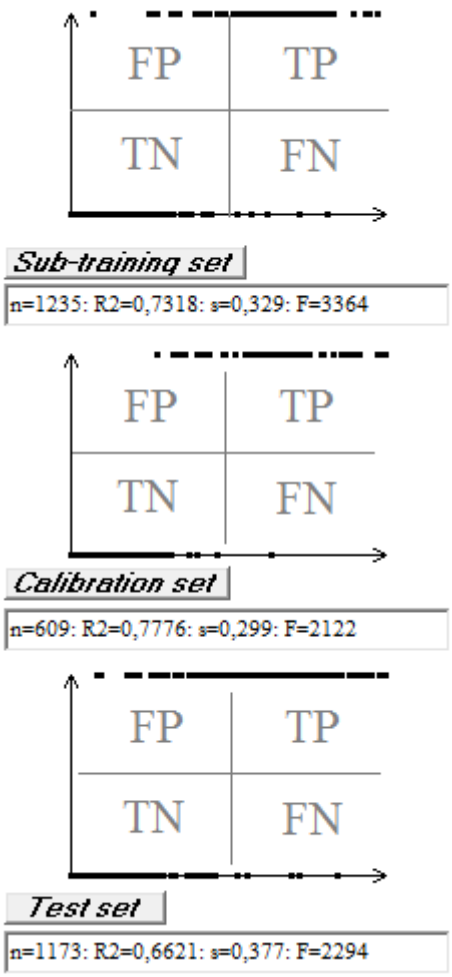


FIGURE 33

Measure of statistical quality of this model is expressed by

Sensitivity = TP / (TP+FN) (6)
Specificity = TN / (TN+FP) (7)
Accuracy = (TP+TN) / (TP+FP+TN+FN) (8)
MCC) = (TP*TN - FP*FN)/sqrt((TP+FP)(TP+FN)(TN+FP)(TN+FN)) (9)

MCC is Matthews correlation coefficient.
One can activate this kind of models by selecting of box ☑ Classification (see page 13).

In the case of classification model m-files have two changes in comparison with above description (page 23, FIGURE 16).
(i) The following addition (after delta R_m^2)

```
Subtraining set:
TP=   26 TN=  195 FP=   21 FN=   24 N=  266
Sensitivity= 0.5200
Specificity= 0.9028
Accuracy   = 0.8308
MCC        = 0.4331

Calibration set:
TP=   24 TN=  140 FP=    9 FN=   21 N=  194
Sensitivity= 0.5333
Specificity= 0.9396
Accuracy   = 0.8454
MCC        = 0.5313

Test set:
TP=   15 TN=  139 FP=   21 FN=   22 N=  197
Sensitivity= 0.4054
Specificity= 0.8688
Accuracy   = 0.7817
MCC        = 0.2771
```

(ii) The scheme of the representation of the classification model can be expressed as the following

```
:         DCW:        Expr:        Calc:   Expr-Calc Blk/All:
:    21.33650:           1:       0.820:        TP    0/ 85:
:    19.12750:           1:       0.687:        TP    0/ 43:
:    23.14350:           1:      -0.856:        FN    0/ 61:
:     7.89500:           1:       0.006:        TP    0/105:
:    19.00100:           1:       0.679:        TP    0/ 63:
:    30.48500:           1:      -0.804:        FN    0/ 73:
:    12.00550:           1:       0.303:        TP    0/ 31:
:    14.35300:           1:       0.505:        TP    0/ 57:
:    23.84900:           1:       0.973:        TP    0/ 49:
:    27.77950:           1:       1.211:        TP    0/ 33:
:    16.16500:           1:      -0.863:        FN    0/ 35:
:     4.86550:           1:      -0.919:        FN    0/ 37:
:    24.56250:           1:       1.016:        TP    0/ 57:
:    22.58450:           1:      -1.006:        FN    0/ 63:
:    34.08850:           1:       1.593:        TP    0/ 53:
:    47.81200:           1:       2.425:        TP    0/ 89:
:    24.94950:           1:      -1.136:        FN    0/ 47:
:    31.40850:           1:       1.431:        TP    0/ 61:
:     7.78850:           1:      -1.030:        FN    0/ 49:
:    22.81150:           1:      -0.897:        FN    0/ 51:
:     6.32700:           1:      -1.146:        FN    0/ 43:
:     2.36550:          -1:      -0.329:        TN    0/ 35:
:    -7.86800:          -1:      -0.944:        TN    0/ 67:
:    -4.76400:          -1:      -0.761:        TN    0/ 57:
:     2.76250:          -1:       1.166:        FP    0/ 41:
:    -9.52000:          -1:      -0.843:        TN    0/ 89:
:    -6.68300:          -1:      -1.012:        TN    0/ 25:
:    -5.80350:          -1:      -0.824:        TN    0/111:
:    -8.66000:          -1:      -0.997:        TN    0/337:
:   -10.94550:          -1:       1.039:        FP    0/ 67:
:    -8.62100:          -1:      -0.995:        TN    0/103:
:    -9.08900:          -1:      -0.677:        TN    0/185:
       ...              ...         ...         ...      ...
```

where TP, TN, FP, and FN are quality of the prediction i.e. true positive, true negative, false positive, and false negative, respectively (FIGURE 33).

A3. Split Information

The split into sub-training, calibration, and test sets is important fragment of the QSPR/QSAR analyses.

In the case of building up QSPR/QSAR by the CORALSEA software there are possibility to compare various splits as well as methods via criterion denoted as W% (work percentage).

$$W\% = \frac{N_{111}}{N_{ALL}} \quad (9)$$

where N_{All} is the total number of structural attributes which are involved in the modeling process (i.e. which are not blocked); and N_{111} is the number of structural attributes which are taking place in all sets, i.e. which are taking place in sub-training, calibration, and test sets.

In the case of "classic" scheme

$$W\% = \frac{N_{101}}{N_{ALL}} \quad (10)$$

where N_{101} is the number of structural attributes which are taking place in training, and test sets.

Various method are characterized by different values of W% (compare FIGURE 34 and FIGURE 35).

The threshold also influences W%.

