Monte Carlo Optimization

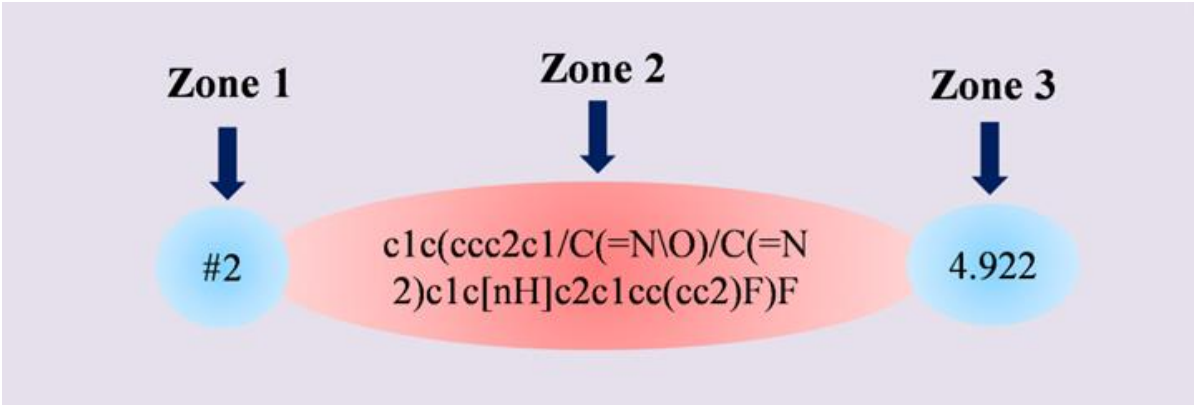
Acquisition of dataset

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.



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)Oc1ccccc1)Oc1ccccc1  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(c2)F)F  4.672
12      c1cc(cc2c1/C(=N\O)/C(=N2)c1c[nH]c2c1ccc(c2)C(=O)OC)C(=O)OC      4.586
13      c1ccc(c2c1/C(=N\O)/C(=N2)c1c[nH]c2c1cccc2C)C    4.756
14      c1(cccc2c1/C(=N\O)/C(=N2)c1c[nH]c2c1c(ccc2)C)C  4.348
15      c1(cccc2c1/C(=N\O)/C(=N2)c1c[nH]c2c1c(ccc2)C#N)C#N      4.473
16      c1cccc2c1/C(=N\OCc1ccccc1)/C(=N2)c1c[nH]c2c1cccc2       4.713
17      c1c(ccc2c1/C(=N\OCc1ccccc1)/C(=N2)c1c[nH]c2c1cc(cc2)F)F         4.977
18      c1c(ccc2c1/C(=N\OCc1ccccc1)/C(=N2)c1c[nH]c2c1cc(cc2)C#N)C#N     4.848
19      c1cc(cc2c1/C(=N\OCc1ccccc1)/C(=N2)c1c[nH]c2c1ccc(c2)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

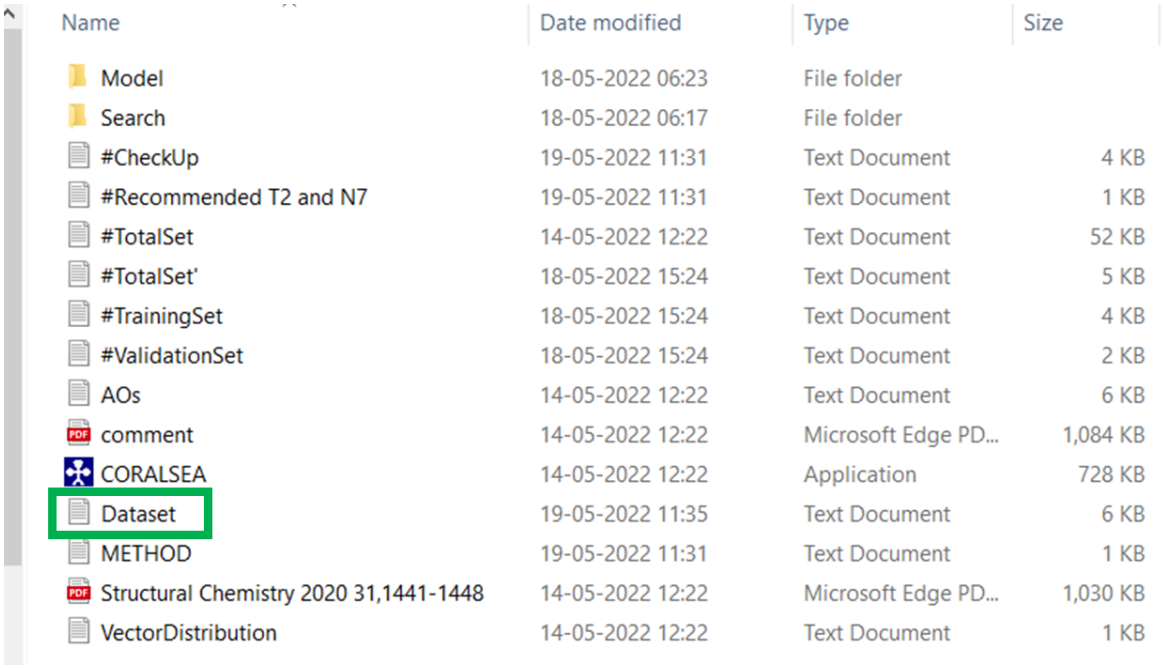
Software required

Coral (CORrelation and Logic)

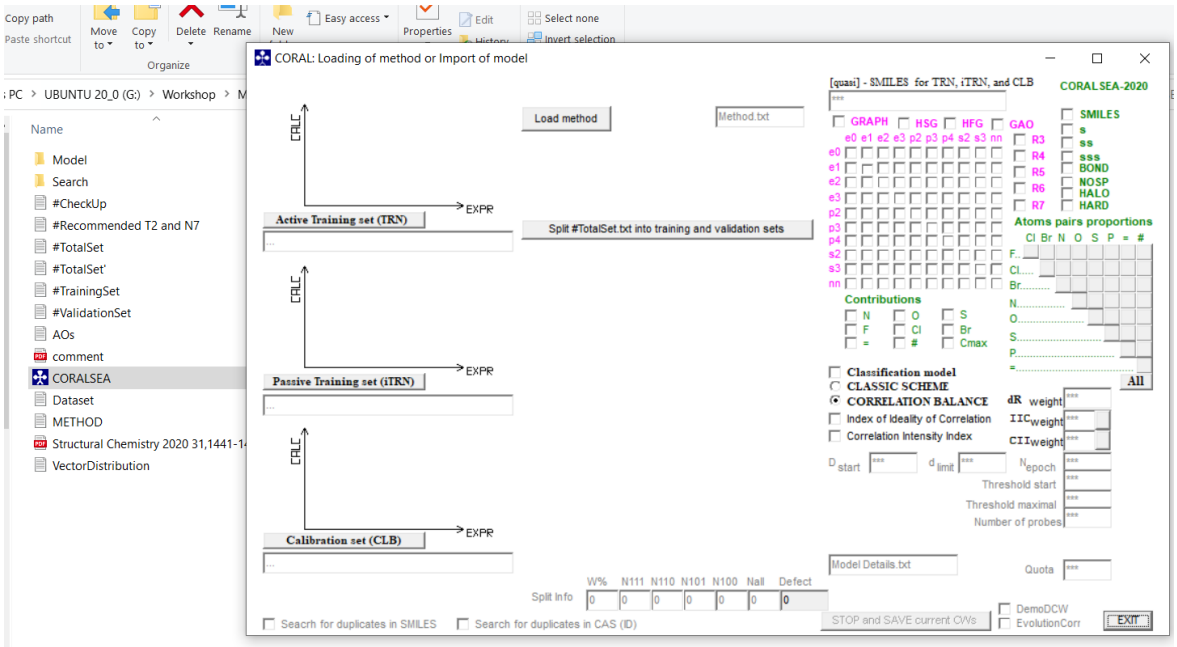
http://www.insilico.eu/coral/SOFTWARECORAL.html

