



Monte Carlo Optimization

Acquisition of dataset

There are different sources for obtaining the data required to carry out QSAR studies

- From various database like ChEMBL Database, ExCAPE-DB, PubChem and etc.
- From literature search.
- In-house synthesized molecules and their biological activity.

Preparation of the dataset (Step 1)

The input format for CORAL software was separated into three zones. Each zone is separated by a single space as per the criteria of CORAL software format: zone 1: the indicator sign for various sets along with the compound number; zone 2 the SMILES format of the compound; zone 3: the biological activity (endpoint) as pIC_{50} of the compounds.

Zone 1 **Zone 2** **Zone 3**

#2 c1c(ccc2c1/C(=N\O)/C(=N2)c1c[nH]c2c1cc(cc2)F)F 4.922

Dataset - Notepad

File	Edit	Format	View	Help
1	c1c(ccc2c1/C(=N\O)/C(=N2)c1c[nH]c2c1cc(cc2)Br)Br	4.791		
2	c1c(ccc2c1/C(=N\O)/C(=N2)c1c[nH]c2c1cc(cc2)F)F	4.922		
3	c1c(ccc2c1/C(=N\O)/C(=N2)c1c[nH]c2c1cc(cc2)C)C	4.951		
4	c1c(ccc2c1/C(=N\O)/C(=N2)c1c[nH]c2c1cc(cc2)OC)OC	5.215		
5	c1c(ccc2c1/C(=N\O)/C(=N2)c1c[nH]c2c1cc(cc2)Oc1cccc1)Oc1cccc1	4.71		
6	c1c(ccc2c1/C(=N\O)/C(=N2)c1c[nH]c2c1cc(cc2)N)N	4.829		
7	c1c(ccc2c1/C(=N\O)/C(=N2)c1c[nH]c2c1cc(cc2)NC(=O)C)NC(=O)C	4.659		
8	c1c(ccc2c1/C(=N\O)/C(=N2)c1c[nH]c2c1cc(cc2)[N+](=O)[O-])[N+](=O)[O-]	4.981		
9	c1c(ccc2c1/C(=N\O)/C(=N2)c1c[nH]c2c1cc(cc2)C#N)C#N	4.826		
10	c1c(ccc2c1/C(=N\O)/C(=N2)c1c[nH]c2c1cc(cc2)C(=O)OC)C(=O)OC	4.688		
11	c1cc(cc2c1/C(=N\O)/C(=N2)c1c[nH]c2c1ccc(cc2)F)F	4.672		
12	c1cc(cc2c1/C(=N\O)/C(=N2)c1c[nH]c2c1ccc(cc2)C(=O)OC)C(=O)OC	4.586		
13	c1ccc(cc2c1/C(=N\O)/C(=N2)c1c[nH]c2c1cccc2C)C	4.756		
14	c1(ccccc2c1/C(=N\O)/C(=N2)c1c[nH]c2c1c(ccc2)C)C	4.348		
15	c1(ccccc2c1/C(=N\O)/C(=N2)c1c[nH]c2c1c(ccc2)C#N)C#N	4.473		
16	c1cccc2c1/C(=N\OCc1cccc1)/C(=N2)c1c[nH]c2c1cccc2	4.713		
17	c1c(ccc2c1/C(=N\OCc1cccc1)/C(=N2)c1c[nH]c2c1cc(cc2)F)F	4.977		
18	c1c(ccc2c1/C(=N\OCc1cccc1)/C(=N2)c1c[nH]c2c1cc(cc2)C#N)C#N	4.848		
19	c1cc(cc2c1/C(=N\OCc1cccc1)/C(=N2)c1c[nH]c2c1ccc(cc2)F)F	4.804		
20	c1cccc2c1/C(=N\OCc1ccc(cc1)C)/C(=N2)c1c[nH]c2c1cccc2	4.648		
21	c1c(ccc2c1/C(=N\OCc1ccc(cc1)C)/C(=N2)c1c[nH]c2c1cc(cc2)F)F	4.867		
22	c1c(ccc2c1/C(=N\OCc1ccc(cc1)C)/C(=N2)c1c[nH]c2c1cc(cc2)Br)Br	4.877		
23	c1cccc2c1/C(=N\OCc1ccc(cc1)F)/C(=N2)c1c[nH]c2c1cccc2	4.568		

Input file format for the CORAL software

Software required

Coral (CORrelation and Logic)

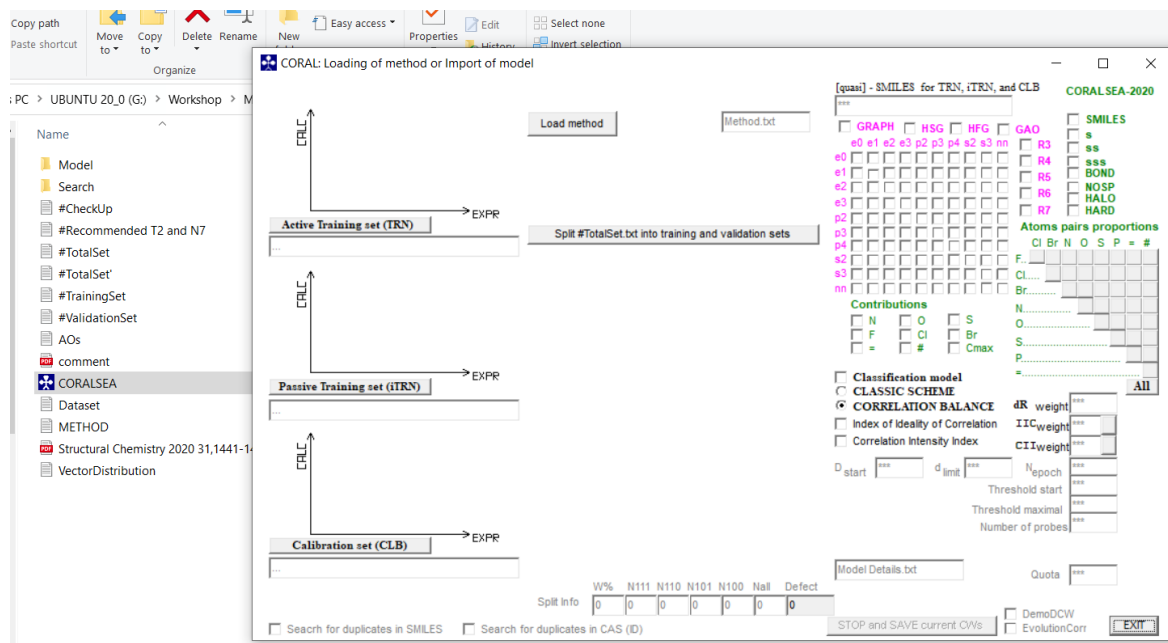
<http://www.insilico.eu/coral/SOFTWARECORAL.html>