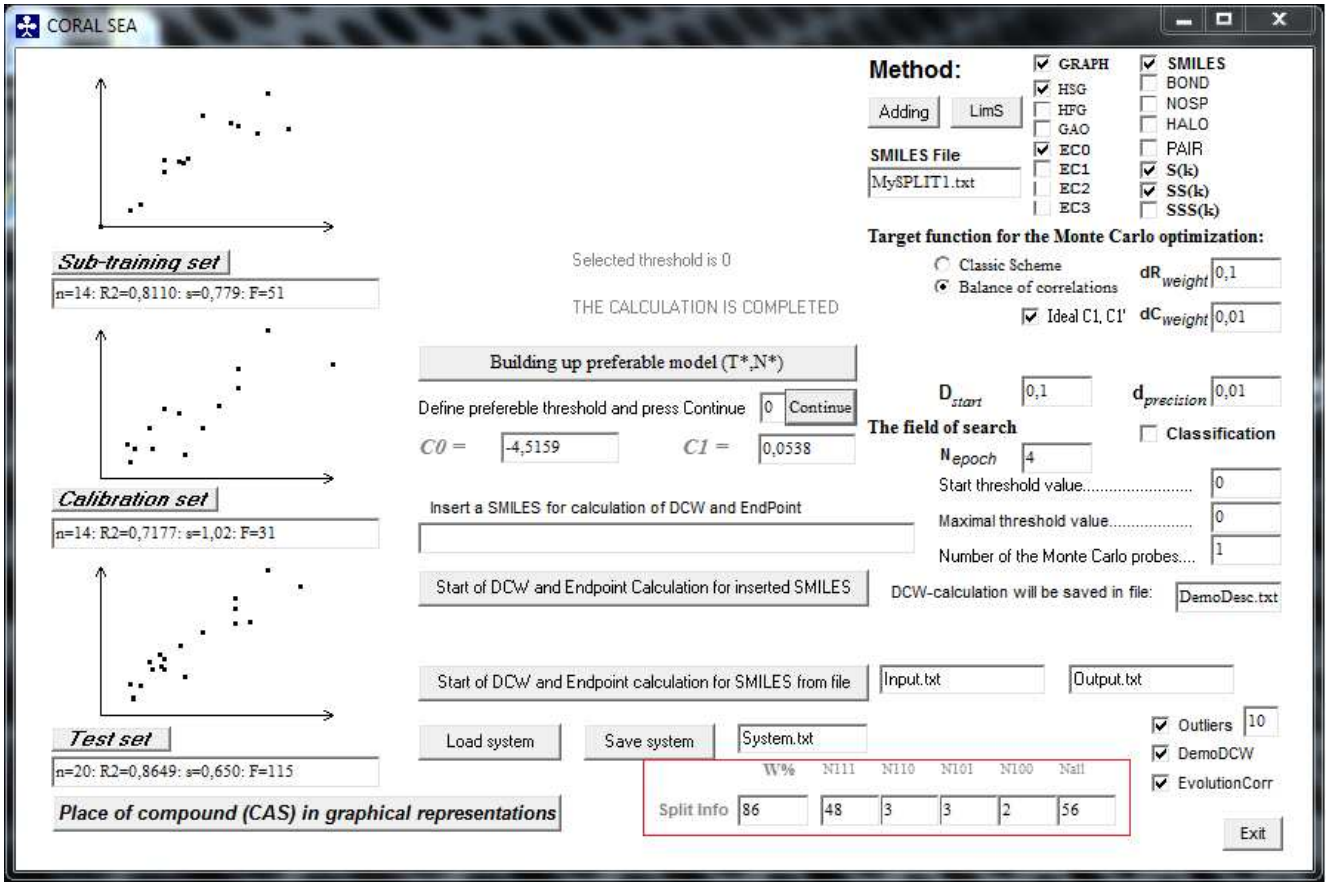


FIGURE 34

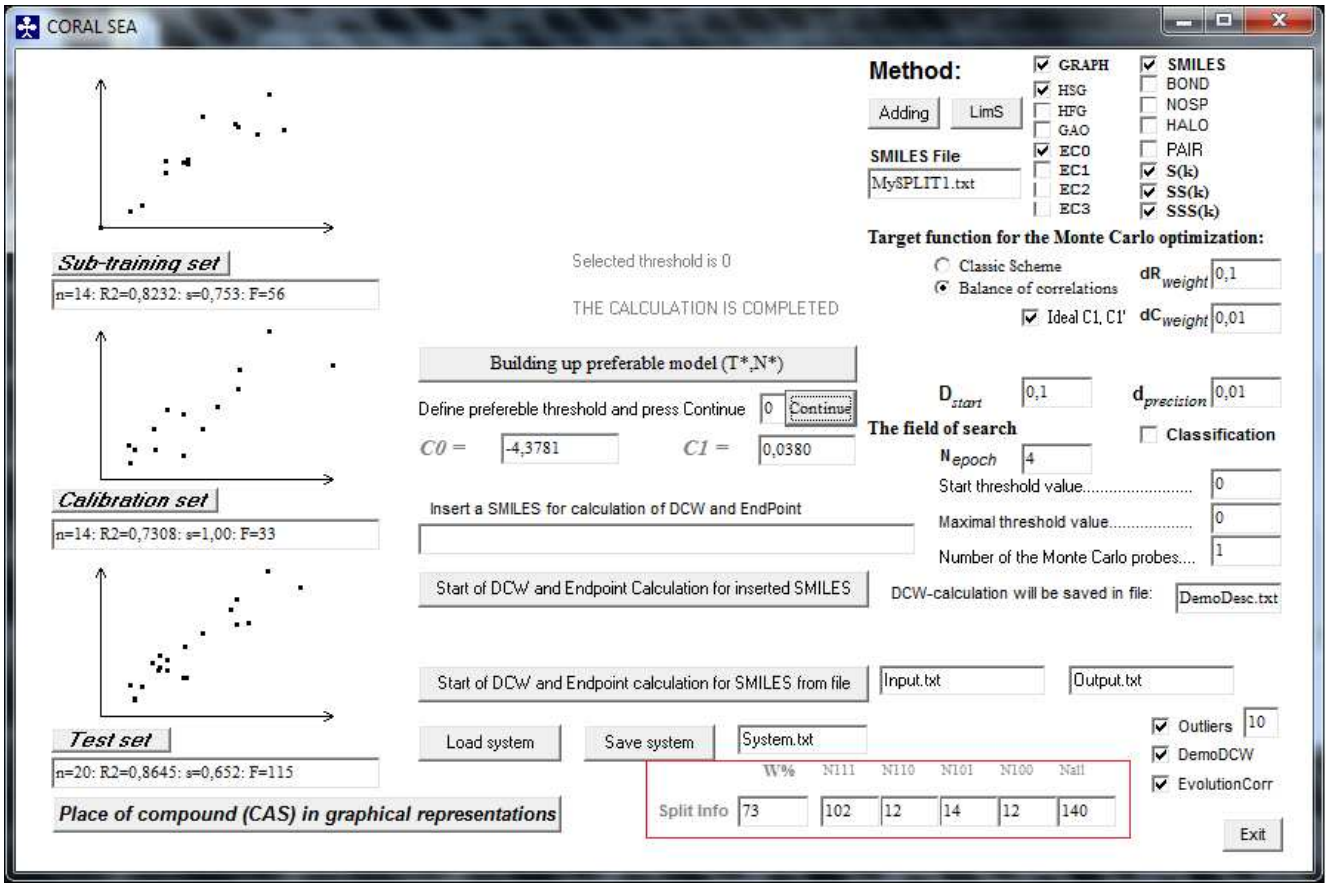


FIGURE 35

A4. Sketch of praxis

1. Molecular structure of the majority of substances can be represented by SMILES.
2. SMILES is provider of molecular attributes which are representing *local* and *global* molecular features.
3. The building up of QSPR/QSAR model for an arbitrary split into the training and test sets should be qualified as a random event.
4. The statistical quality of each QSPR/QSAR model is a mathematical function of split into the training and test sets.
5. The average statistical quality of QSPR/QSAR models that is obtained for several splits into training and test sets is more robust criterion for the estimation of an approach than statistical quality for solely one split.
6. The average statistical quality of a models *for external test sets* is more significant data than the average statistical quality for training sets.
7. The correlation weights for molecular features (which are extracted from graph and/or SMILES) can be used for classification of the above-mentioned features according to their values for several models into three categories: features with stable positive values of correlation weights (promoters of increase for an endpoint); features with stable negative values of correlation weights (promoters of decrease of an endpoint); and undefined features which have positive values of correlation weights together with negative correlation weights values for series of runs of the Monte Carlo optimization.
8. Data on the correlation weights for molecular features which are calculated with graph and/or SMILES (which are promoters of increase of an endpoint and promoters of its decrease) give possibility to define the applicability domain (a set of compounds): ideal applicability domain is a set of compounds which have not molecular features with undefined role (which are not stable promoters of increase or decrease of endpoint).
9. Most simple method as rule gives models with highest predictive potential.

A5. Semi-Optimal Descriptors

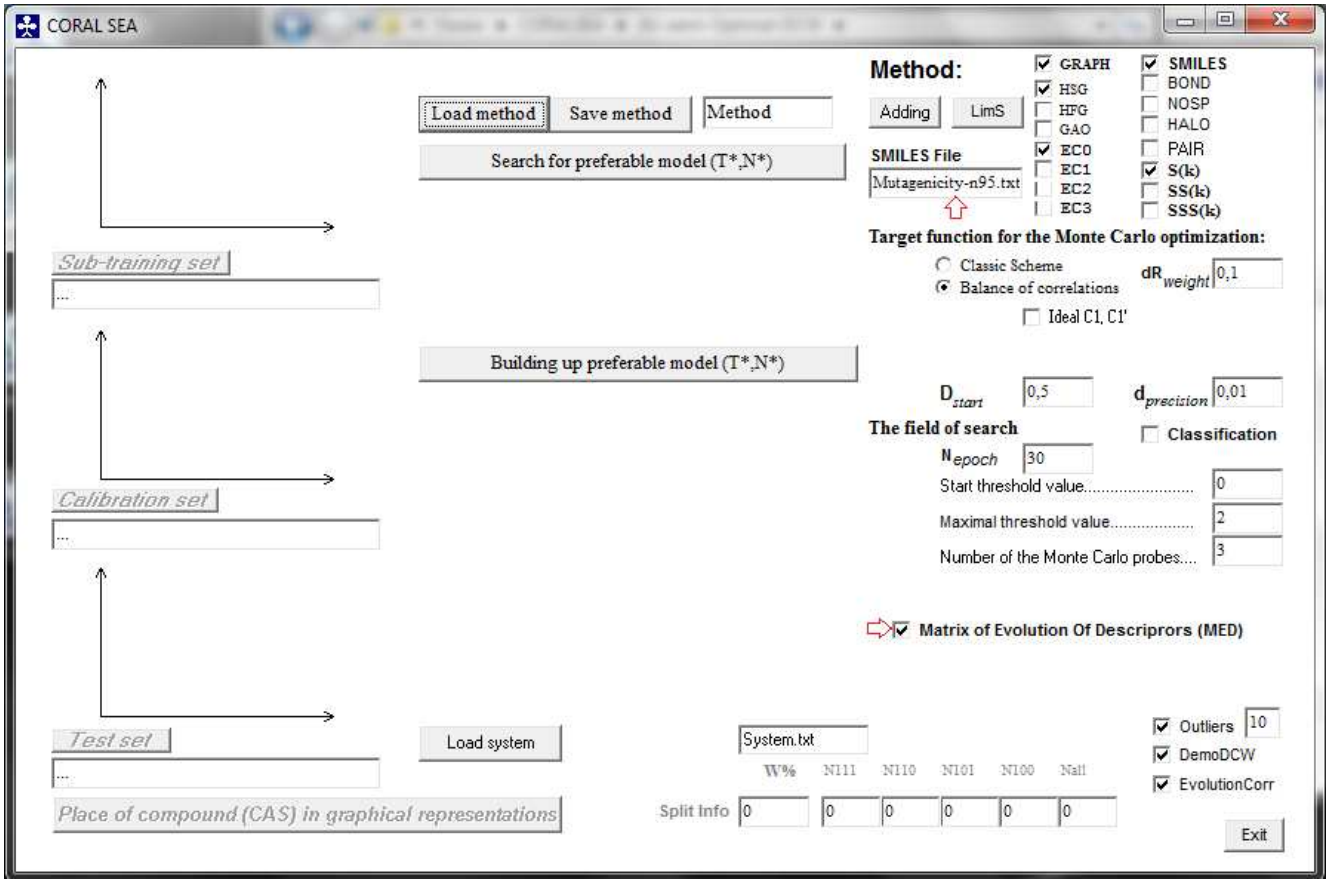
At the beginning, (epoch 1, epoch 2, ...), the status of optimal descriptors can be characterized as "random values". When, the Monte Carlo optimization is completed, the optimal descriptors are forced to be correlated with endpoint, as good as possible. But in the middle of the process, the optimal descriptors can be useful, as participants of the multiple regression analysis (MRA) together with widely used descriptors, such as topological indices, 3D-descriptors, descriptors of the quantum mechanics, etc. The folder "semi-optimal DCW" contains the program that gives possibility to analyze the optimal descriptors obtained at the middle phase of the Monte Carlo optimization. These data are represented in files “MED<threshold>-<probe>” (in other words, in files with names such as MED1-2.txt, MED3-1.txt, etc.). These values can be translated in MS-word file (x.doc) and further into the excel file (x.xls), in order to use these in procedures of the MRA.

This idea has been suggested by Dr. Pablo R. Duchowicz (INIFTA, La Plata, Argentina).

The folder contains two datasets: Mutagenicity-n95.txt [see ChemometrIntellLab109(2011)94] and Mutagencity-n48.txt [see CBDD-73(2009)94].

In order to carry out calculation with one of these data, one can do the following steps:

1. Run CORALSEA.exe;
2. Click “Load method”;
2. Modify method using as the SMILES-file “Mutagenicity-n95.txt” or “Mutagenicity-n48.txt”;



3. Select "Matrix of Evolution of Descriptors"
4. Click "Save method"
5. Click "Search of preferable model..."

Apparently, one can use this program for analysis of arbitrary data, if this data will be prepared in the form of analogical SMILES-file. In addition, one can select other options related to graph and / or to SMILES.