Step 2

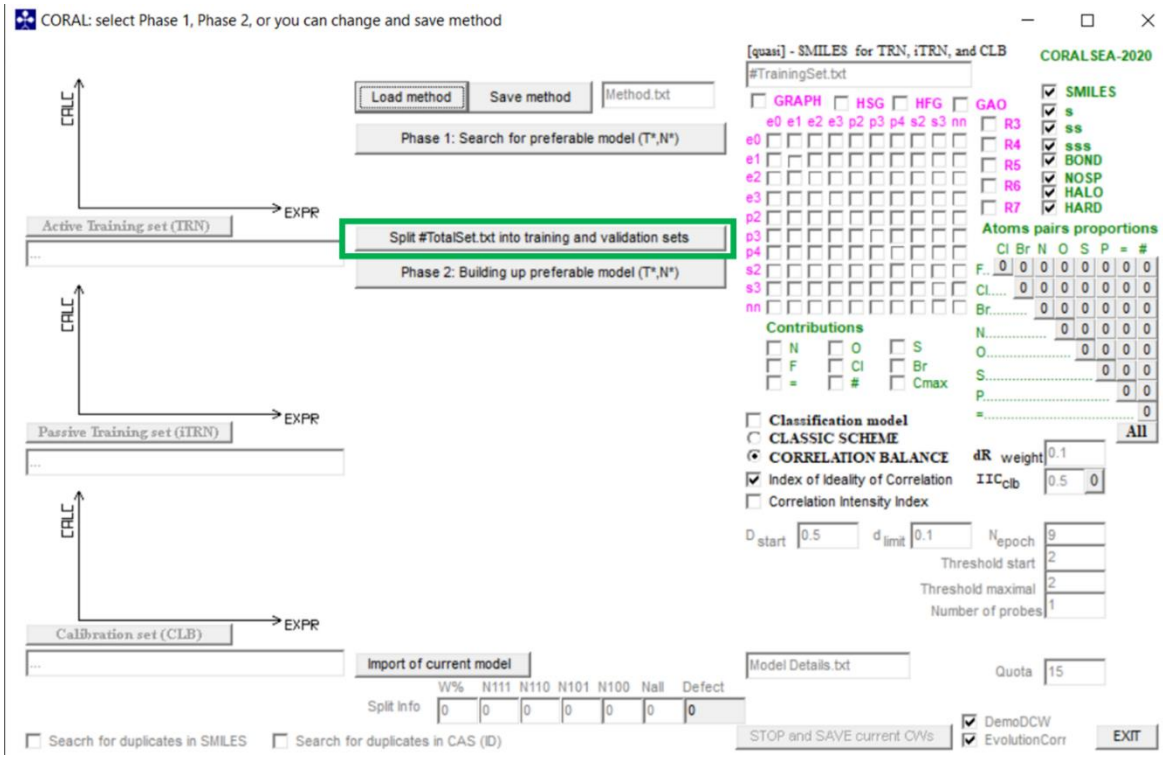
From the generated SMILES attributes, a complete list of compounds containing training, invisible training, calibration and validation sets for different splits was prepared to run in the CORAL software.



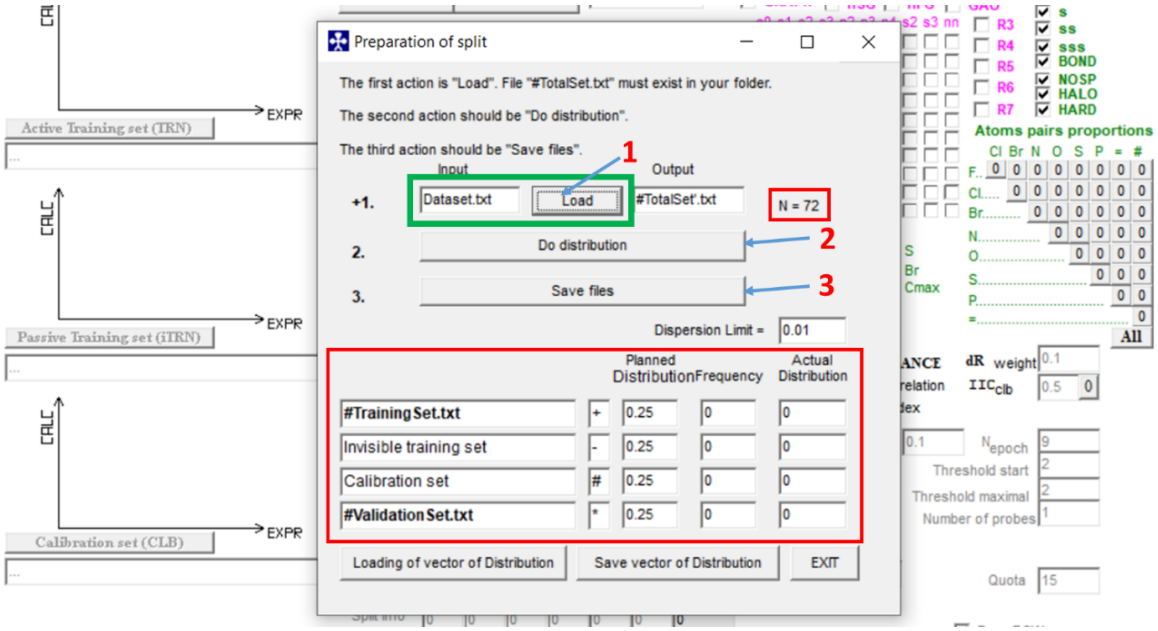
Click on the CORALSEA icon and a floating window will appear

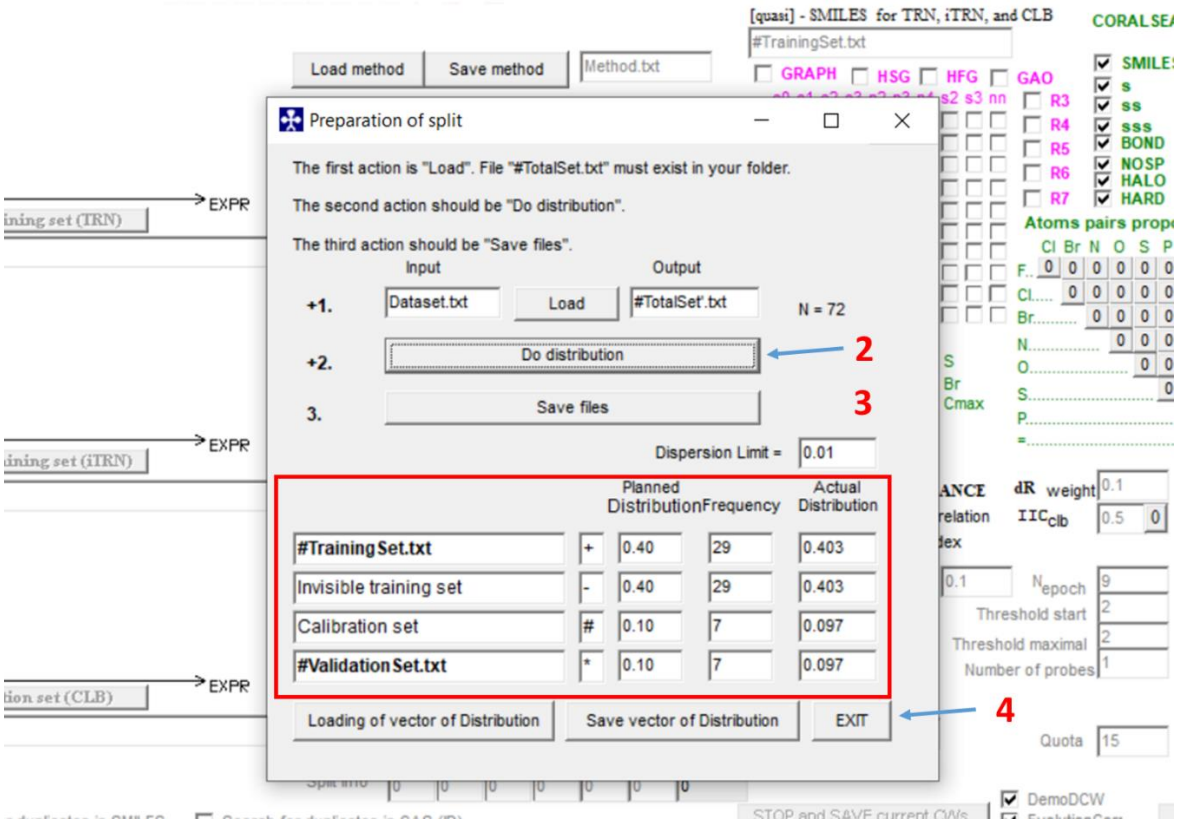


First we have to split our dataset for that click on Split into training and validation set icon, a popup will appear