Step 2

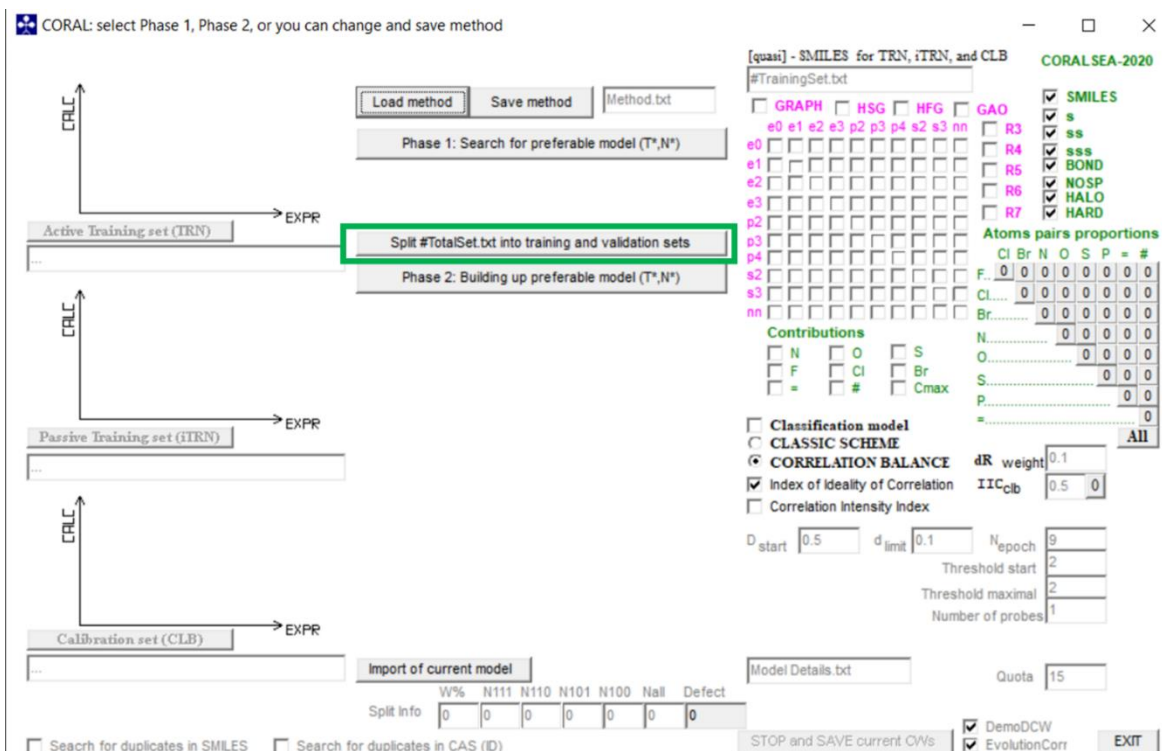
From the generated SMILES attributes, a complete list of compounds containing **training**, **invisible training**, **calibration** and **validation sets** for different splits was prepared to run in the CORAL software.

Name	Date modified	Type	Size
Model	18-05-2022 06:23	File folder	
Search	18-05-2022 06:17	File folder	
#CheckUp	19-05-2022 11:31	Text Document	4 KB
#Recommended T2 and N7	19-05-2022 11:31	Text Document	1 KB
#TotalSet	14-05-2022 12:22	Text Document	52 KB
#TotalSet'	18-05-2022 15:24	Text Document	5 KB
#TrainingSet	18-05-2022 15:24	Text Document	4 KB
#ValidationSet	18-05-2022 15:24	Text Document	2 KB
AOs	14-05-2022 12:22	Text Document	6 KB
comment	14-05-2022 12:22	Microsoft Edge PD...	1,084 KB
CORALSEA	14-05-2022 12:22	Application	728 KB
Dataset	19-05-2022 11:35	Text Document	6 KB
METHOD	19-05-2022 11:31	Text Document	1 KB
Structural Chemistry 2020 31,1441-1448	14-05-2022 12:22	Microsoft Edge PD...	1,030 KB
VectorDistribution	14-05-2022 12:22	Text Document	1 KB

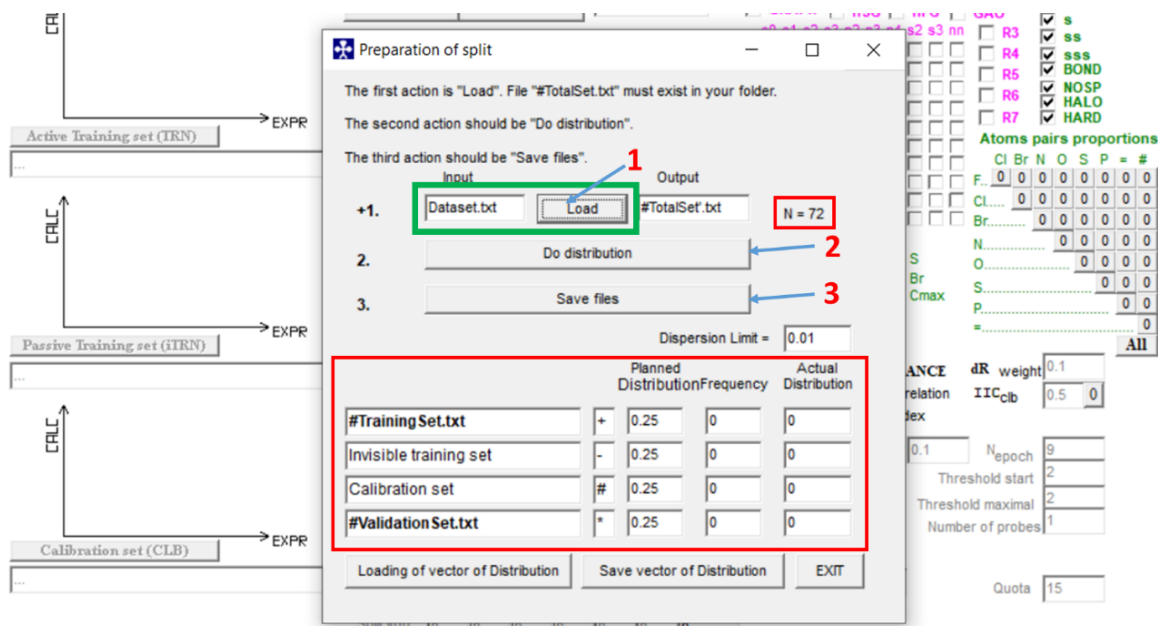
Click on the **CORALSEA** icon and a floating window will appear