A6. Version oriented to organometallic compounds

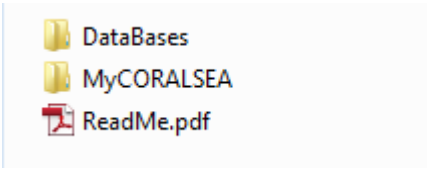
The folder (7)-Metals-and-Ions contains version of the CORAL software where the representation of SMILES attributes is based on 18 characters separated into three zones which contain 6 symbols (see page 11).

It gives possibility to detect the following SMILES fragments:
[AB]
[ABC]
[ABCD]
In fact these can be [Ni], [SiH], [NH4+], [C@H], [C@@H], and so on.

This version is temporary. This program cannot involve molecular graphs for the QSPR/QSAR analyses. However, this version can be useful for the case of QSPR/QSAR analysis of organometallic compounds.

The folder **(7)-Metals-and-Ions** contains ten random splits into sub-training, calibration, and test sets for 132 organometallic compounds (see also folder **(1)-Enthalpy-kJ-mol**).

A7. Contains of CORALSEA folder (comments)



DataBases contains the following folders:

(1)-Enthalpy-kJ-mol
This folder contains model for enthalpy of formation from elements for organometallic compounds and PDF of article where similar model is discussed.

(2)-logBeeToxicity
This folder contains model for toxicity towards bee and PDF of article where similar model is discussed.

(3)-lnR-TA98-mutagenicity
This folder contains model for mutagenicity and PDF of article where similar model is discussed.

(4)-LD50-Rat
This folder contains model for toxicity in rats and PDF of article where similar model is discussed.

(5)-DaphniaMagna
This folder contains model for toxicity towards *Daphnia magna* and PDF of article where similar model is discussed.

(6)-semi-Optimal DCW
This folder contains program that gives possibility to analyze optimal descriptors obtained at the any epoch of the Monte Carlo optimization (epoch 1, epoch 2, …, epoch 5, …) in role of possible participants of the multiple regression analysis (MRA).

(7)-Metals-and-Ions
This folder contains program that gives possibility to analyze substances which contain metals (e.g. [Cu], [Ni], etc.) as well as ions (e.g. [NH4+], [Cl-], [Na+], etc.). However this program cannot involve invariants of the molecular graph.

(8)-Anti-Sarcoma
This folder contains qualitative database on anti-sarcoma activity and PDF of article where these data have been taken.

(9)-Rate Constants
This folder contains models for predictions of rate constants of hydroxyl radical reaction and galley proofs of article where similar models are discussed.

(10)-LD50-Rat-NOSP
This folder contains model for toxicity in rats of organic compounds which are containing nitrogen, oxygen, sulfur, and phosphorus.

(11)-Duchowicz's CORAL
This folder contains version of CORAL that gives preferable T* and N* (see **2.5. Sketch of theory)** for both training set and test set (i.e., not only for test set).

(12)-quasi-SMILES-for-nano-QSAR-demo
Example of nano-QSAR based on quasi-SMILES

MyCORALSEA
This folder contains two sub-folders which are examples of (i) a linear regression model (REGRESSION); and (ii) a classification model (CLASSIFICATION). User can modify the containing of these sub-folders according to his /her tasks by means of modification of SMILES.txt and METHOD.txt.

ReadMe.pdf
File that contains this Reference Manual.

A8. Updates April 2014

1. MyCORALSEA folder contains two sub-folders: (i) Example of the linear regression model (REGRESSION); and (ii) Example of the classification model (CLASSIFICATION).
2. The CORALSEA.exe calculates addition criteria of the quality of distribution of available data into the sub-training set, calibration set, test set, and validation set.
3. The file “input.txt” (see 4.2) should be placed in the same folder where CORALSEA.exe, not in the folder “Model”.
4. Matthews Correlation Coefficient (MCC) is added for the classification model.
5. Brief instructions which appear during of the calculations are added in the modified version of the program.
6. Paths of length 2 are available in the new version of the CORAL (pt2).
7. Paths of length 3 are available in the new version of the CORAL (pt3).
8. Valence shells of second range are available in new version of the CORAL (S2).
9. Valence shells of third range are available in new version of the CORAL (S3).
10. Nearest neighbors codes are available in the new version of the CORAL (NNC).

Quality of an attribute SA_k

The measure of quality of molecular features which are extracted from SMILES or from molecular graph is calculated as the following:

$$SA_k\text{-Defect} = \begin{cases} \dfrac{\left| P_{TRN}(SA_k) - P_{TST}(SA_k) \right|}{N_{TRN}(SA_k) + N_{TST}(SA_k)}, & \text{if } N_{TST}(SA_k) > 0 \\ 1, & \text{otherwise} \end{cases} \qquad (11)$$

where the $P_{TRN}(SA_k)$ is the probability of presence of the SA_k in SMILES of the sub-training set, i.e.

$P_{TRN}(SA_k) = N_{TRN}(SAk) \ / \ N_{TRN}$

The $P_{TRN}(SA_k)$ is the probability of presence of the SAk in SMILES of the test set, i.e.

$P_{TST}(SA_k) = N_{TST}(SAk) \ / \ N_{TST}$

The $N_{TRN}(SAk)$ is the number (frequency) of SMILES which contain SA_k in the sub-training set;
The N_{TRN} is the total number of SMILES in the sub-training set;
The $N_{TST}(SAk)$ is the number (frequency) of SMILES which contain SAk in the test set;
The N_{TST} is the total number of SMILES in the test set.

The logic: if the probability of SA_k in the sub-training set is equal to the probability of SA_k in the test set it is the ideal situation and the defect is zero. However, this situation is not typical, i.e. the difference between the probability of SA_k in the sub-training set and the probability of SA_k in the test set is not zero. Under such circumstances, the frequency of SA_k in the sub-training set and in the test set also should be taken into account: if these are small then the defect of SA_k must be larger. Finally, if SAk is absent in the test set, the SAk-defect is maximal. Thus, the measure calculated with Eq. 11 can be used for the classification of the active (not blocked) attributes

Split-Defect

Having the numerical data on the defects of SA_k which are involved in building up model one can estimate the defect of a split (i.e. the distribution into the visible sub-training, calibration, and test sets and invisible external validation set) based on the

$$\text{Split-Defect} = \sum \text{SAk-Defect} \qquad (12)$$

It is to be noted that blocked SA_k are not involved in the calculation with Eq. 12.

The criterion calculated with Eq. 12 gives possibility to compare two splits. If Split-Defect for SplitX is equal to X and Split-Defect for SplitY is equal to Y then

(i) SplitX is better than SplitY if X<Y;
(ii) SplitY is better than SplitX if X>Y;
(iii) SplitX and SplitY are identical if X=Y.

The selection of substances into the domain of applicability

Having the numerical data on the defects of SAk one can compare reliability of the prediction for an substance, using the following criterion (DefectSMILES):

$$\frac{DCW(T^*,N^*,SMILES) = \sum CW(SAk)}{DefectSMILES = \sum SAk\text{-}Defect}$$

The domain of applicability can be defined as the following: Substance is fall into the domain of applicability if its DefectSMILES obeys the condition:

$$DefectSMILES < 2 * \overline{DefectSMILES}$$

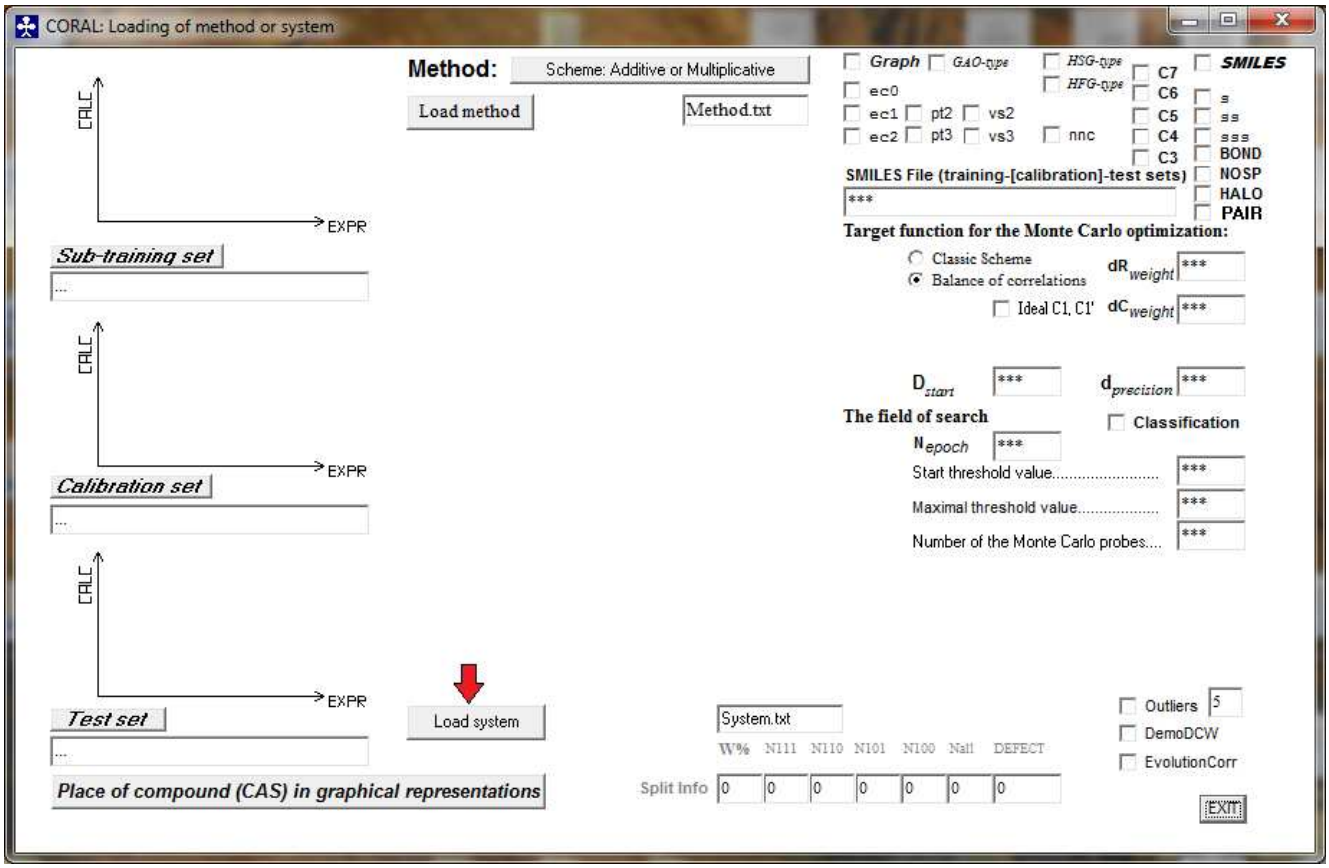
where $\overline{DefectSMILES}$ is average for visible set (sub-training, calibration, and test sets).