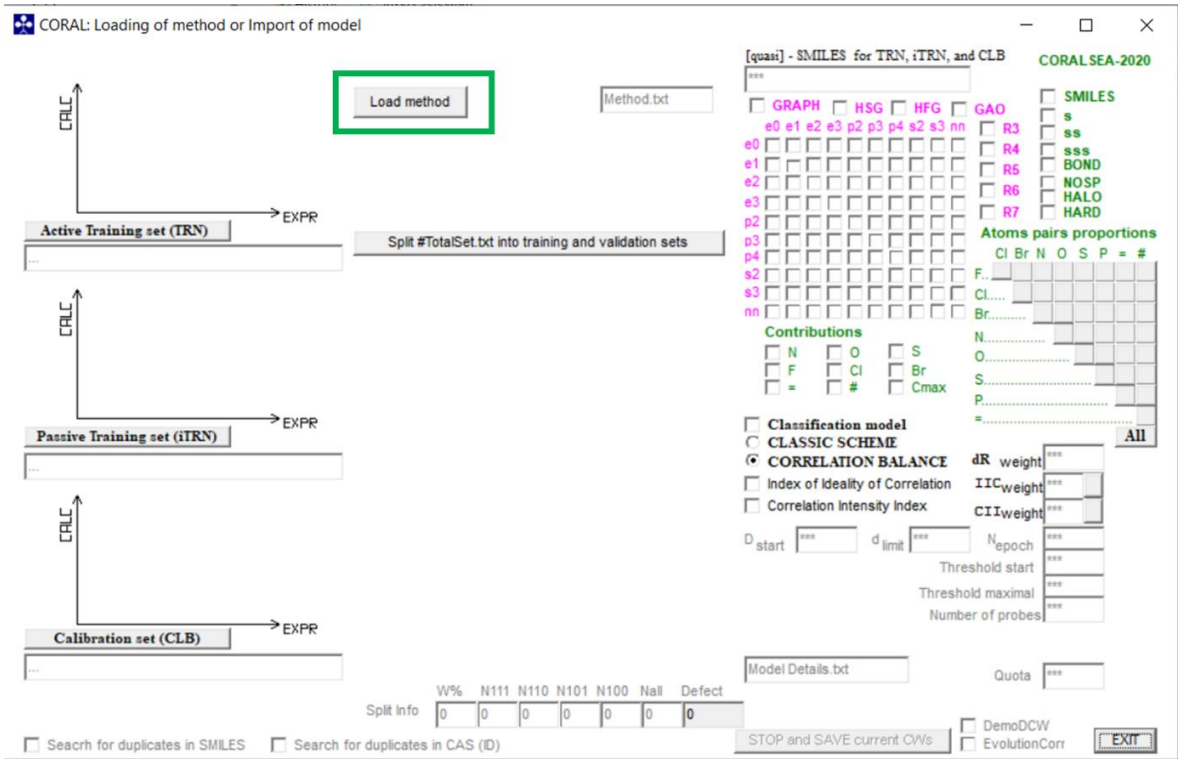
- In the input section enter the name of your text file containing the SMILES format along with its biological endpoint.
- Then press load icon and total number of compounds will be displayed on the right side.
- Manually adjust the percentage distribution among your split.
- Click on do distribution.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