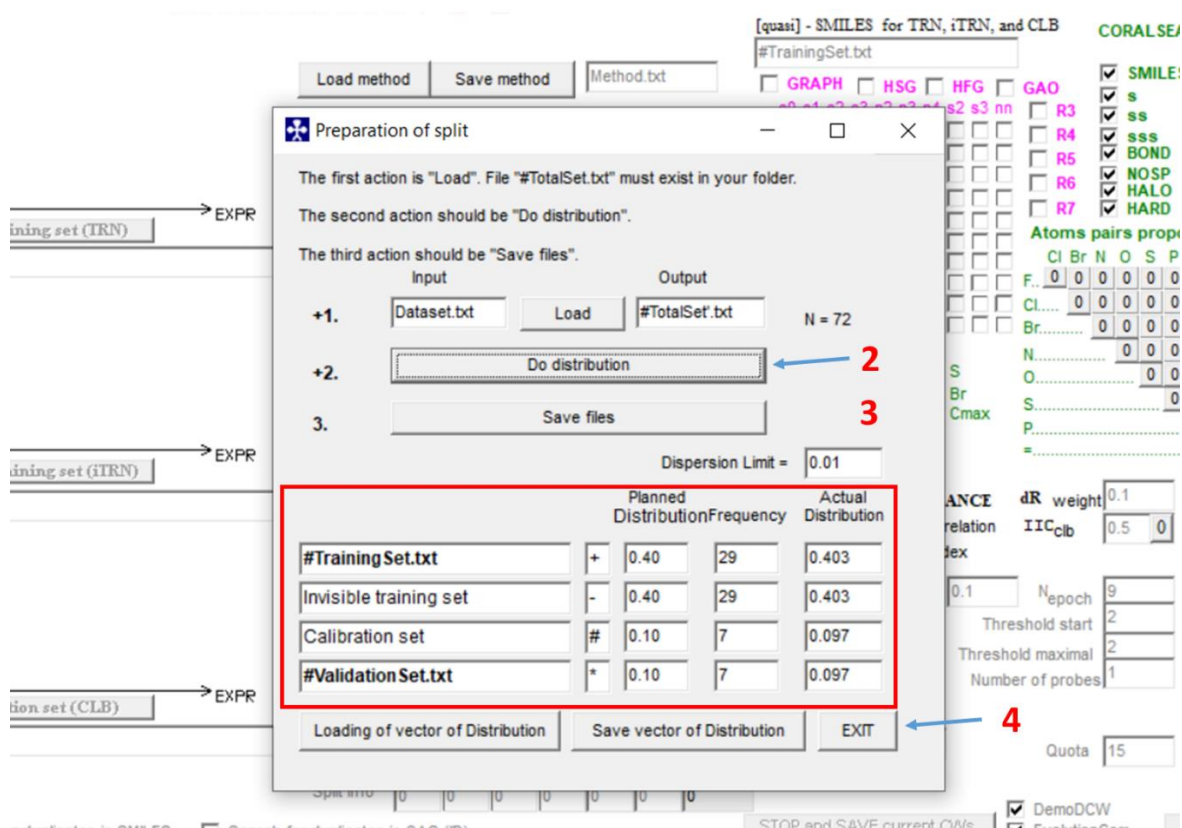


First we have to split our dataset for that click on Split into training and validation set icon, a popup will appear



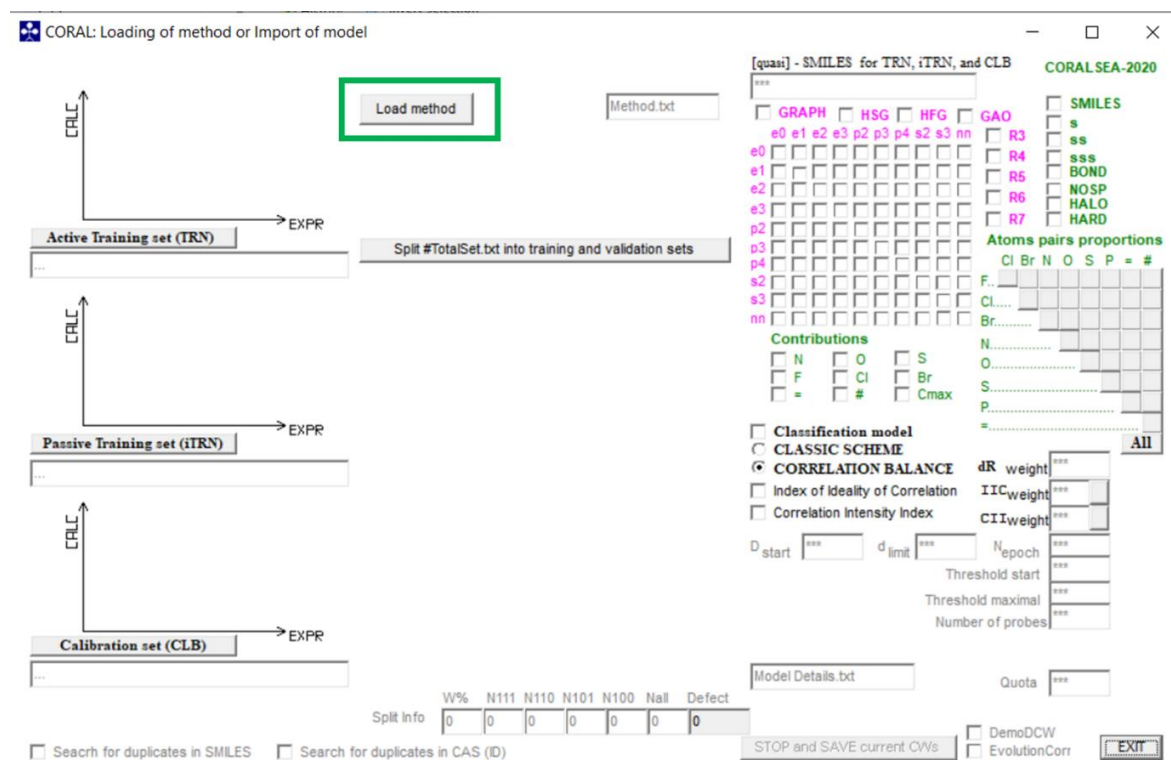
- In the input section enter the name of your text file containing the SMILES format along with its biological endpoint.
- Then press load icon and total number of compounds will be displayed on the right side.
- Manually adjust the percentage distribution among your split.
- Click on do distribution.
- Finally click on save files.





Step 3

Search for preferable model



Click on the **load method**, here in this tutorial we are going to run the **SMILES** based QSAR,

so we will tick on all the **SMILES** attributes, then set the **Nepoch** value, the **threshold** value range and the number of **probes**. Once done click on **save method**.

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

Load method Save method Method.txt

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

Split #TotalSet.txt into training and validation sets

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

Active Training set (TRN) → EXPR

Passive Training set (ITRN) → EXPR

Calibration set (CLB) → EXPR

Import of current model

Model Details.txt

Quota 15

STOP and SAVE current CWs

DemocDW EvolutionCorr EXIT

[quasi] - SMILES for TRN, iTRN, and CLB

#TrainingSet.txt

GRAPH HSG HFG GAO

e0 e1 e2 e3 p2 p3 p4 s2 s3 nn

R3 R4 R5 R6 R7

SMILES

ss

sss

BOND

NOSP

HALO

HARD

Atoms pairs proportions

	Cl	Br	N	O	S	P	#
F	0	0	0	0	0	0	0
Cl	0	0	0	0	0	0	0
Br	0	0	0	0	0	0	0
N	0	0	0	0	0	0	0
O	0	0	0	0	0	0	0
S	0	0	0	0	0	0	0
P	0	0	0	0	0	0	0
#	0	0	0	0	0	0	0

Contributions

N O S

F Cl Br

= # Cmax

Classification model

CLASSIC SCHEME

CORRELATION BALANCE

Index of Ideality of Correlation

Correlation Intensity Index

d start 0.5 d limit 0.1 Nepoch 30

Threshold start 1

Threshold maximal 5

Number of probes 3

W% N111 N110 N101 N100 Nall Defect

Split Info 0 0 0 0 0 0 0

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

Now we will search for the preferable model the best Threshold and Nepoch value. Click on the icon **Phase 1: search for preferable model**. A popup will appear, click on **yes** and the search should start.