Thus the DefectSMILES gives possibility to define the domain of applicability for the CORAL-models. This information is represented in file model/#Output.txt which contains prediction of the endpoint for the external (invisible) validation set (see 4.2). Unfortunately, the above criteria are not garantie, but the probabilistic measure of quality of distribution into the visible training and invisible validation sets. SMILES with large DefectSMILES should be estimated as “suspect” ones, however their categorization in role of outliers should be based on additional examination.

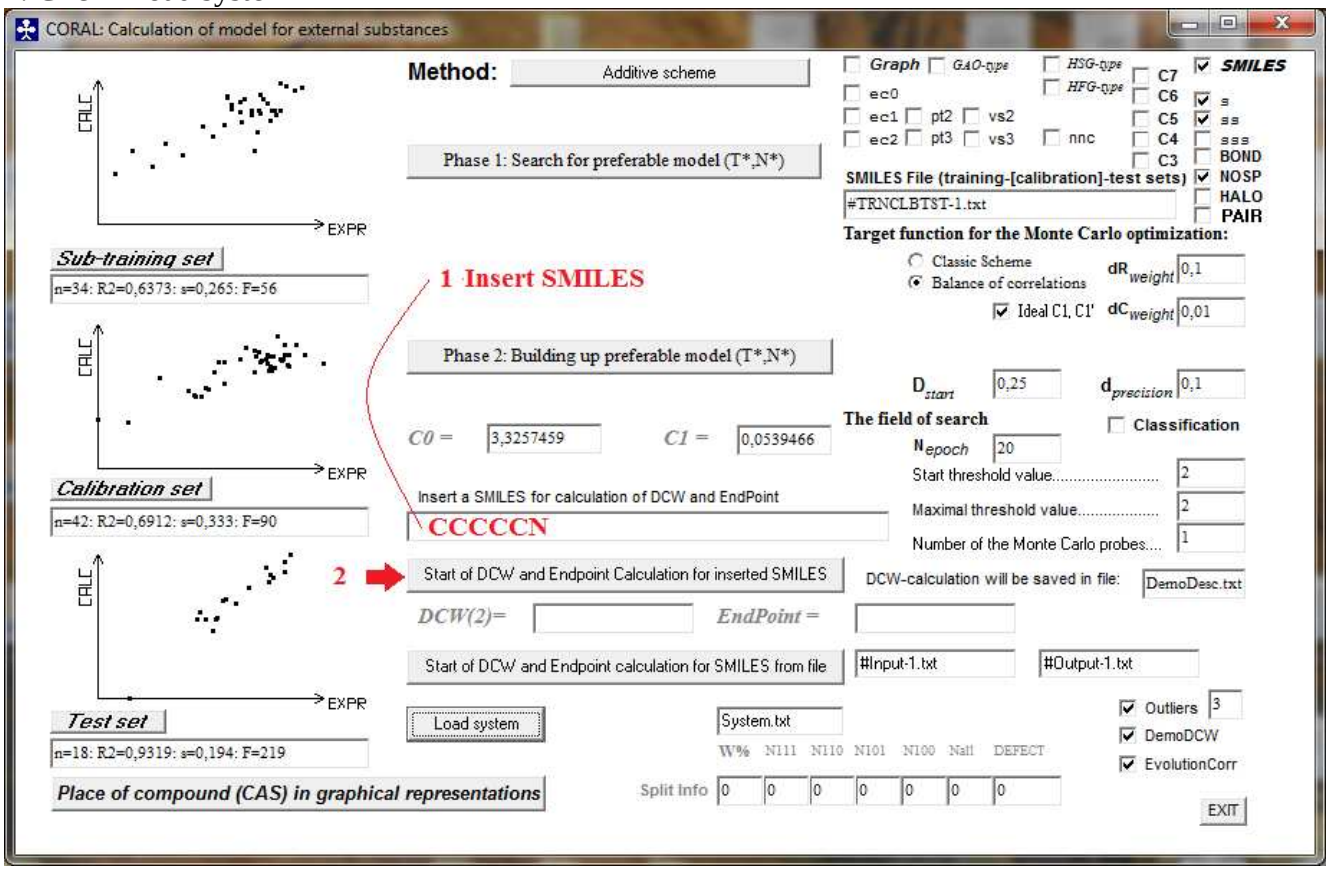
File "SMILESdefect.txt" contains data on the defectSMILES for external set taken from "Input.txt"

Having the calculated model, one can check up whether a given SMILES falls into the domain of applicability:

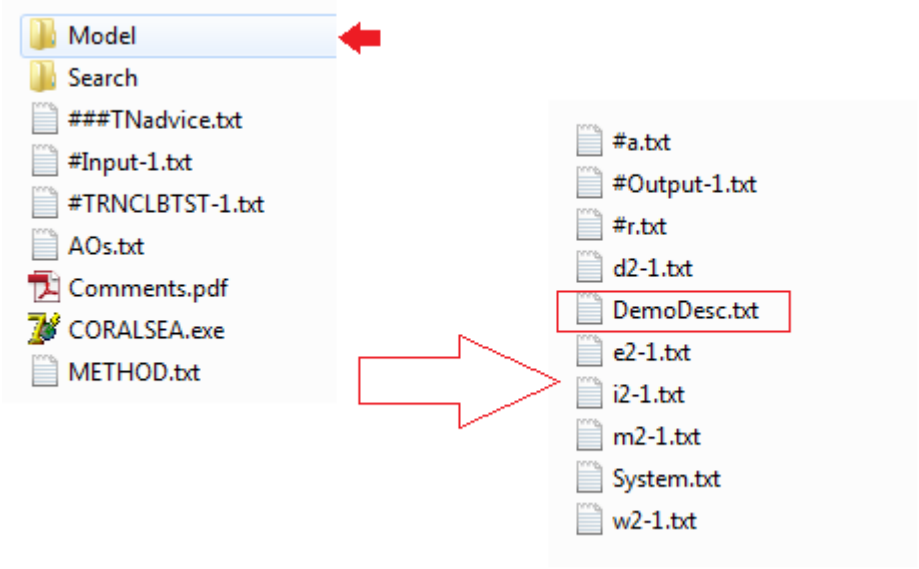
1. Run CORALSEA.exe



2. Click “Load system”



3. Insert a SMILES, e.g. CCCCCN and click ”Start of DCW and Endpoint Calculation…”
4. Open the file “DemoDesc.txt” which is placed in folder “Model”:



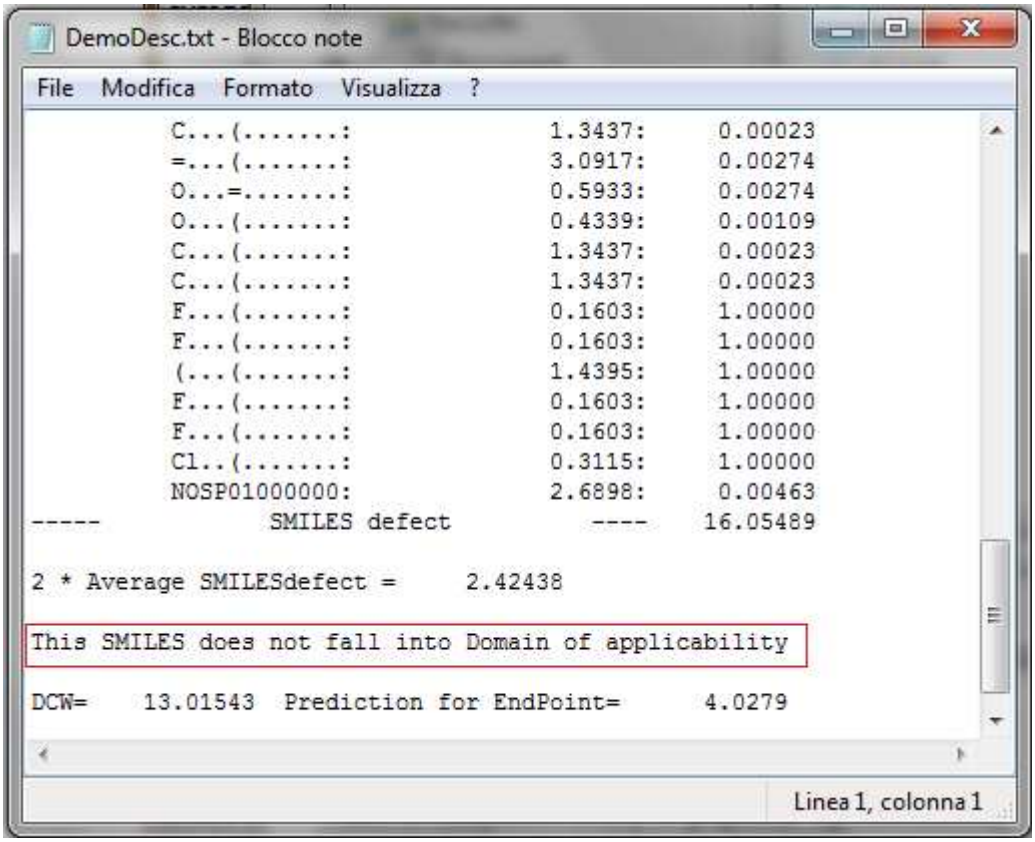
The file “DemoDesc.txt” has the following contant:

DemoDesc.txt - Blocco note

File Modifica Formato Visualizza ?

```
This file contains an example of calculations of DCW and endpoint
for SMILES that is inserted in dialog box
CCCCCN
Selected threshold is 2

SMILES attribute (SA) : Correlation Weight :   SA defect
          C...........:              -0.9961:     0.00000
          C...........:              -0.9961:     0.00000
          C...........:              -0.9961:     0.00000
          C...........:              -0.9961:     0.00000
          C...........:              -0.9961:     0.00000
          N...........:               0.7488:     0.00206
          C...C.......:               1.1249:     0.00308
          C...C.......:               1.1249:     0.00308
          C...C.......:               1.1249:     0.00308
          C...C.......:               1.1249:     0.00308
          N...C.......:              -3.4358:     0.00189
          NOSP10000000:               9.1263:     0.00327
-----            SMILES defect          ----     0.01952

2 * Average SMILESdefect =     2.42438

This SMILES falls into Domain of applicability

DCW=     5.95821  Prediction for EndPoint=      3.6472
```

If “FC(F)(Cl)C(=O)OC(=O)C(F)(F)Cl” is inserted, the contant will be other:



File "Expr-Calc.txt" contains data on the delta (experimental endpoint minus calculated endpoint) for external set taken from "Input.txt"

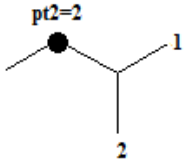
There are some changes of the dialog window and files of the CORALSEA.exe, however we hope these are apparent and do not need additional comments. For example: (i) Split info is added by DEFECT described in the previous page; (ii) button “Rare?” (LimS / LimN) is deleted, because in the new version only LimS is used to define rare and not rare attributes (see, also, page 14); (iii) #Output.txt contains additional information related to criteria described above (pages 47 and 48).

The MyCORALSEA folder contains updates files, whereas files in Databases folder are not modified.

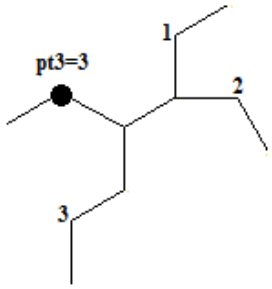
Examples of the CORAL-models in folders “CLASSIFICATION” and “REGRESSION” contain brief comments.

A9. Comments for additional attributes which can be extracted from graph

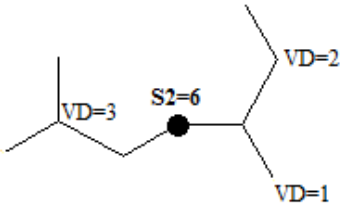
pt2[k]: the number of paths of length 2 which are starting from k-th vertex



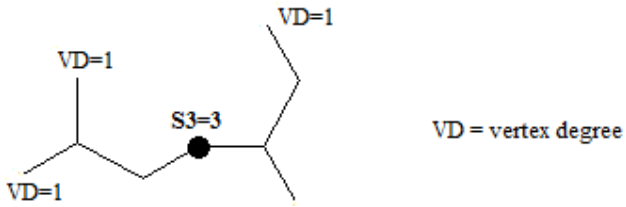
pt3[k]: the number of paths of length 3 which are starting from k-th vertex



S2[k]: the sum of vertex degrees which take place at topological distance 2 relatively to k-th vertex



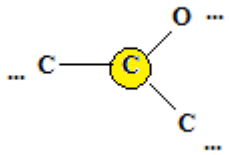
S3[k]: the sum of vertex degrees which take place at topological distance 3 relatively to k-th vertex



For more details please see: Toropov, A.A., Benfenati, E. Correlation weighting of valence shells in QSAR analysis of toxicity (2006) Bioorganic and Medicinal Chemistry, 14 (11), pp. 3923-3928

NNC[k]: The nearest neighboring codes are calculated as the following:

In general NNC[k] = $100*N_{all} + 10*N_{carbon} + N_{noncarbon}$ (N_{all}, N_{carbon}, and $N_{noncarbon}$ are the total number of neighbors for k-th vertex, the number of vertices which are carbon, and the number of vertices which are not carbon, respectively)

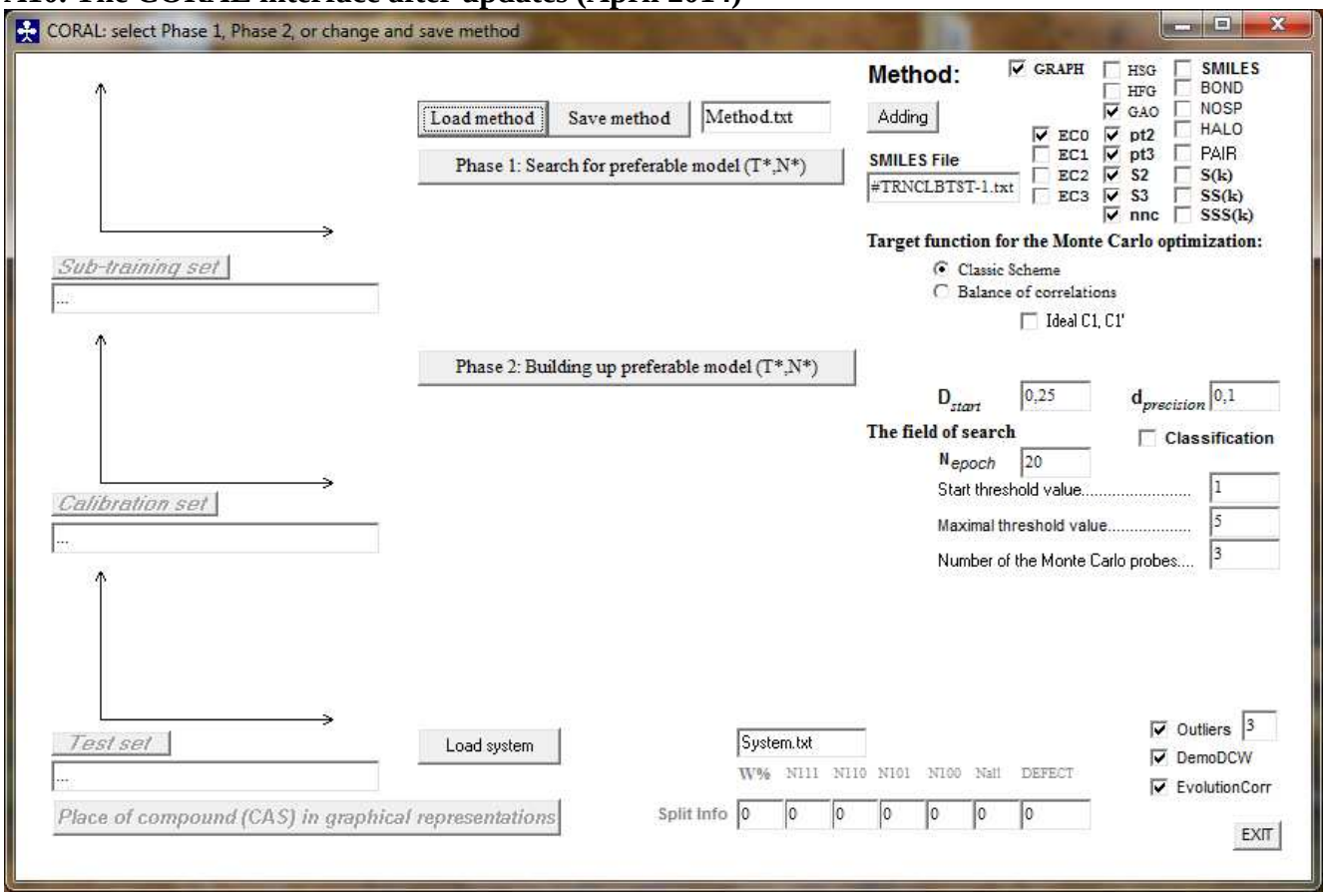


For the example

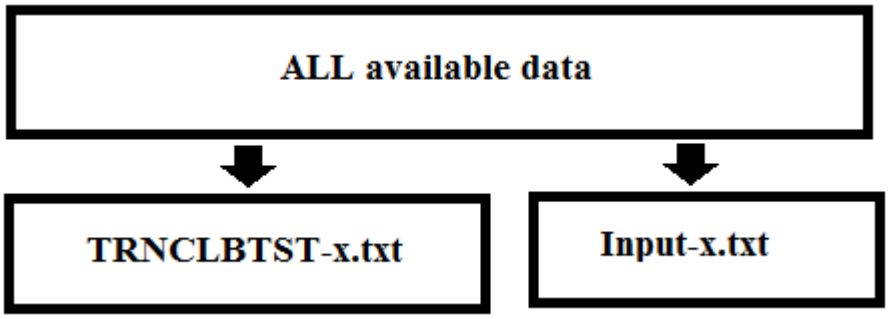
NNC[k] = 3*100 + 10*2 + 1 = 321

In the case of graph of atomic orbitals NNC[k]=$100*N_{all} + 10*N(2p^2) + N(non\ 2p^2)$, [$N(2p^2)$ is the number of neighboring vertices which are $2p^2$ and $N(non\ 2p^2)$ is the number of neighboring vertices which are not $2p^2$].