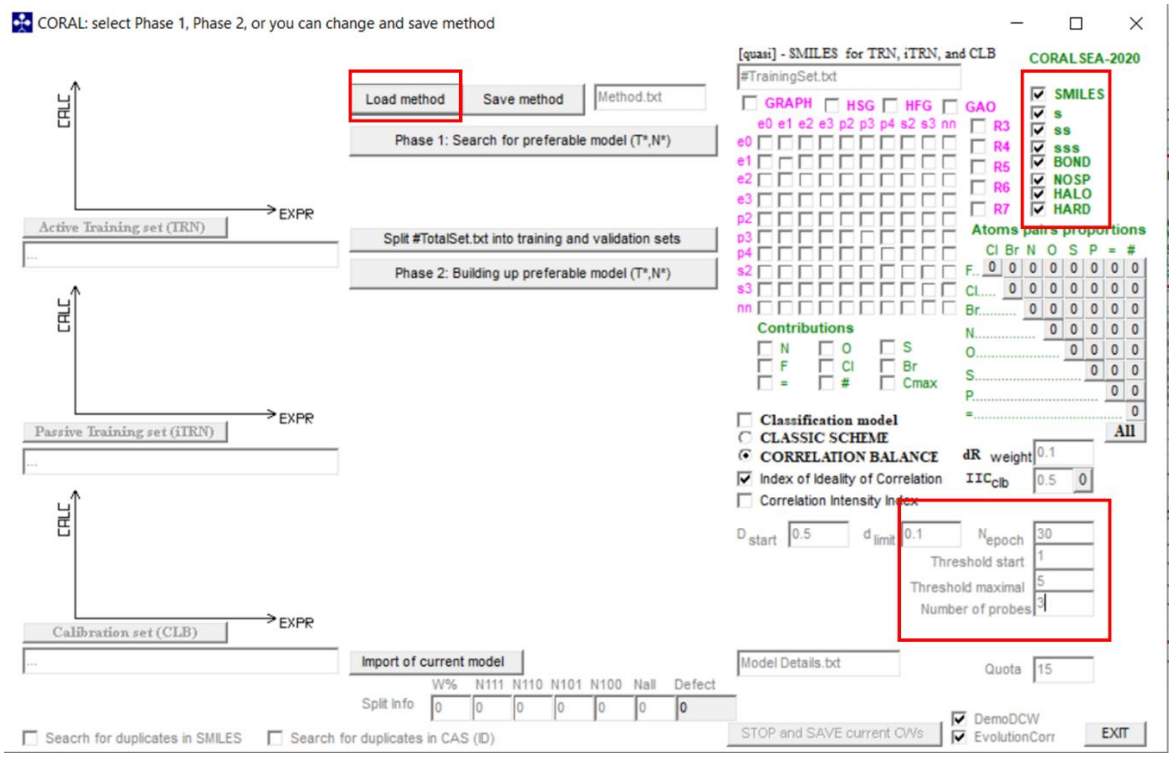
Step 3

Search for preferable model

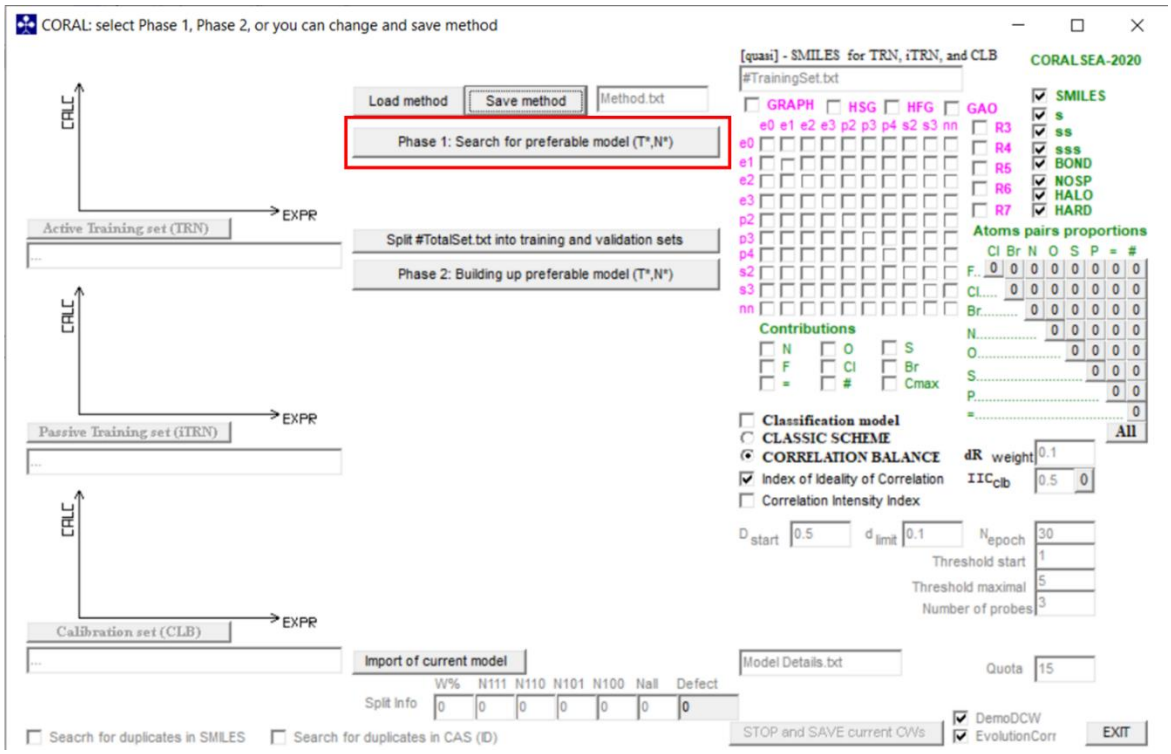


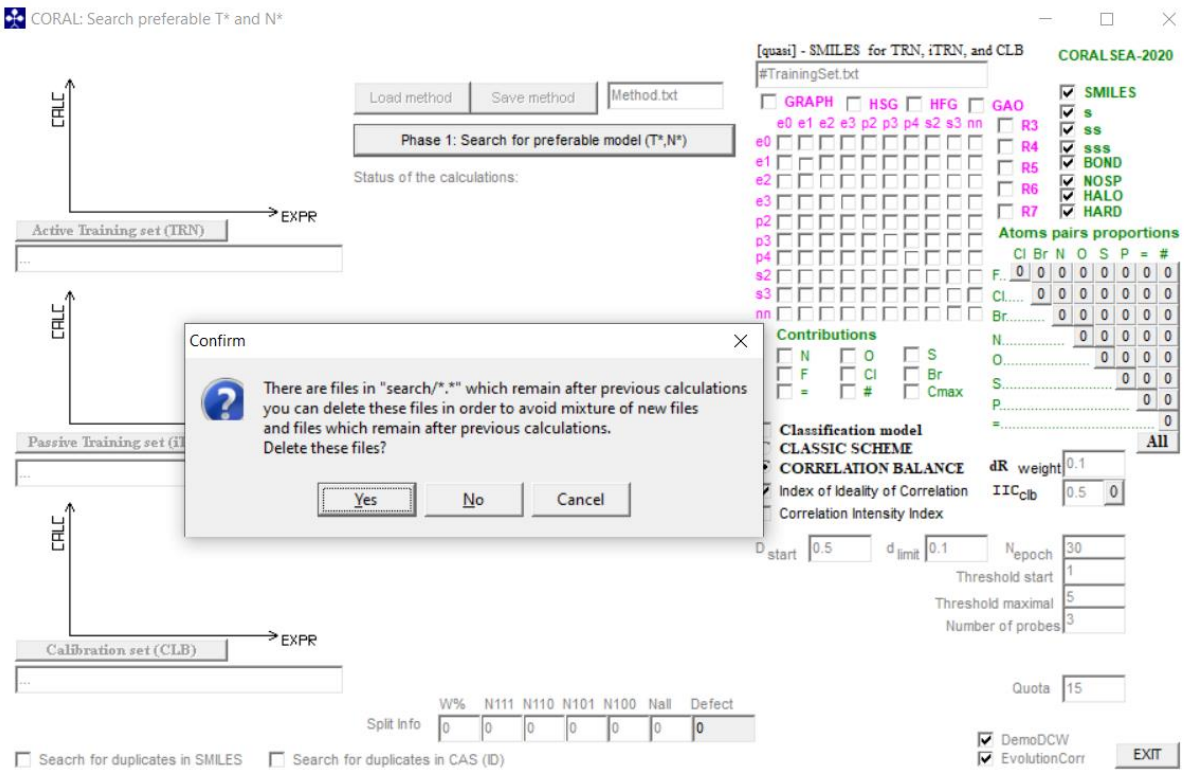
Click on the load method, here in this tutorial we are going to run the SMILES based QSAR,

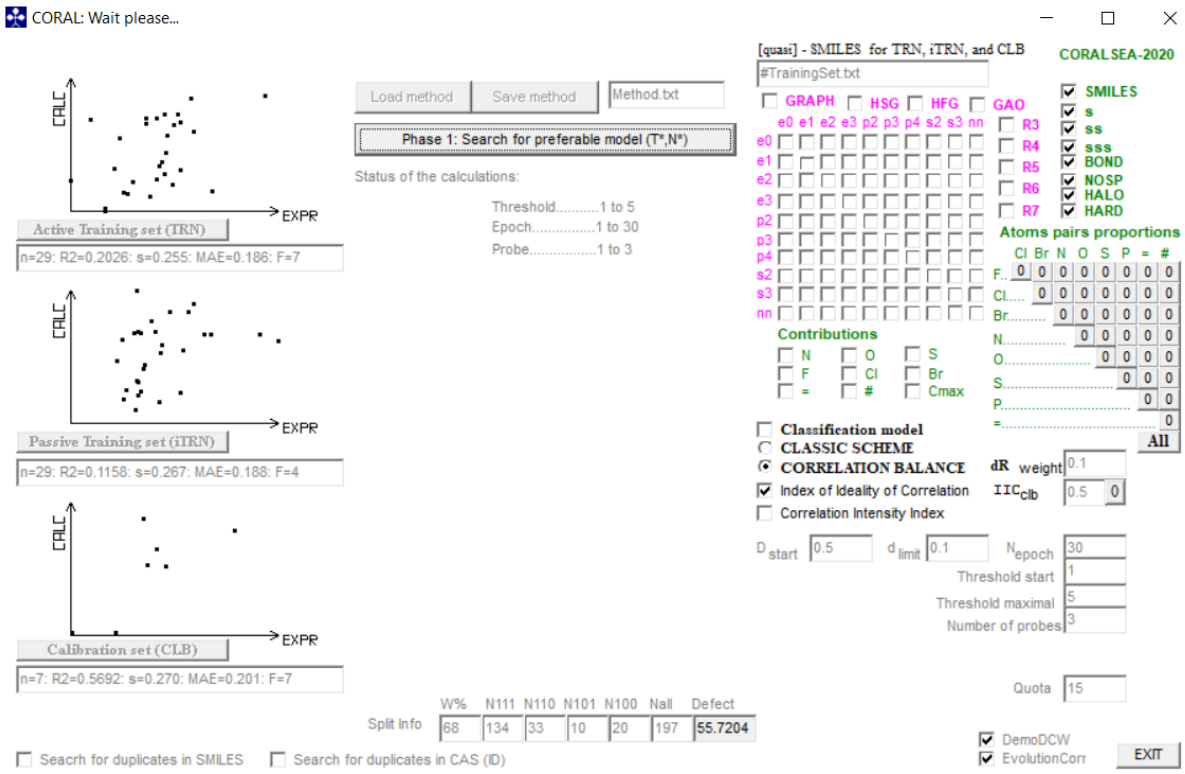
so we will tick on all the SMILES attributes, then set the Nepoch value, the threshold value range and the number of probes. Once done click on save method.