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

Load method Save method Method.txt

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

Split #TotalSet.txt into training and validation sets

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

Active Training set (TRN) → EXPR

Passive Training set (ITRN) → EXPR

Calibration set (CLB) → EXPR

Import of current model

Model Details.txt

Quota 15

STOP and SAVE current CWs

DemocDW EvolutionCorr EXIT

[quasi] - SMILES for TRN, iTRN, and CLB

#TrainingSet.txt

GRAPH HSG HFG GAO

e0 e1 e2 e3 p2 p3 p4 s2 s3 nn

R3 R4 R5 R6 R7

SMILES

ss

sss

BOND

NOSP

HALO

HARD

Atoms pairs proportions

	Cl	Br	N	O	S	P	#
F	0	0	0	0	0	0	0
Cl	0	0	0	0	0	0	0
Br	0	0	0	0	0	0	0
N	0	0	0	0	0	0	0
O	0	0	0	0	0	0	0
S	0	0	0	0	0	0	0
P	0	0	0	0	0	0	0
#	0	0	0	0	0	0	0

Contributions

N O S

F Cl Br

= # Cmax

Classification model

CLASSIC SCHEME

CORRELATION BALANCE

Index of Ideality of Correlation

Correlation Intensity Index

d start 0.5 d limit 0.1 Nepoch 30

Threshold start 1

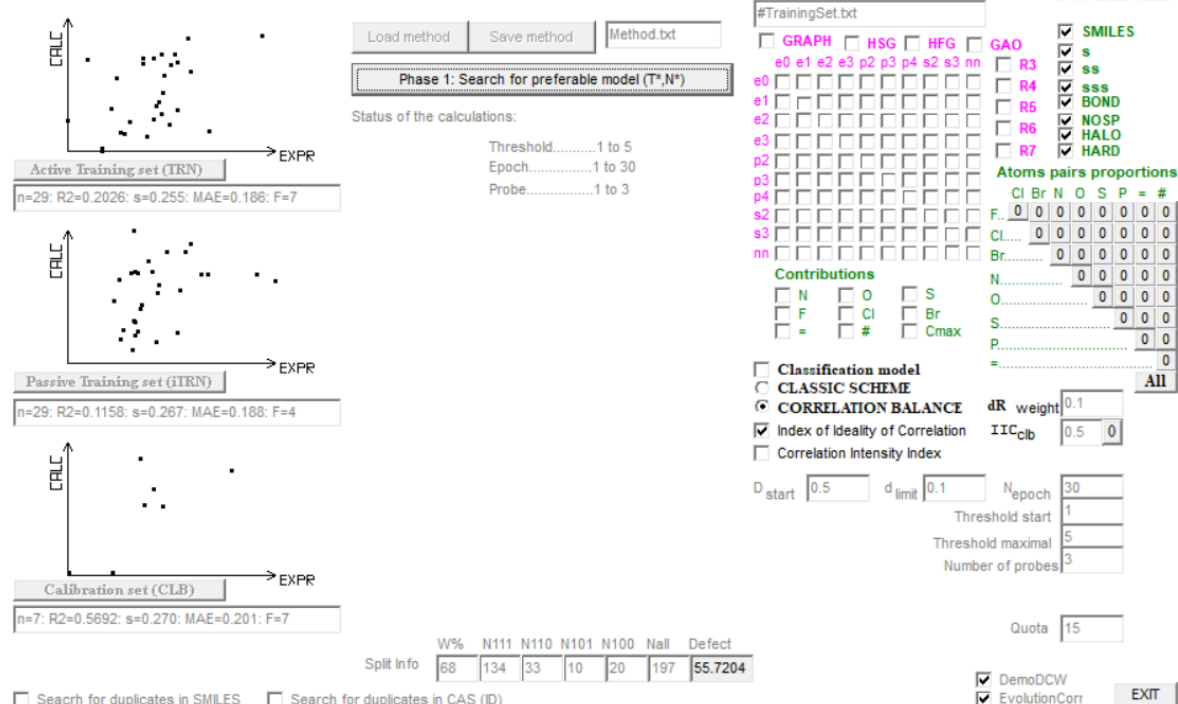
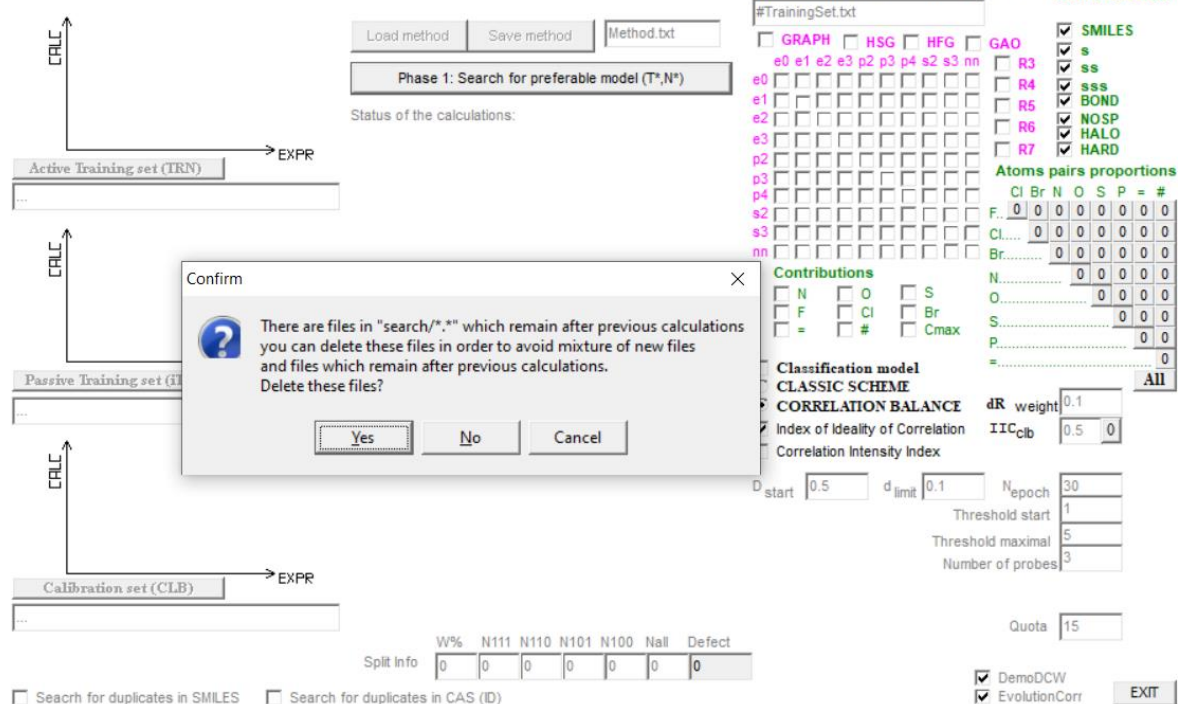
Threshold maximal 5