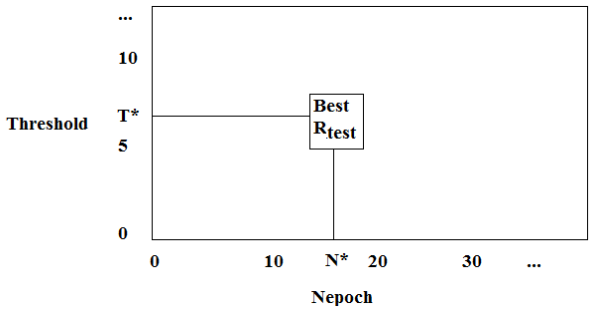
A10. The CORAL interface after updates (April 2014)



In spite of changes, the logic of building up CORAL model remains the same.
The first: with using available data one should prepare two files TRNCLBTST-x.txt and Input-x.txt



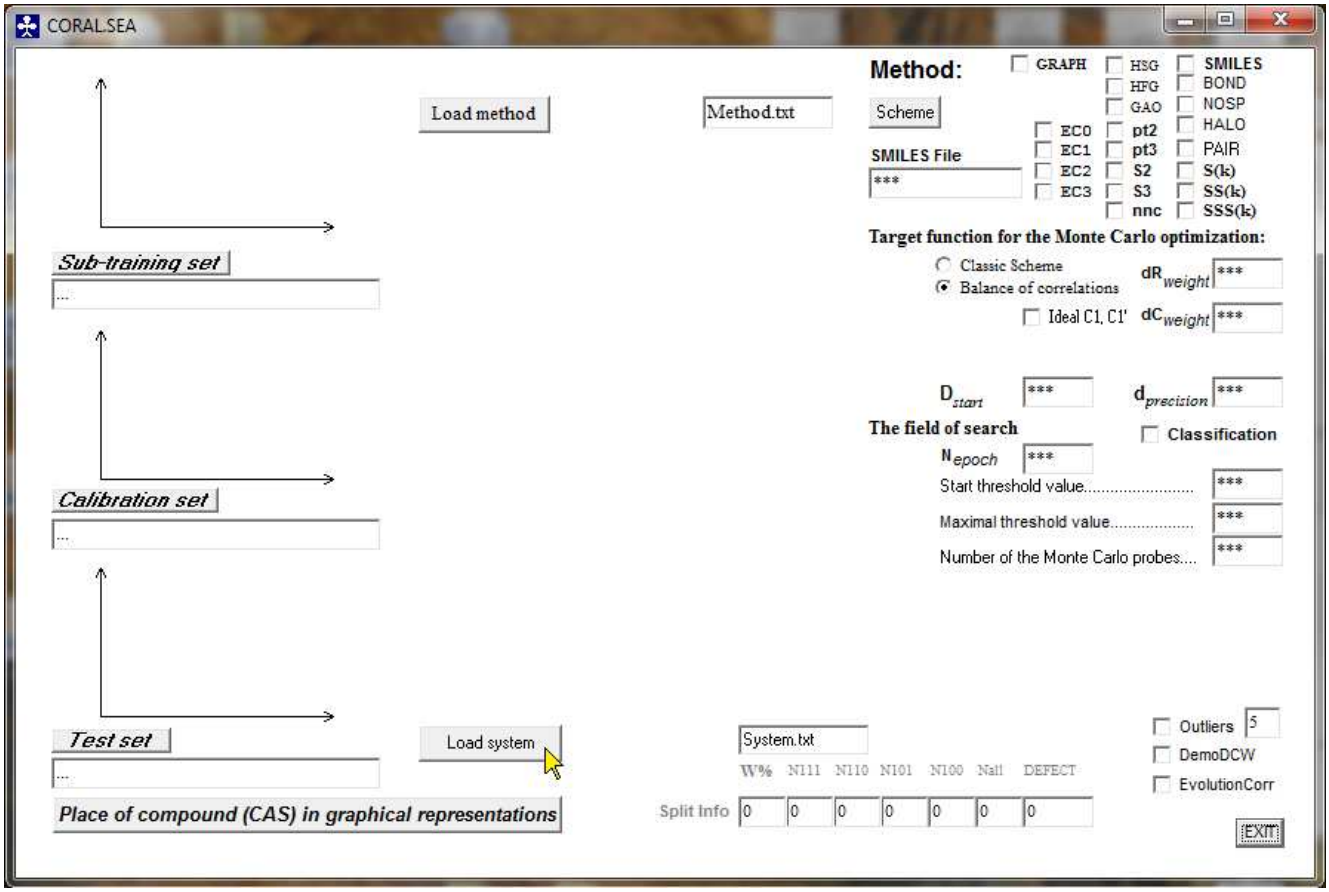
The second: the definition of the preferable values of the threshold (T*) and the number of epochs of the Monte Carlo optimization (N*), which give best statistics for the test set:



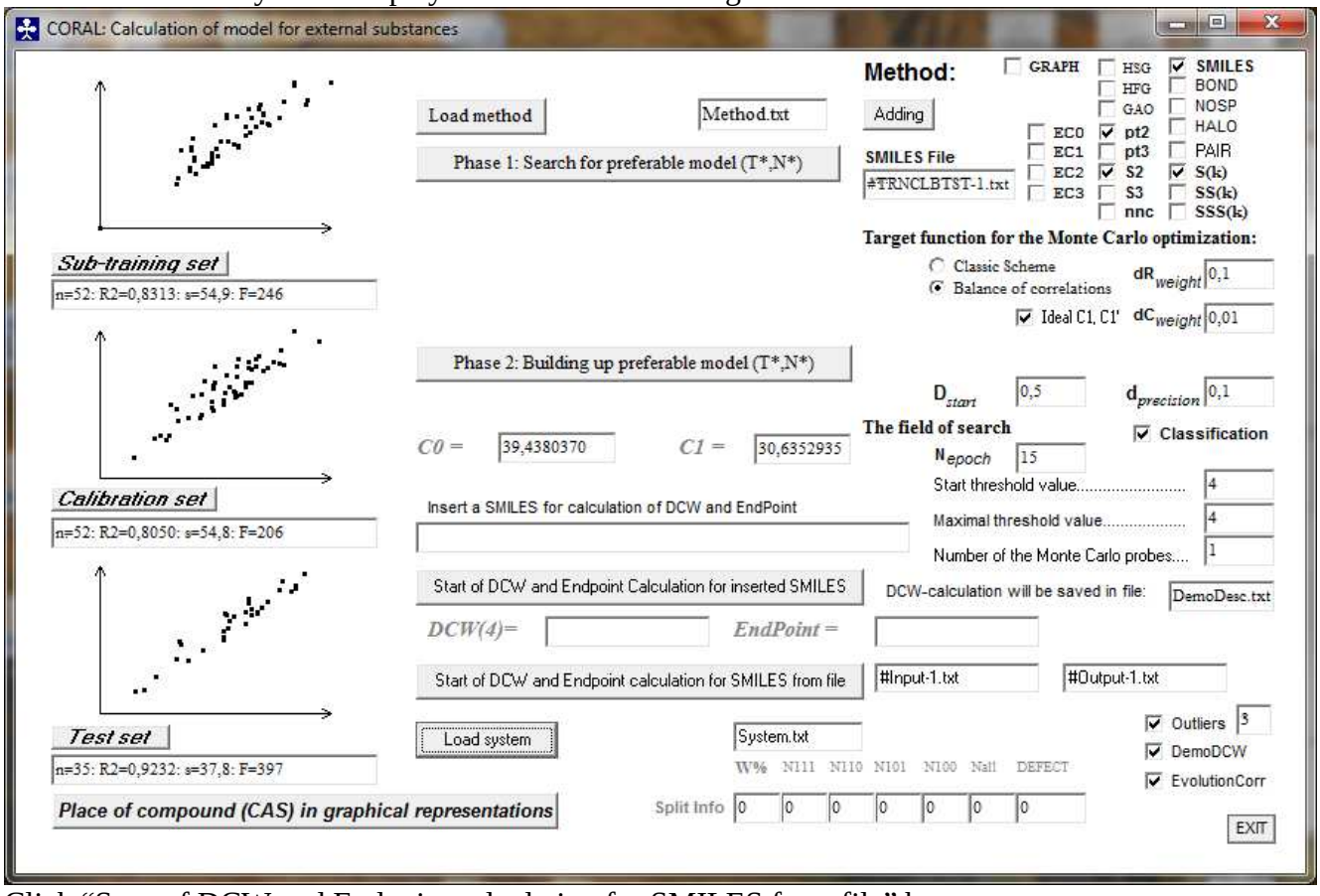
The third: the checking up (T*,N*) model with invisible set from file Input-x.txt

A11. Graphical representation of model for external validation set

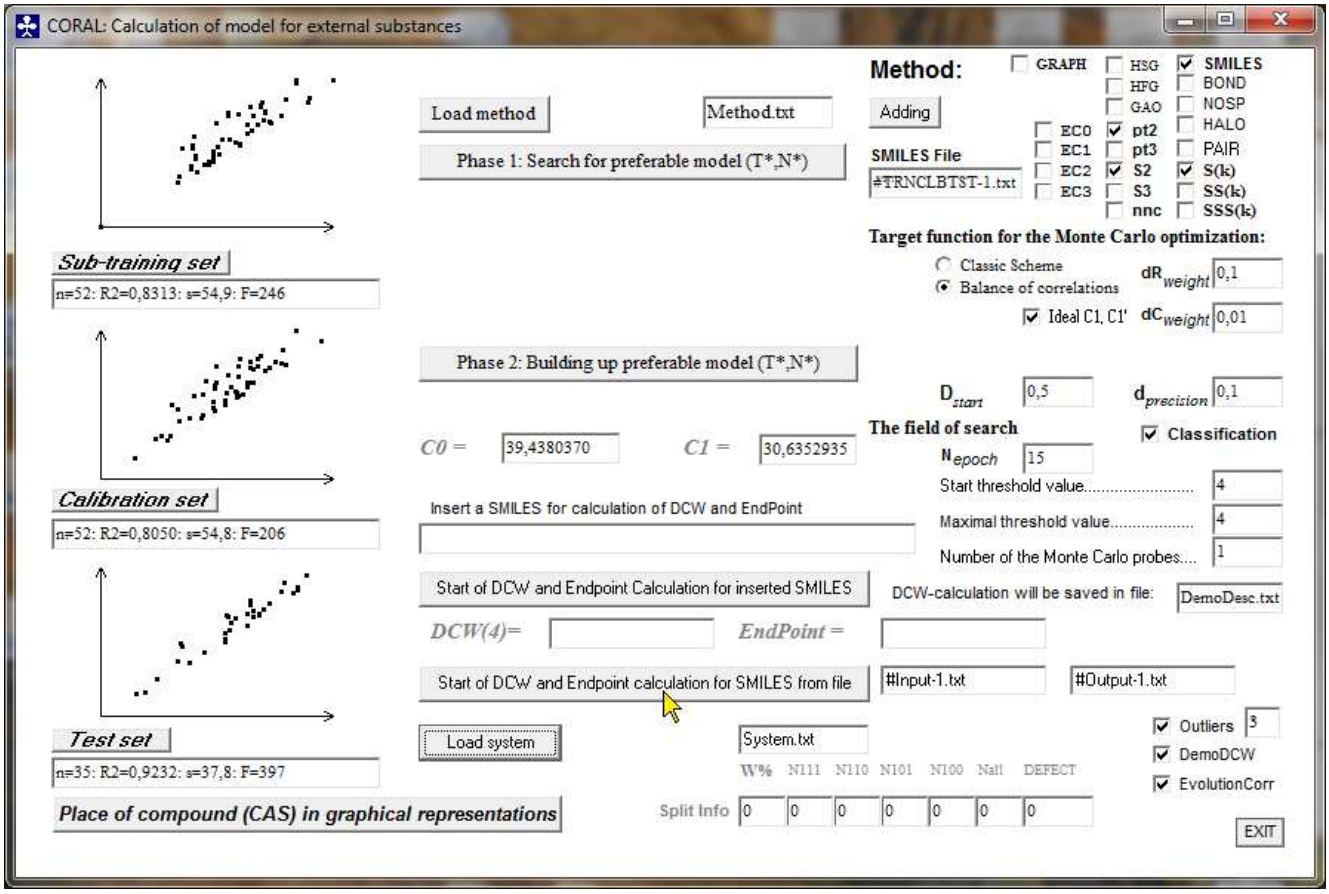
If you have prepared a model (step 4, page 29) you can use the model: click “Load system” button