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.



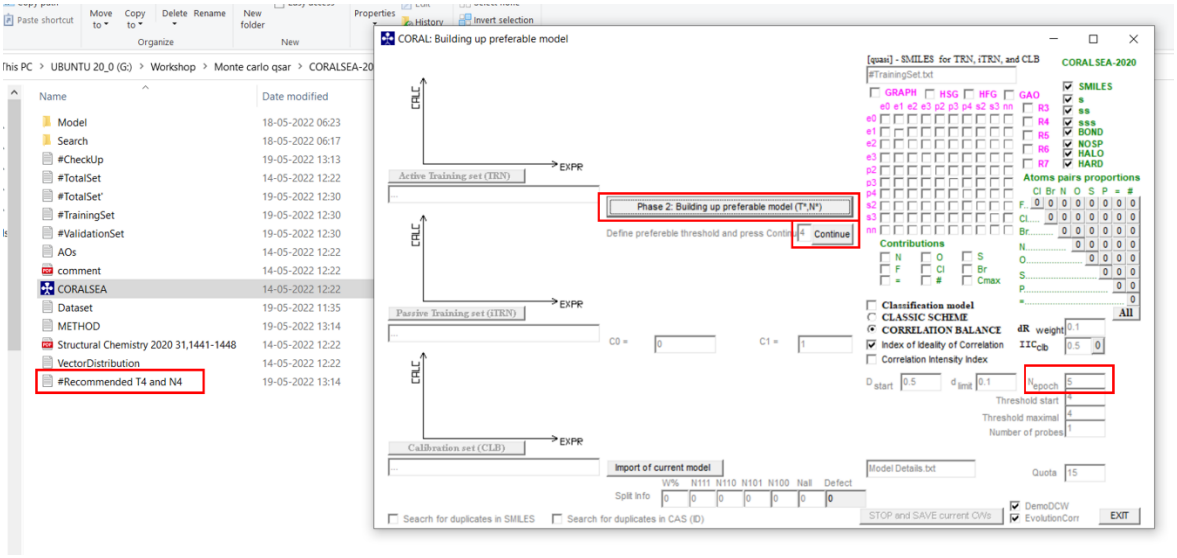




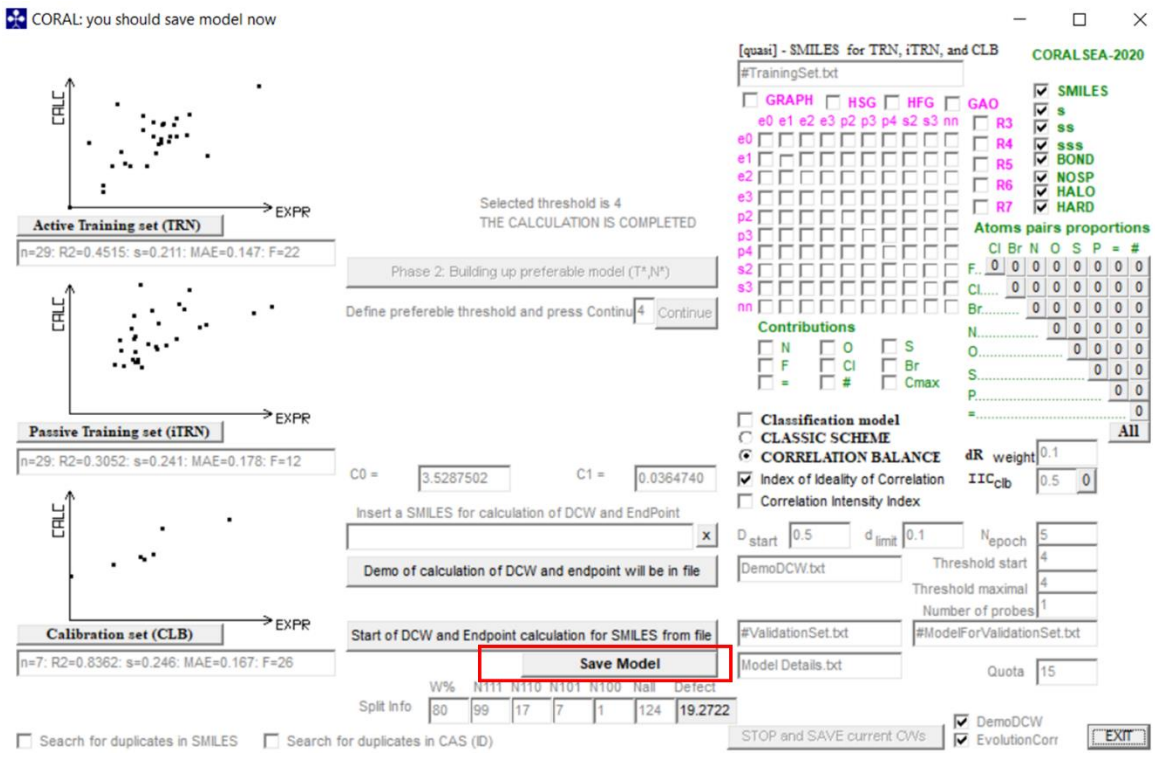
Once the search has been completed it will recommend the best threshold and Nepoch value with which we can now finally build our preferable model.

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.



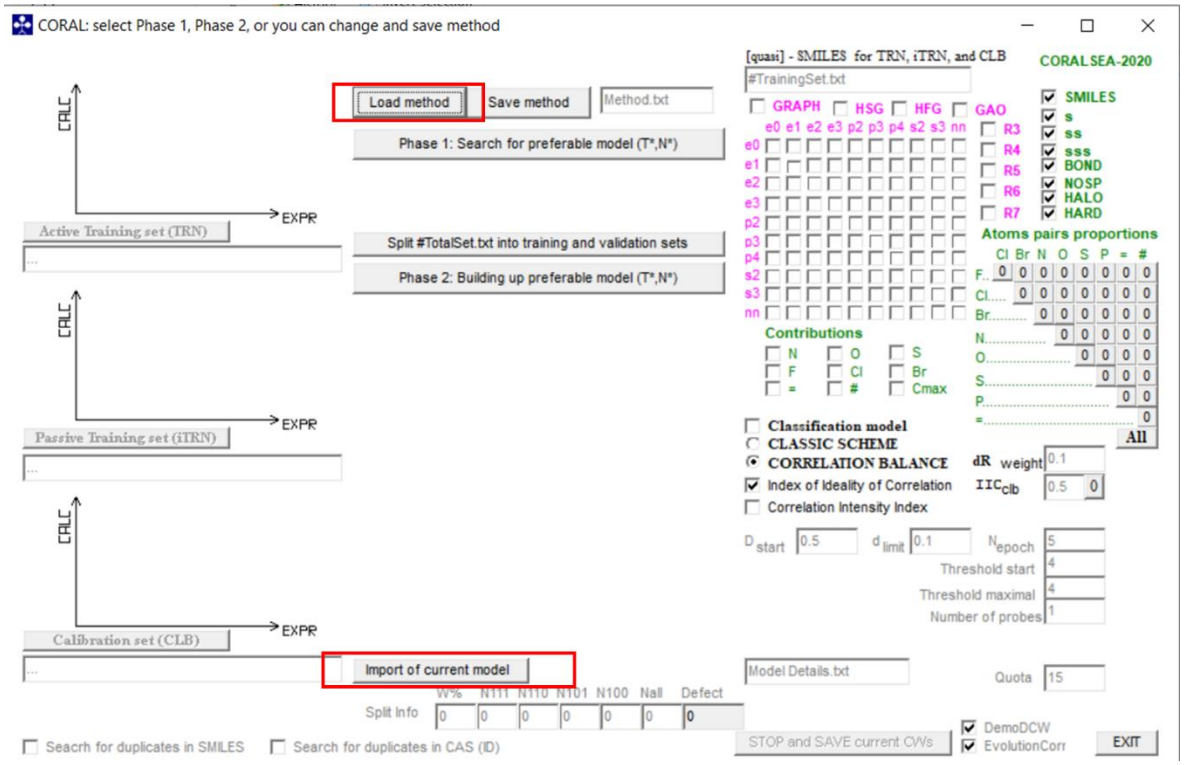
One the model is built press on save model at the bottom of the window.