Number of probes 3

W% N111 N110 N101 N100 Nall Defect

Split Info 0 0 0 0 0 0 0

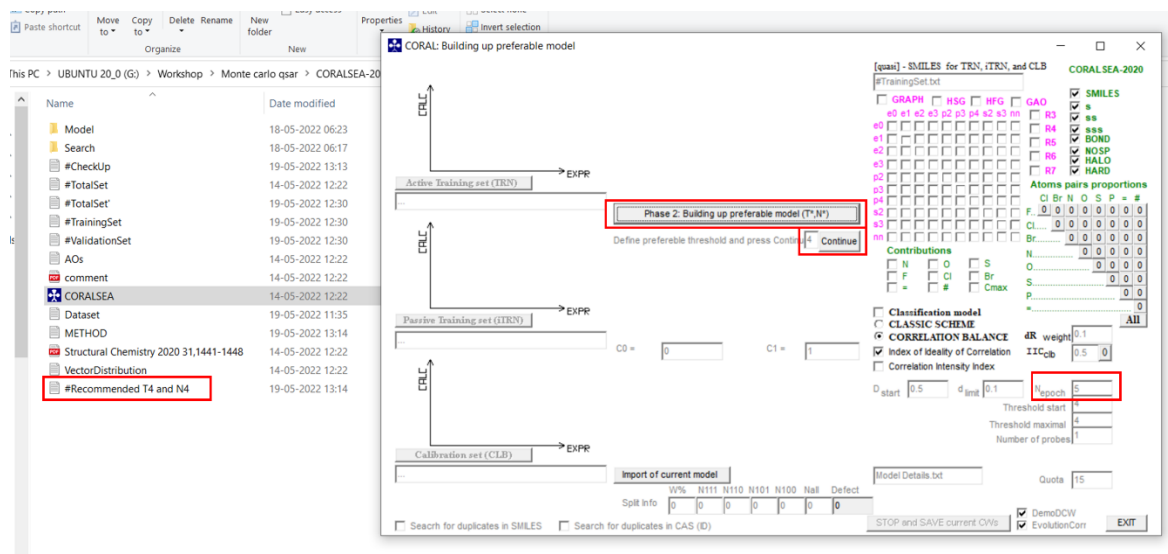
Search for duplicates in SMILES Search for duplicates in CAS (ID)



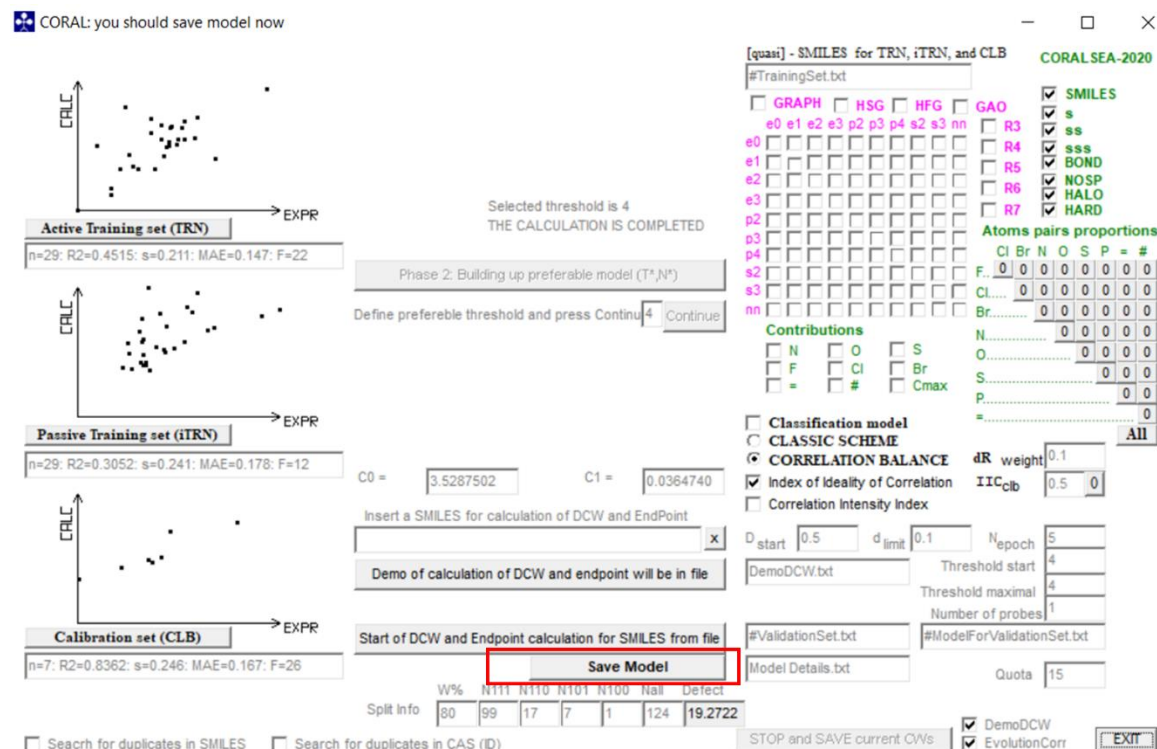
Once the search has been completed it will recommend the best threshold and Nepoch value with which we can now finally build our preferable model.

Step 4

Build up preferable model, set the recommended Nepoch and Threshold values and click on **Phase 2: Building up preferable model** and press **continue**.



Once the model is built press on **save model** at the bottom of the window.



Step 5

Validation of the model

Now again click on **load method** and press on **import of current model**. It will automatically load the last saved model and press on the **start of DCW and endpoint calculation for SMILES from the**. A popup will appear click on **yes** and the validation will be carried out.

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

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

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

Active Training set (TRN) → EXPR

Passive Training set (iTRN) → EXPR

Calibration set (CLB) → EXPR

Import of current model

Model Details.txt

Quota: 15

STOP and SAVE current CWs

DemoDCW EvolutionCorr

EXIT

CORAL: Calculation of model for external substances

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

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

Active Training set (TRN) → EXPR

Passive Training set (iTRN) → EXPR

Calibration set (CLB) → EXPR

Calculation Endpoint for dark SMILES

list.txt

Endpoint.txt

DCW(4,5)=

EndPoint =

Start of DCW and Endpoint calculation for SMILES from file

Import of current model

Model Details.txt

Quota: 15

Continue optimization

STOP and SAVE current CWs

DemoDCW EvolutionCorr

EXIT

Please wait! You will be informed when calculations will be completed...