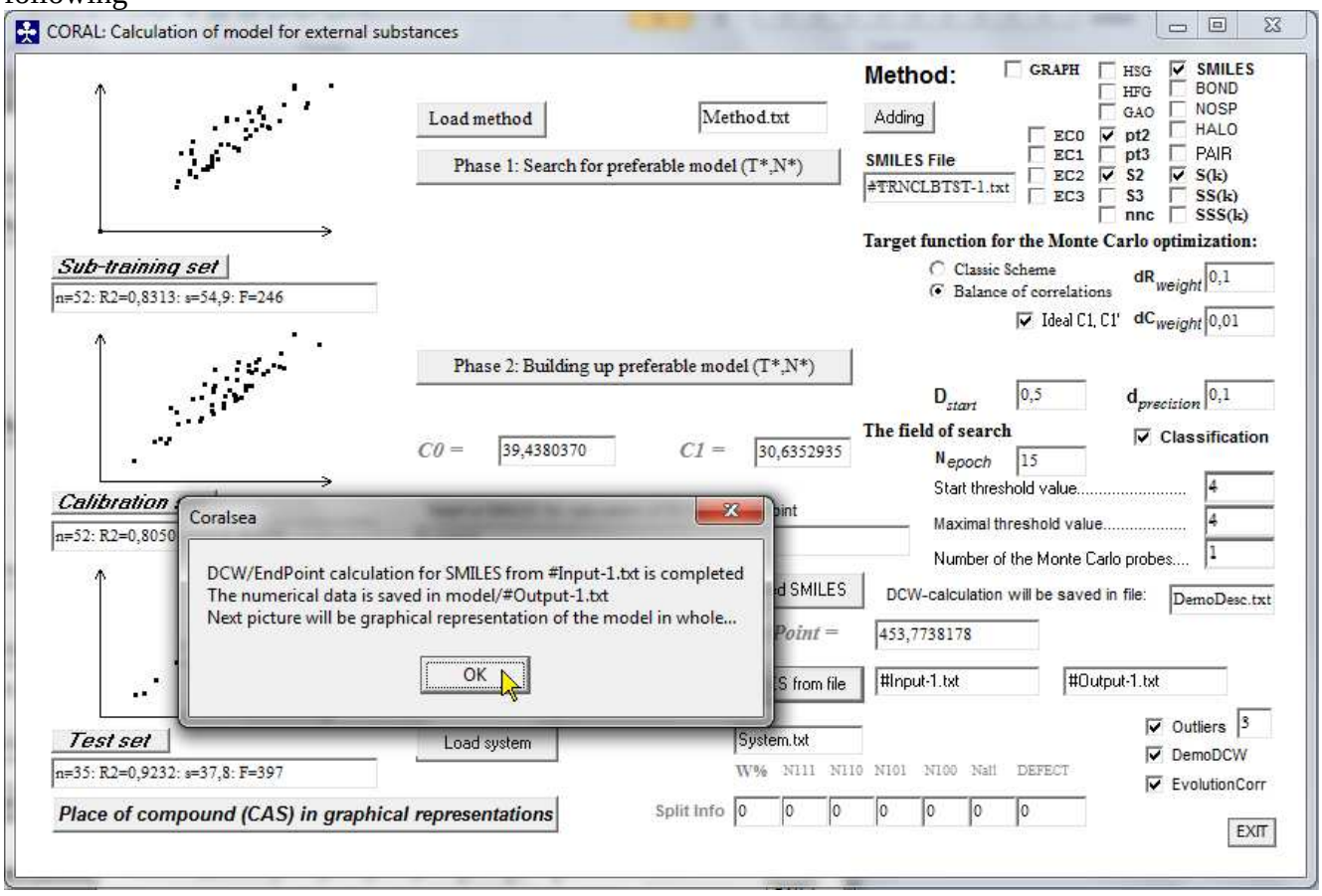
After click “Load System” display becomes the following:



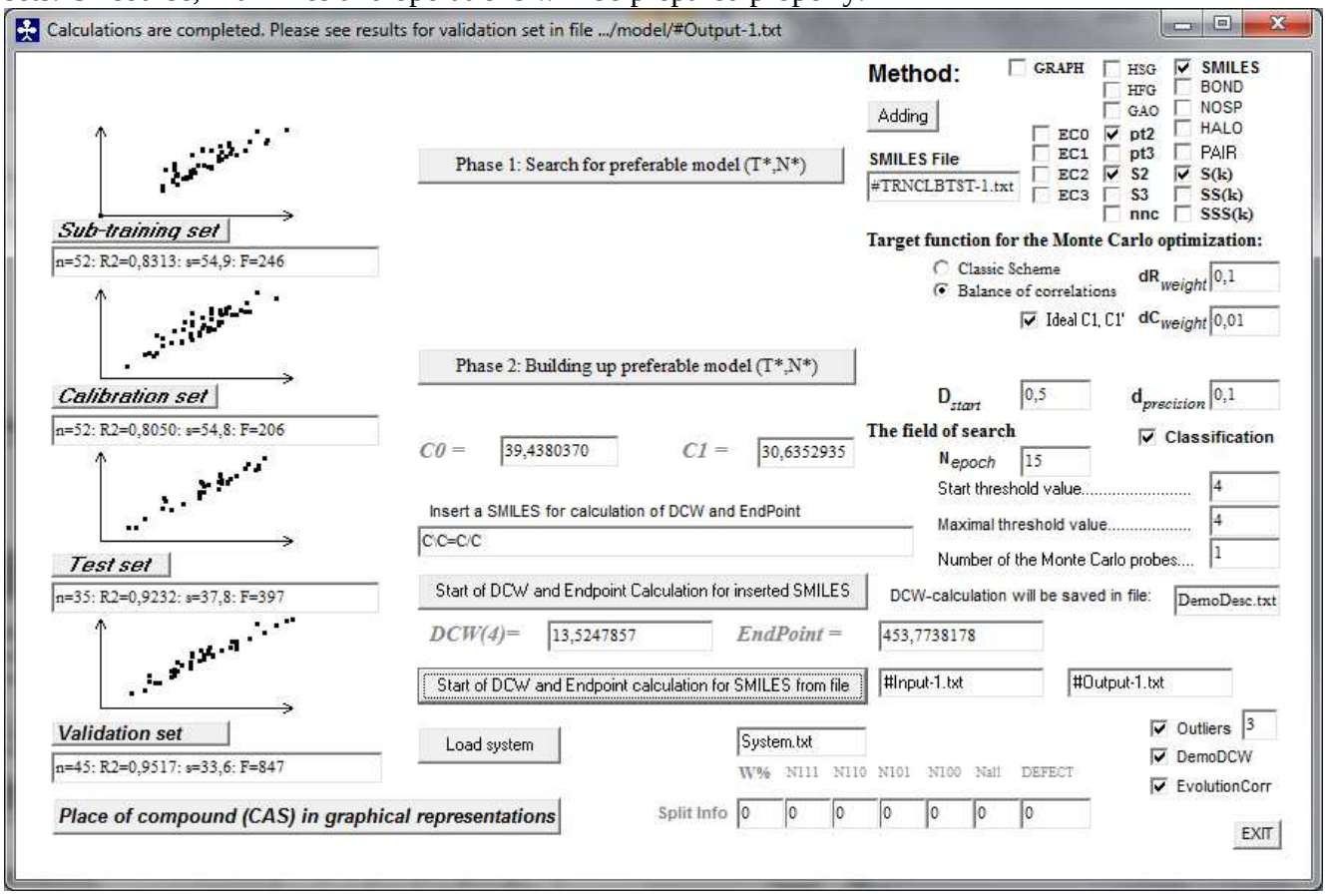
Click “Start of DCW and Endpoint calculation for SMILES from file” button.



After click “Start of DCW and Endpoint calculation for SMILES from file” display becomes the following

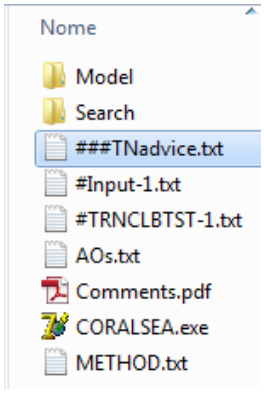


Click “OK” and you will see graphical representation for sub-training, calibration, test, and validation sets. Of course, if all files and operations will be prepared properly.

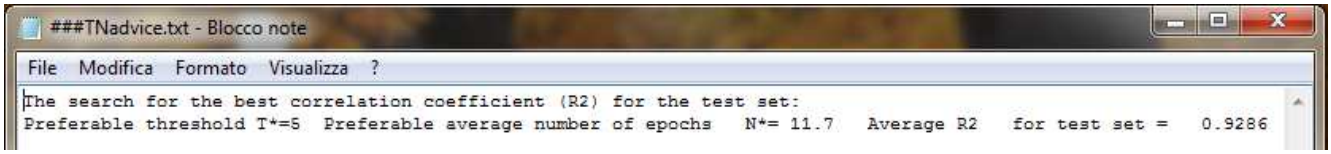


If everything is OK, you can use service (A1, page 37) for sub-training, calibration, test, and validation sets.