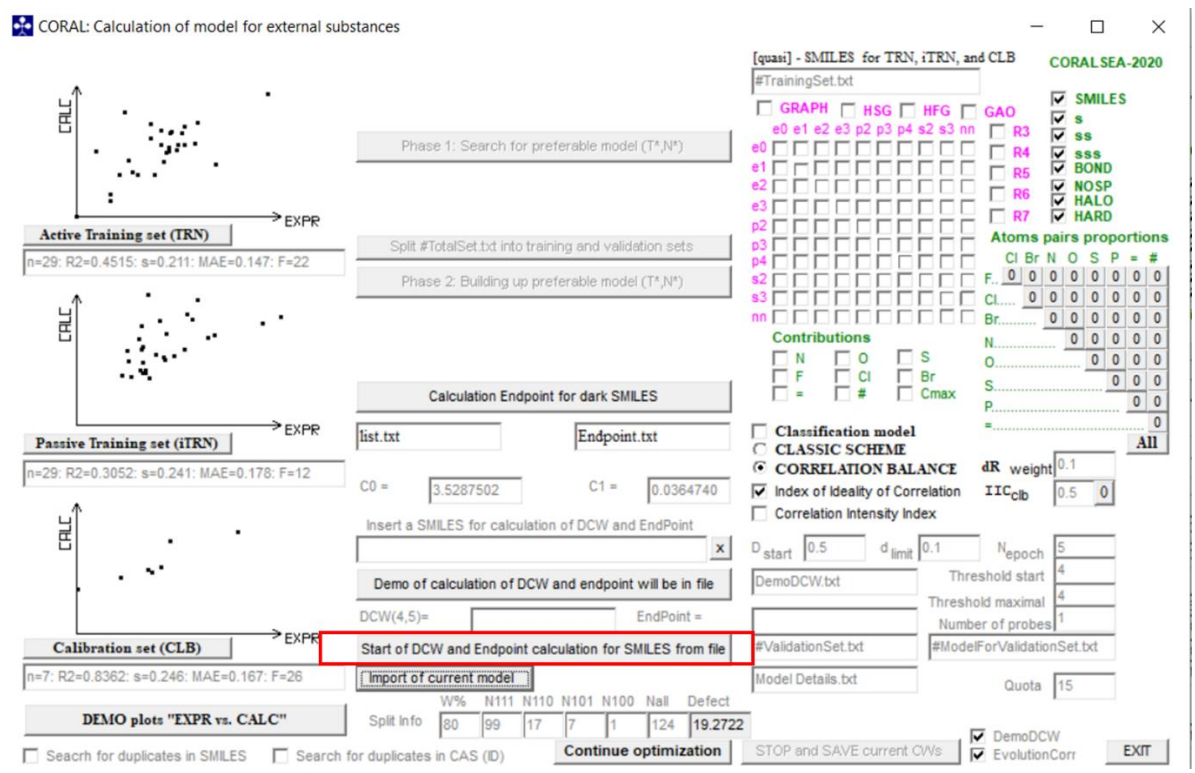


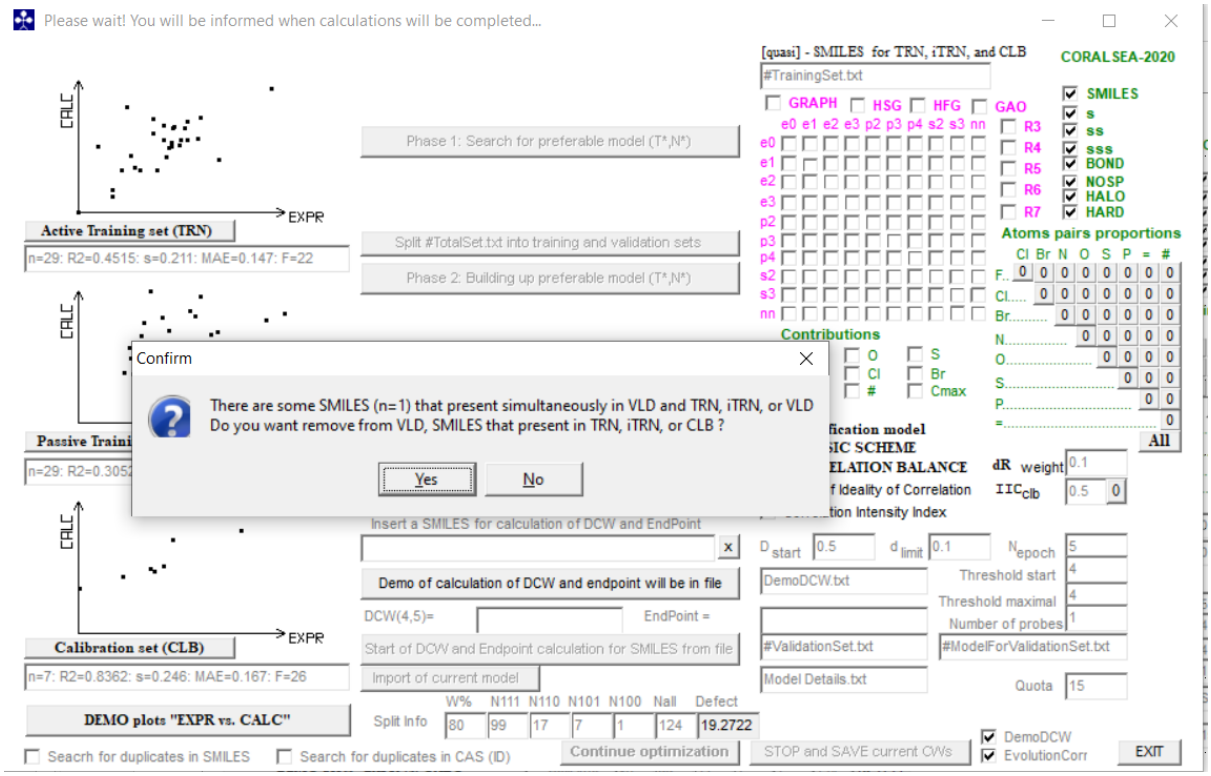
Step 5

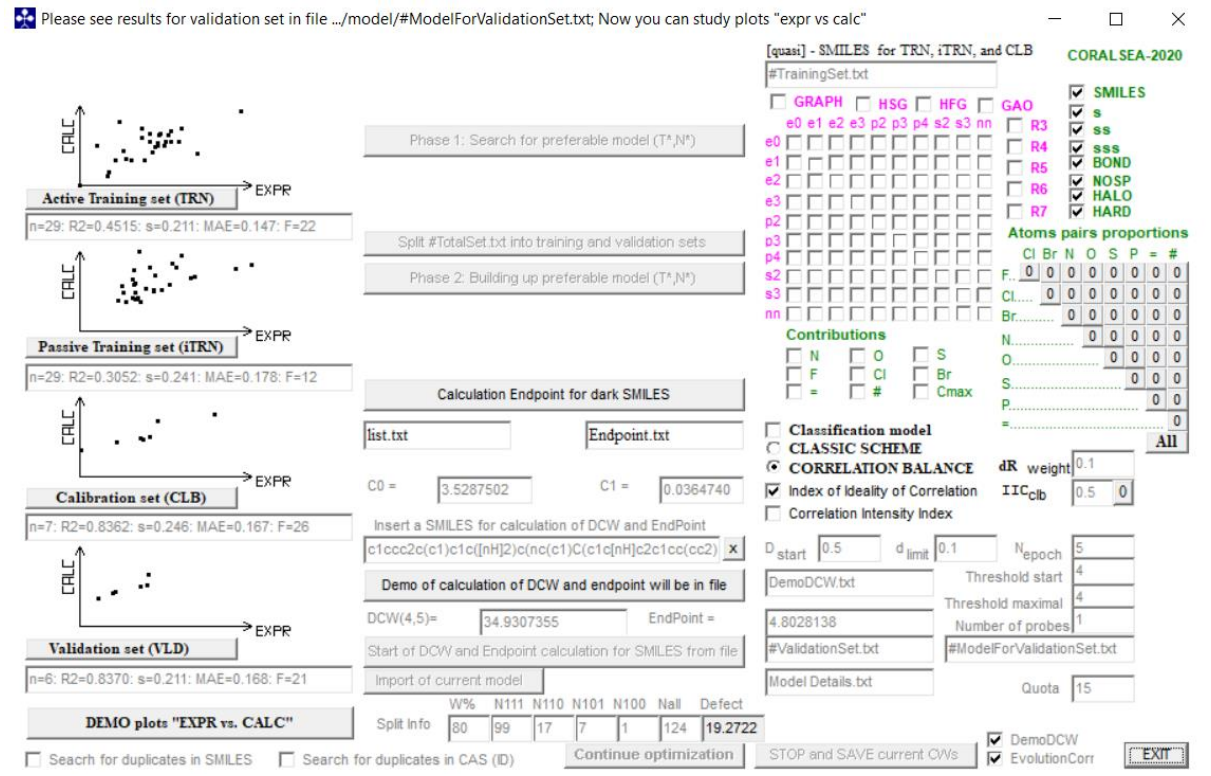
Validation of the model

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
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*)
Import of current model
Active Training set (TRN)
Passive Training set (iTRN)
Calibration set (CLB)
CALC
EXPR
[quasi] - SMILES for TRN, iTRN, and CLB
CORAL SEA-2020
#TrainingSet.txt
GRAPH
HSG
HFG
GAO
SMILES
Atoms pairs proportions
Contributions
Classification model
CLASSIC SCHEME
CORRELATION BALANCE
Index of Ideality of Correlation
Correlation Intensity Index
dR weight
Model Details.txt
Quota
Split Info
W% N111 N110 N101 N100 Nall Defect
Seacrh for duplicates in SMILES
Search for duplicates in CAS (ID)
STOP and SAVE current CWs
DemoDCW
EvolutionCorr
EXIT


CORAL: Calculation of model for external substances
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"