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

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

Active Training set (TRN)
n=29: R2=0.4515: s=0.211: MAE=0.147: F=22

Passive Training set (iTRN)
n=29: R2=0.3052: s=0.241: MAE=0.178: F=12

Calibration set (CLB)
n=7: R2=0.8362: s=0.246: MAE=0.167: F=26

DEMO plots "EXPR vs. CALC"

Confirm

There are some SMILES (n=1) that present simultaneously in VLD and TRN, iTRN, or CLB
Do you want remove from VLD, SMILES that present in TRN, iTRN, or CLB?

Yes No

Insert a SMILES for calculation of DCW and EndPoint

Demo of calculation of DCW and endpoint will be in file

DCW(4,5)= EndPoint =

Start of DCW and Endpoint calculation for SMILES from file

Import of current model

W% N111 N110 N101 N100 Nall Defect

Split Info 80 99 17 7 1 124 19.2722

Search for duplicates in SMILES Search for duplicates in CAS (ID) Continue optimization STOP and SAVE current CWs DemoDCW EvolutionCorr EXIT

[quasi] - SMILES for TRN, iTRN, and CLB

#TrainingSet.txt

GRAPH HSG HFG GAO

e0 e1 e2 e3 p2 p3 p4 s2 s3 nn

R3 R4 R5 R6 R7

SMILES

s

ss

sss

BOND

NOSP

HALO

HARD

Atoms pairs proportions

Cl Br N O S P = #

F. 0 0 0 0 0 0

Cl. 0 0 0 0 0 0

Br. 0 0 0 0 0 0

N. 0 0 0 0 0 0

O. 0 0 0 0 0 0

S. 0 0 0 0 0 0

P. 0 0 0 0 0 0

= 0

Contributions

N O S

F Cl Br

= # Cmax

Classification model

CLASSIC SCHEME

CORRELATION BALANCE

Index of Ideality of Correlation

Correlation Intensity Index

dR weight 0.1

IIC_{clb} 0.5 0

D start 0.5 d limit 0.1 Nepoch 5

DemoDCW.txt Threshold start 4

Threshold maximal 4

Number of probes 1

#ValidationSet.txt #ModelForValidationSet.txt

Model Details.txt Quota 15

Please see results for validation set in file .../model/#ModelForValidationSet.txt; Now you can study plots "expr vs calc"

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

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

Active Training set (TRN)
n=29: R2=0.4515: s=0.211: MAE=0.147: F=22

Passive Training set (iTRN)
n=29: R2=0.3052: s=0.241: MAE=0.178: F=12

Calibration set (CLB)
n=7: R2=0.8362: s=0.246: MAE=0.167: F=26

Validation set (VLD)
n=6: R2=0.8370: s=0.211: MAE=0.168: F=21

DEMO plots "EXPR vs. CALC"

Calculation Endpoint for dark SMILES

list.txt Endpoint.txt

C0 = 3.5287502 C1 = 0.0364740

Insert a SMILES for calculation of DCW and EndPoint

c1ccc2c(c1)c1c[nH]2)c(nc(c1)C(c1c[nH]c2c1cc(cc2))

Demo of calculation of DCW and endpoint will be in file

DCW(4,5)= 34.9307355 EndPoint = 4.8028138

Start of DCW and Endpoint calculation for SMILES from file

Import of current model

W% N111 N110 N101 N100 Nall Defect

Split Info 80 99 17 7 1 124 19.2722

Search for duplicates in SMILES Search for duplicates in CAS (ID) Continue optimization STOP and SAVE current CWs DemoDCW EvolutionCorr EXIT

[quasi] - SMILES for TRN, iTRN, and CLB

#TrainingSet.txt

GRAPH HSG HFG GAO

e0 e1 e2 e3 p2 p3 p4 s2 s3 nn

R3 R4 R5 R6 R7

SMILES

s

ss

sss

BOND

NOSP

HALO

HARD

Atoms pairs proportions

Cl Br N O S P = #

F. 0 0 0 0 0 0

Cl. 0 0 0 0 0 0

Br. 0 0 0 0 0 0

N. 0 0 0 0 0 0

O. 0 0 0 0 0 0

S. 0 0 0 0 0 0

P. 0 0 0 0 0 0

= 0

Contributions

N O S

F Cl Br

= # Cmax

Classification model

CLASSIC SCHEME

CORRELATION BALANCE

Index of Ideality of Correlation

Correlation Intensity Index

dR weight 0.1

IIC_{clb} 0.5 0

D start 0.5 d limit 0.1 Nepoch 5

DemoDCW.txt Threshold start 4

Threshold maximal 4

Number of probes 1

#ValidationSet.txt #ModelForValidationSet.txt

Model Details.txt Quota 15

Step 6

Now we can have a look at our model in the folders

Three necessary files are to be looked at for interpreting the model quality and to get the structural attributes.

In the **Model** folder we will have file named with **m....** and another **#ModelForValidationSet\$** which will contain the statistical values for the model built.

```

File Edit Format View Help
r02 = 0.5730
rr02 = -1.6421
(r2-r02)/r2 = 0.0000 should be < 0.1 [1]
(r2-rr02)/r2 = 0.0000 should be < 0.1 [1]
k = 0.9861 should be 0.85 < k < 1.15 [1]
kk = 1.0120 should be 0.85 < kk < 1.15 [1]
R*m2(test) = 0.4072 should be > 0.5 [2]

Average Rm2 = -0.0365 should be larger 0.5 [3]
Delta Rm2 = 0.8874 should be lower 0.2 [3]

: n : R2 : CCC : IIC : CII : Q2 : Q2F1 : Q2F2 : Q2F3 : <Rm2> : RMSE : MAE : F
ActivTRN: 29: 0.4515: 0.6222: 0.6272: 0.6637: 0.3404: : : : : 0.211: 0.147: 22
Pass TRN: 29: 0.3052: 0.4768: 0.4325: 0.7884: 0.1648: : : : : 0.241: 0.178: 12
Calib : 7: 0.8362: 0.6144: 0.9144: 0.8295: 0.5963: 0.5529: 0.5326: 0.3391: -0.0365: 0.246: 0.167: 26

Training set is indicated by +;
Invisible training set is indicated by -;
Calibration set is indicated by #

Balance of correlations:
Active training set - Passive training set - Calibration set

DefectsSMILES should be less than 2 x Average Defect SMILES = 11.1779
:SMILES
+:c1c(ccc2c1/C(=N\O)/C(=N2)c1c[nH]c2c1cc(cc2)F)F : DCW(4,
+:c1c(ccc2c1/C(=N\O)/C(=N2)c1c[nH]c2c1cc(cc2)OC)OC : 34.751
+:c1c(ccc2c1/C(=N\O)/C(=N2)c1c[nH]c2c1cc(cc2)[N+](=O)[O-]) : 31.835
+:c1c(ccc2c1/C(=N\O)/C(=N2)c1c[nH]c2c1cc(cc2)[N+](=O)[O-] : 35.623
+:c1c(ccc2c1/C(=N\O)/C(=N2)c1c[nH]c2c1cc(cc2)C#N)C#N : 33.226
<
Ln 1, Col 1 10