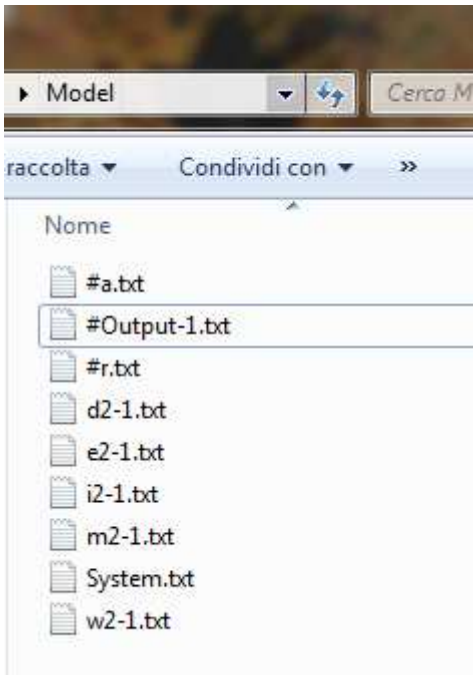
The described version of the CORAL provides additional file “###TNadvice.txt”:



This file contains recommendation for values of T* and N* (see section 2.5).



The file “output.txt” in folder model



after described operations will contain the following information:

#Output-1.txt - Blocco note

File Modifica Formato Visualizza ?

```
This file contains prediction for the invisible validation set
Data taken from file #Input-1.txt

The average of DefectSMILES =     1.21219
Substance falls into domain of applicability if DefectSMILES <      2.42438

          CAS  :SMILES                                                :  DCW       :  Expr      :  Calc      :DefectSMILES: Applicability
              2:FC(F)(Cl)C(=O)OC(=O)C(F)(F)Cl                         :    14.82484:      3.9500:      3.9791:     16.0549:           No
              5:O=C1OC(=O)C=C1                                        :    17.41316:      3.9800:      4.0965:      0.0601:    YES
              6:CC1=CC(=O)OC1=O                                       :     5.87859:      3.5800:      3.5735:      0.0618:    YES
              8:CCCCCC(=O)OC(=O)CCCCC                                 :     7.56649:      3.6500:      3.6500:      0.0620:    YES
             24:O=C1OC(=O)C2C3C=CC(C12)C1C3C(=O)OC1=O                 :     4.81012:      3.6300:      3.5250:      4.1357:           No
             39:O=C1OC(=O)C2=C1CCCC2                                  :    10.34709:      3.7300:      3.7761:      0.0758:    YES
             48:Nc1ccc2C(=O)OC(=O)c3cccc1c23                          :     1.57537:      3.6400:      3.3783:      0.1409:    YES
             57:CCCCCN                                                :     6.53418:      3.7800:      3.6032:      0.0195:    YES
             61:CC(C)(C)N                                             :    11.14732:      3.8600:      3.8124:      1.0124:    YES
             65:CCC(N)CC                                              :     9.66614:      3.8100:      3.7452:      0.0181:    YES
             71:NCCCCCCN                                              :     3.75360:      3.6200:      3.4771:      0.0265:    YES
             87:NC(CO)C(O)=O                                          :    -5.11946:      3.3600:      3.0747:      0.0289:    YES
             91:CC(C)C(N)C(O)=O                                       :    -0.23022:      3.2700:      3.2964:      0.0274:    YES
             93:NC(C(O)=O)c1ccc(Cl)cc1                                :    -8.07619:      3.0600:      2.9406:      2.1184:    YES
             95:NC(CCCNC(N)=N)C(O)=O                                  :   -10.30763:      3.1500:      2.8395:      0.0485:    YES

Rm2(x,y) calculation for validation set from input file
n           =      15
r2          =    0.9055
r02         =    0.8661
rr02        =    0.7644
(r2-r02)/r2 =    0.0436 should be < 0.1
(r2-rr02)/r2=    0.1558 should be < 0.1
k           =    0.9774 should be 0.85 <  k < 1.15
kk          =    1.0218 should be 0.85 < kk < 1.15
Rm2(test)   =    0.7256 should be > 0.5


Rm2(y,x) calculation for validation set from input file
n           =      15
r2          =    0.9055
r02         =    0.7644
rr02        =    0.8661
(r2-r02)/r2 =    0.1558 should be < 0.1
(r2-rr02)/r2=    0.0436 should be < 0.1
k           =    1.0218 should be 0.85 <  k < 1.15
kk          =    0.9774 should be 0.85 < kk < 1.15
R*m2(test)  =    0.5654 should be > 0.5

Average Rm2 = 0.6455 should be larger 0.5
Delta Rm2 = 0.1603 should be lower 0.2

RMSE=   0.1574
MAE =   0.1159

The number of active (not blocked) attributes      =42

Q2-Training set    =   0.5675
Q2-Calibration set =   0.6525
Q2-Test set        =   0.8976
Q2-Validation set  =   0.8752
```

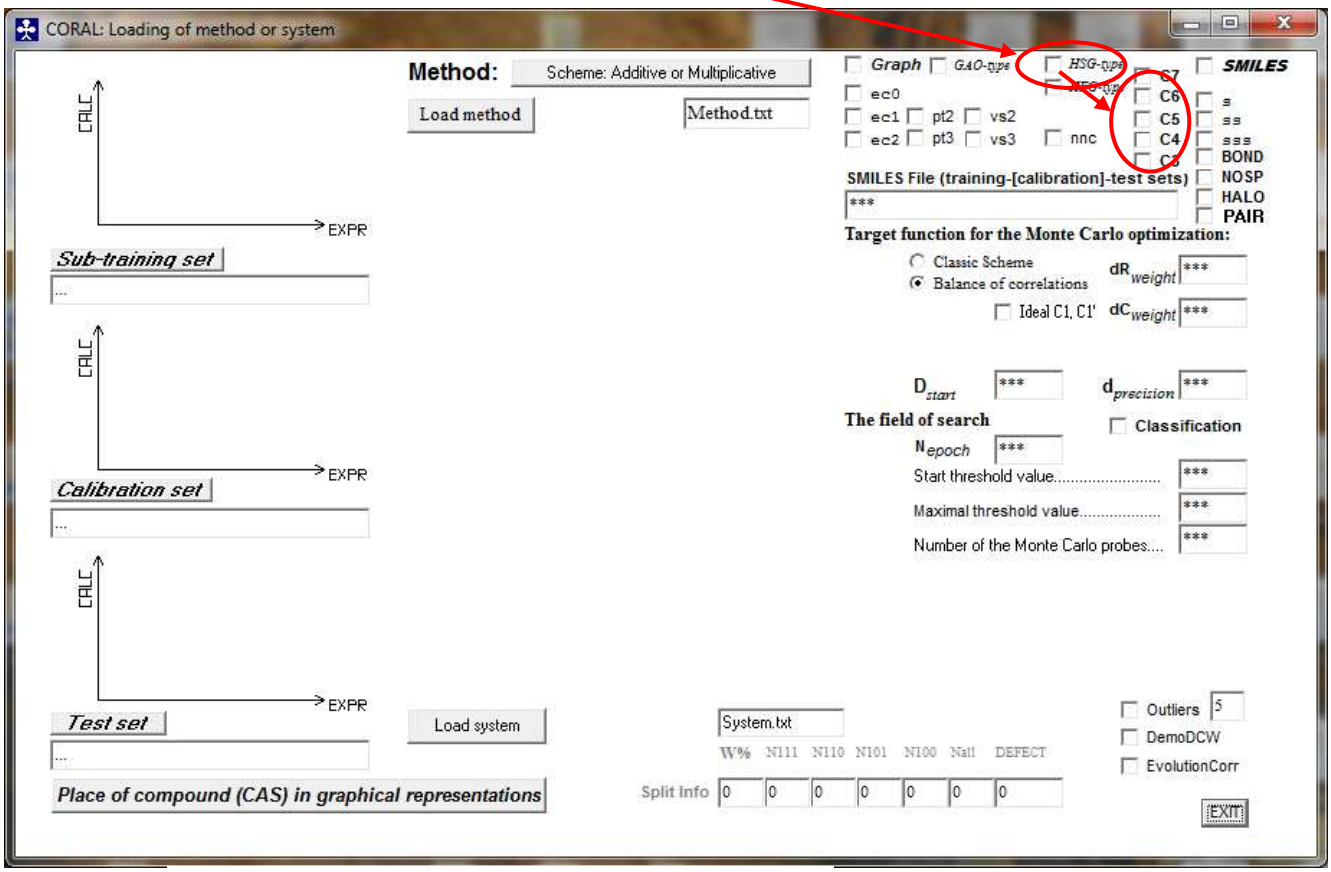
These statistical characteristics are described in files mX-Y.txt (see section 3.6).

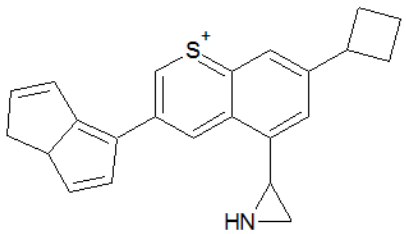
A12. Updates of November 26, 2014. Analysis of cycles

The analysis of cycles is available only for **HSG**.
Options c7, C6, C5, C4 and C3 are a tool to take into account possible influence of cycles. These features of molecular structure are encoded by attributes of view

Cx.....N...
1 2 3 4 5 6 7 8 9 10 11 12
x=3,4,5,6,7

where N is the number of cycles in molecule





c2c1[s+]cc(cc1c(cc2C3CCC3)C4CN4)C=6C=CC5CC=CC5=6

An example,
Cyclic attributes for this structure are the following:

Attribute	Comments
C6...AH.2...	There are **two** six-member cycles with aromaticity and presence of heteroatoms ('A' is indicator of aromaticity or double/triple bonds, 'H' is indicator of heteroatom(s) in cycle)
C5.....2...	There are **two** five-member cycles
C4.....1...	There is **one** four member cycle
C3....H.1...	There is **one** three-member cycle with heteroatom ('H' is indicator of heteroatom in cycle)