Phase 1: Search for preferable model (T*,N*)
Split #TotalSet.txt into training and validation sets
Phase 2: Building up preferable model (T*,N*)
Calculation Endpoint for dark SMILES
list.txt
Endpoint.txt
C0 =
3.5287502
C1 =
0.0364740
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
Continue optimization
CALC
EXPR
[quasi] - SMILES for TRN, iTRN, and CLB
CORAL SEA-2020
#TrainingSet.txt
Atoms pairs proportions
Contributions
Classification model
CLASSIC SCHEME
CORRELATION BALANCE
Index of Ideality of Correlation
Correlation Intensity Index
dR weight
DemoDCW.txt
#ValidationSet.txt
#ModelForValidationSet.txt
Model Details.txt
Quota
Split Info
W% N111 N110 N101 N100 Nall Defect
80 99 17 7 1 124 19.2722
Seacrh for duplicates in SMILES
Search for duplicates in CAS (ID)
STOP and SAVE current CWs
DemoDCW
EvolutionCorr
EXIT



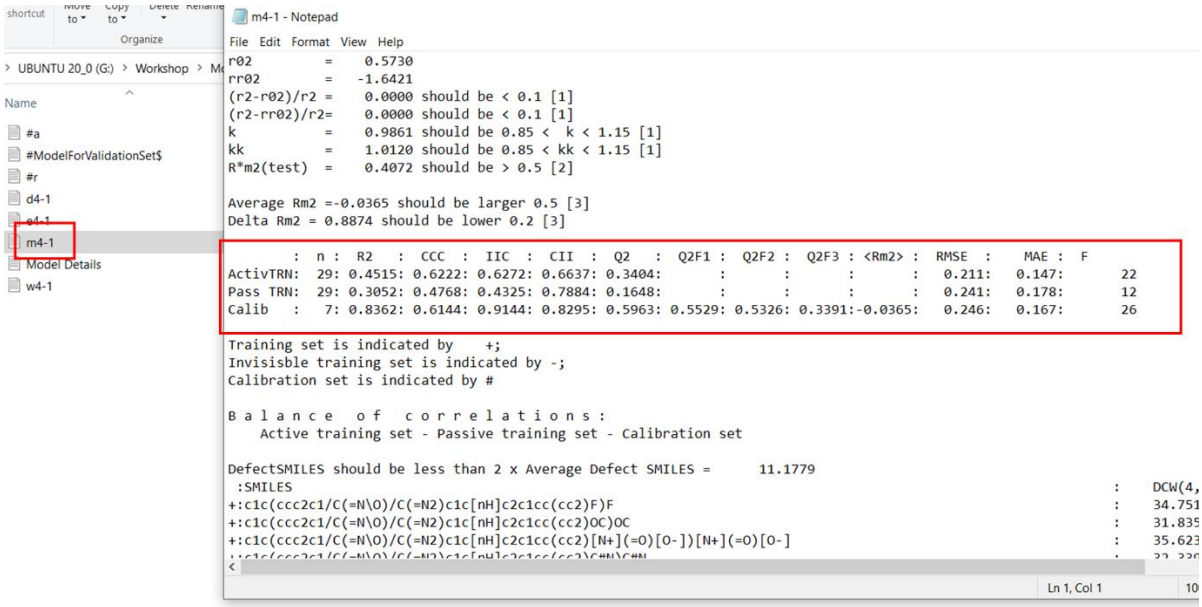


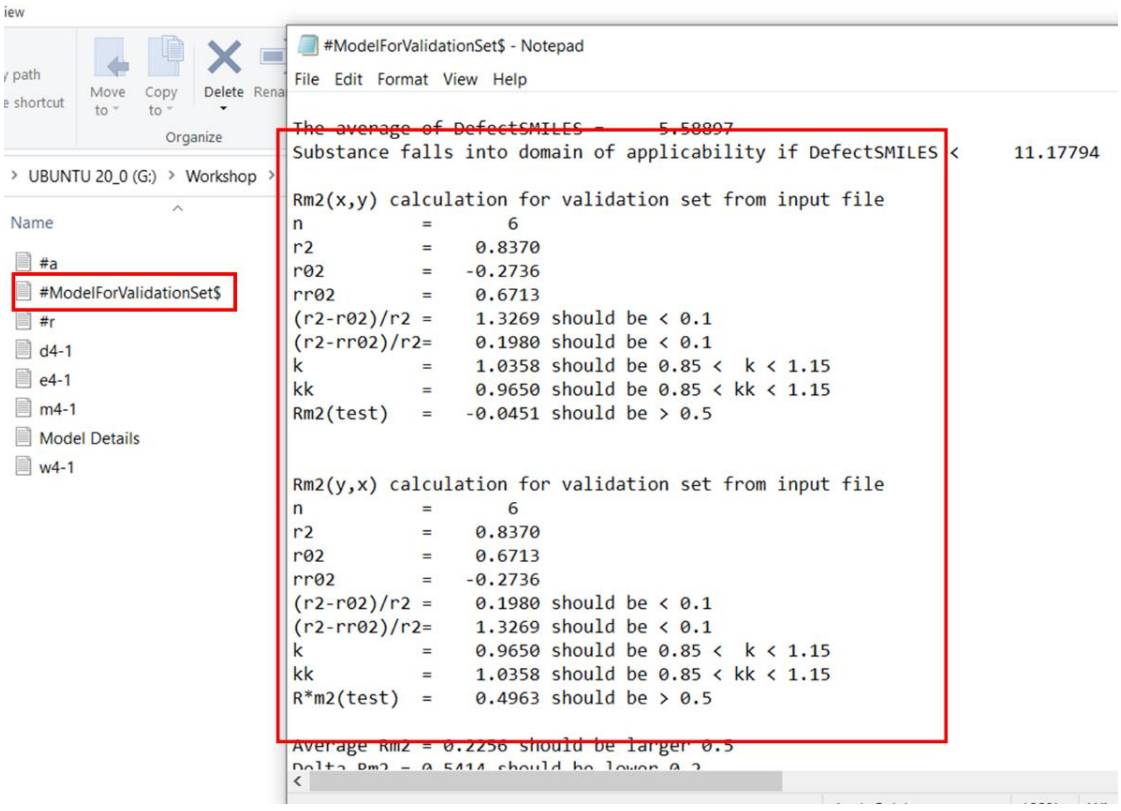
Step 6

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.