```

```

File Edit Format View Help

The average of DefectsSMILES = 5.58097
Substance falls into domain of applicability if DefectsSMILES < 11.17794

Rm2(x,y) calculation for validation set from input file
n = 6
r2 = 0.8370
r02 = -0.2736
rr02 = 0.6713
(r2-r02)/r2 = 1.3269 should be < 0.1
(r2-rr02)/r2 = 0.1980 should be < 0.1
k = 1.0358 should be 0.85 < k < 1.15
kk = 0.9650 should be 0.85 < kk < 1.15
Rm2(test) = -0.0451 should be > 0.5

Rm2(y,x) calculation for validation set from input file
n = 6
r2 = 0.8370
r02 = 0.6713
rr02 = -0.2736
(r2-r02)/r2 = 0.1980 should be < 0.1
(r2-rr02)/r2 = 1.3269 should be < 0.1
k = 0.9650 should be 0.85 < k < 1.15
kk = 1.0358 should be 0.85 < kk < 1.15
R*m2(test) = 0.4963 should be > 0.5

Average Rm2 = 0.2256 should be larger 0.5
Delta Rm2 = 0.5414 should be lower 0.2
<

```

Structural attributes (SA) as promoters and hinderers observed from three CW (Probe)

In the Search folder based on our build model the threshold at which we have built it we select the file named with **S** and followed by the threshold number.

File Edit Format View Help

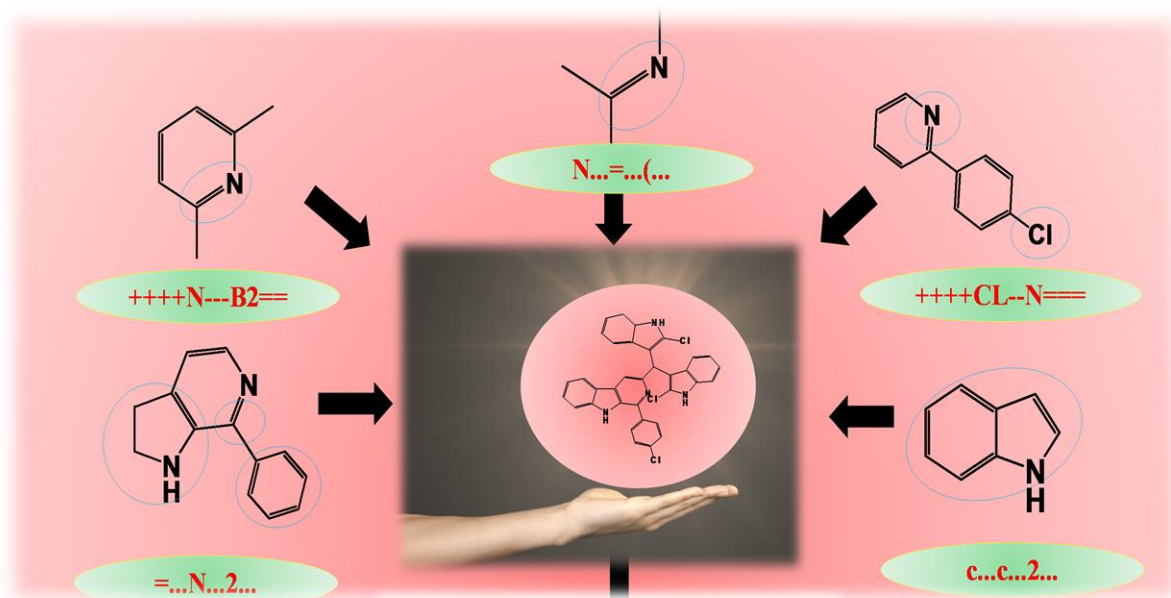
- if SA has CW(SA)>0 in all probes of the Monte Carlo optimization then the SA is a promoter of the Endpoint increase (List 1)
 - if SA has CW(SA)<0 in all probes of the Monte Carlo optimization then the SA is a promoter of the Endpoint decrease (List 2)
 - if SA has CW(SA)>0 together with CW(SA)<0 then 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.

N1, N2, and N3 are numbers of SMILES which contain SA in training, invisible training, and calibration sets, respectively

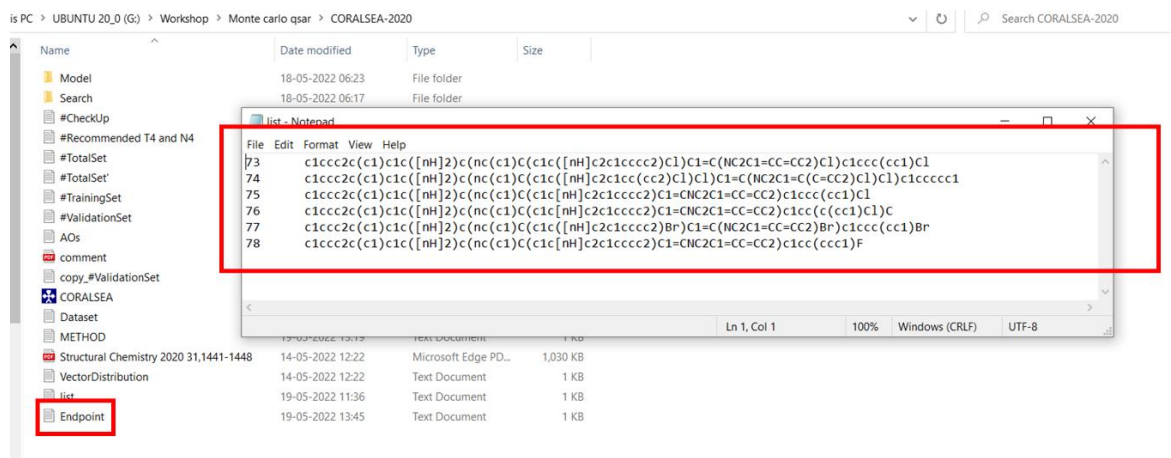
No. :	ID :	SAK :	Cws Probe 1:	Cws Probe 2:	Cws Probe 3:	N1 :	N2 :
1:	44:1...	C...(...	1.24445:	0.36804:	0.07570:	29:	2
2:	45:2...	(.....	0.31689:	0.34234:	1.32720:	29:	2
3:	46:2...	(.....	0.29469:	0.02236:	0.25451:	29:	2
4:	61:C...	(.....	0.04543:	0.19446:	0.14019:	29:	2
5:	178:C...	(.....	0.22257:	0.91174:	0.38270:	29:	2
6:	193:C...	1.....	0.21044:	0.68700:	0.16231:	29:	2
7:	197:C...	1...C...	1.32698:	1.09211:	0.97920:	29:	2
8:	213:C...	C.....	0.42792:	0.06191:	0.03506:	29:	2
9:	50:2...	C...1...	1.40285:	1.11276:	1.11470:	27:	2
10:	65:C...	(...=...	0.23156:	1.52108:	0.47396:	27:	2
11:	212:C...	C...C...	0.04339:	0.16583:	0.44328:	26:	2
12:	180:C...	C...2...	0.27588:	1.89279:	0.85304:	25:	2
13:	93:BOND10000000		0.92535:	1.99469:	0.03280:	24:	2
14:	36:1...	(.....	1.38737:	0.52129:	0.16250:	22:	2
15:	198:C...	C...2...	0.44176:	2.10340:	1.62225:	21:	2
16:	31:/...	(.....	0.46908:	1.16963:	0.34832:	17:	1
17:	32:/...	(.....	0.66990:	0.45952:	0.28837:	17:	1
18:	131:N...	C...2...	0.01815:	0.35684:	1.09866:	17:	1
19:	133:N...	C...=...	0.26905:	0.14346:	0.81527:	17:	1

Now based on these attributes obtained and the model build we can either design newer more potent molecules containing most of the important structural fragments or screen a library of compounds that have not yet been validated for the specific target and predict its activity in-silico.

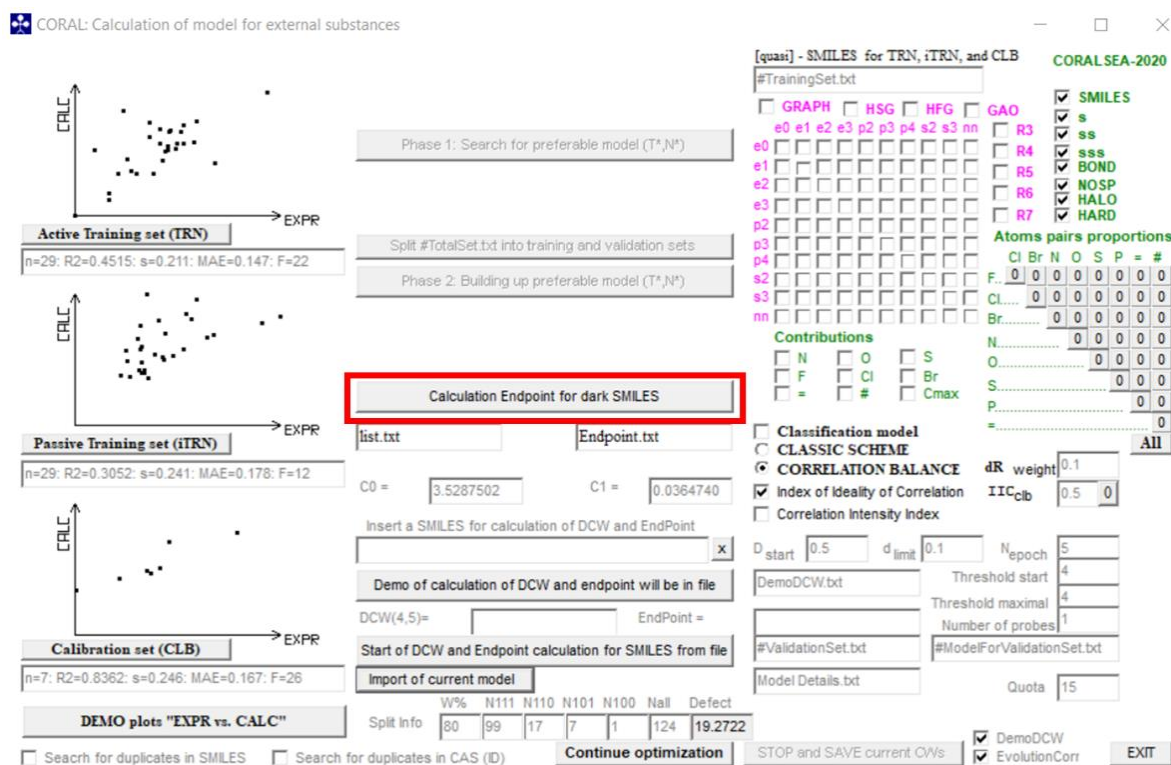


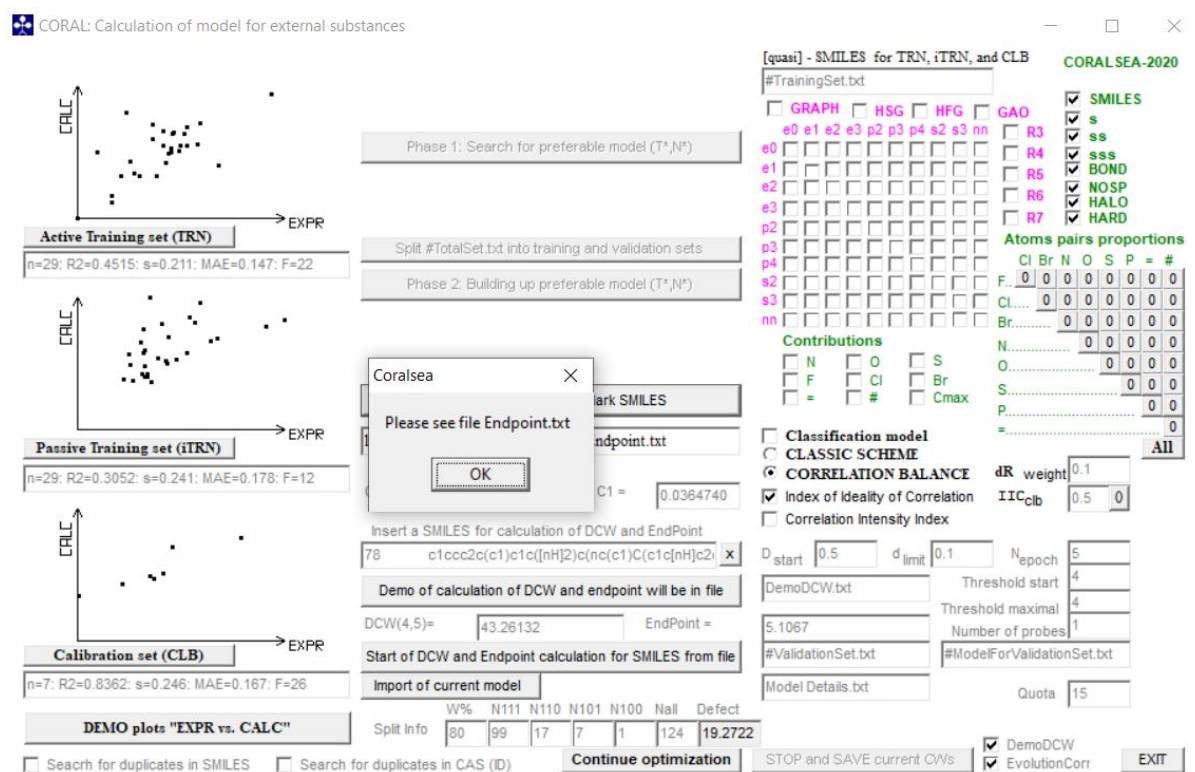
Activity prediction of unknown compounds

For the input we need to prepare a simple text file containing the serial number and the SMILES format.



Once we open the software click on **Load method** and then click on **import of current model**. Then click on the **Calculation Endpoint for dark SMILES**.





After completion it will give us the predicted biological endpoint as a text file named **Endpoint.txt**