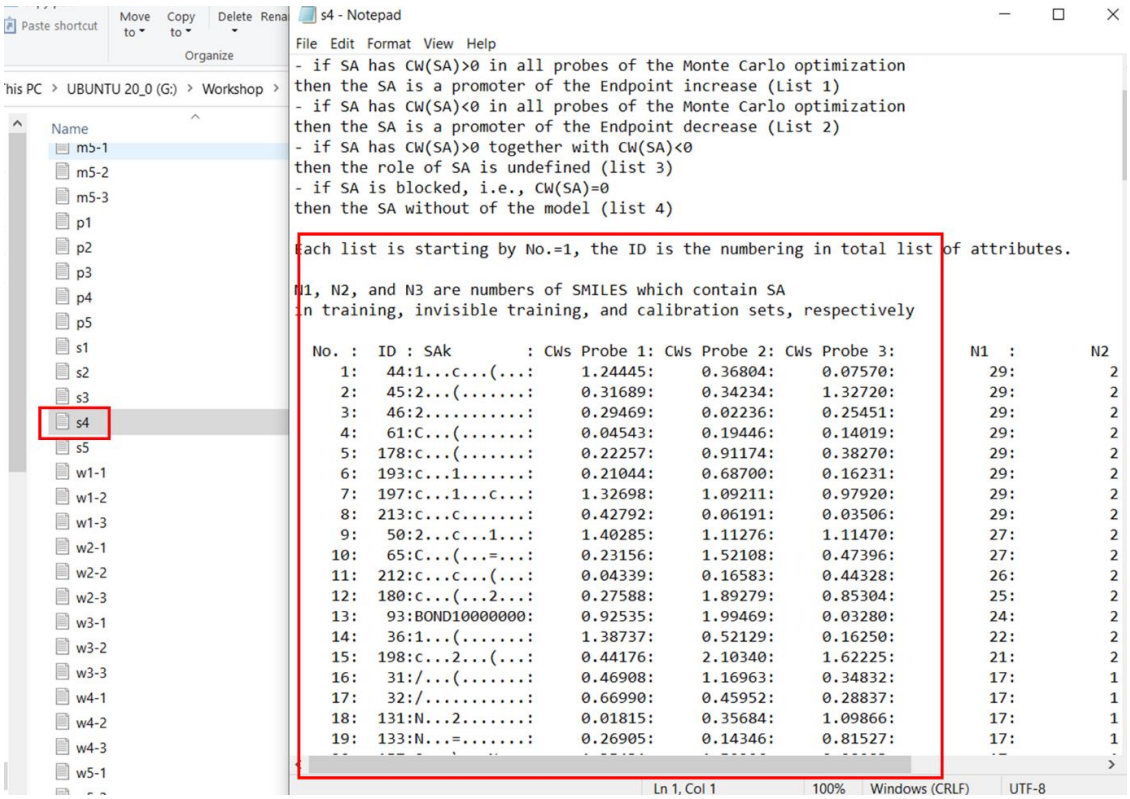
In the Model folder we will have file named with m…. and another #ModelForValidationSet$ which will contain the statistical values for the model built.



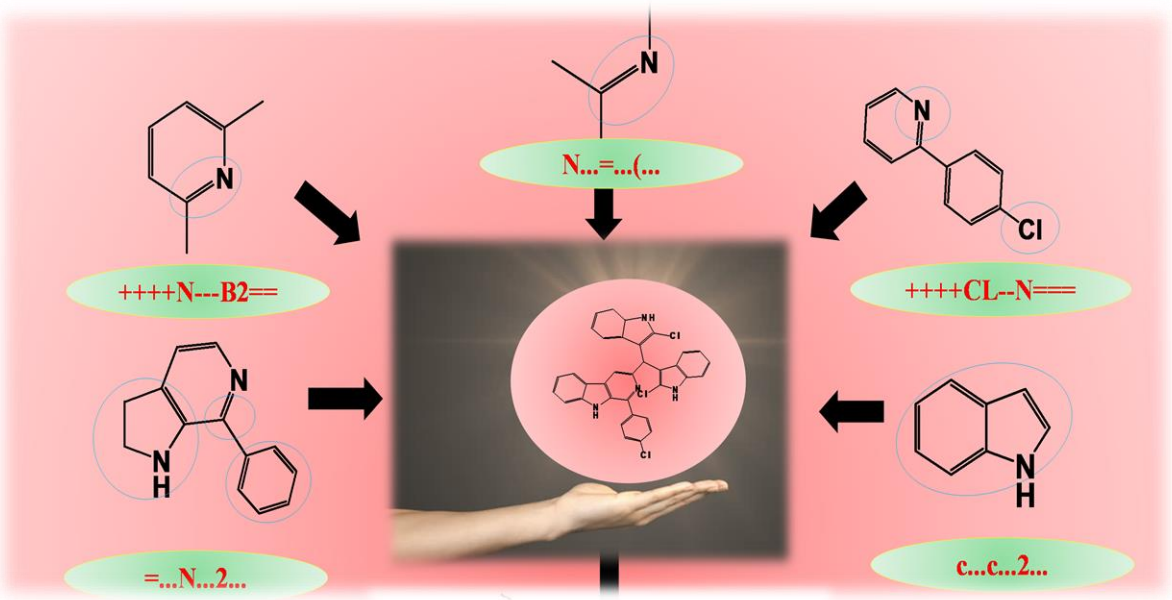


Structural attributes (SA) as promoters and hinderers observed from three CW (Probe)

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.

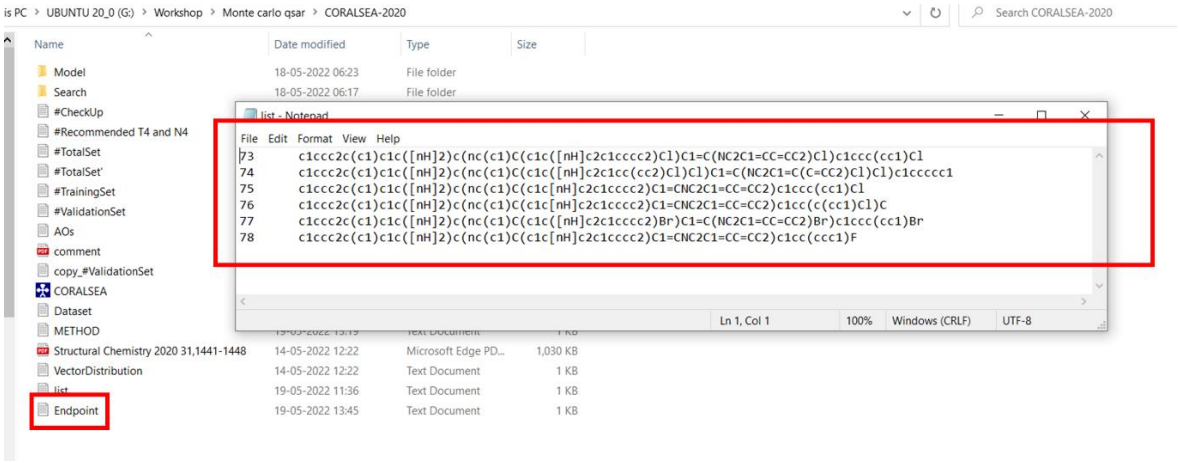


Now based on these attributes obtained and the model build we can either design newer more potent molecules containing most of the important structural fragments or screen a library of compounds that have not yet been validated for the specific target and predict its activity in-silico.



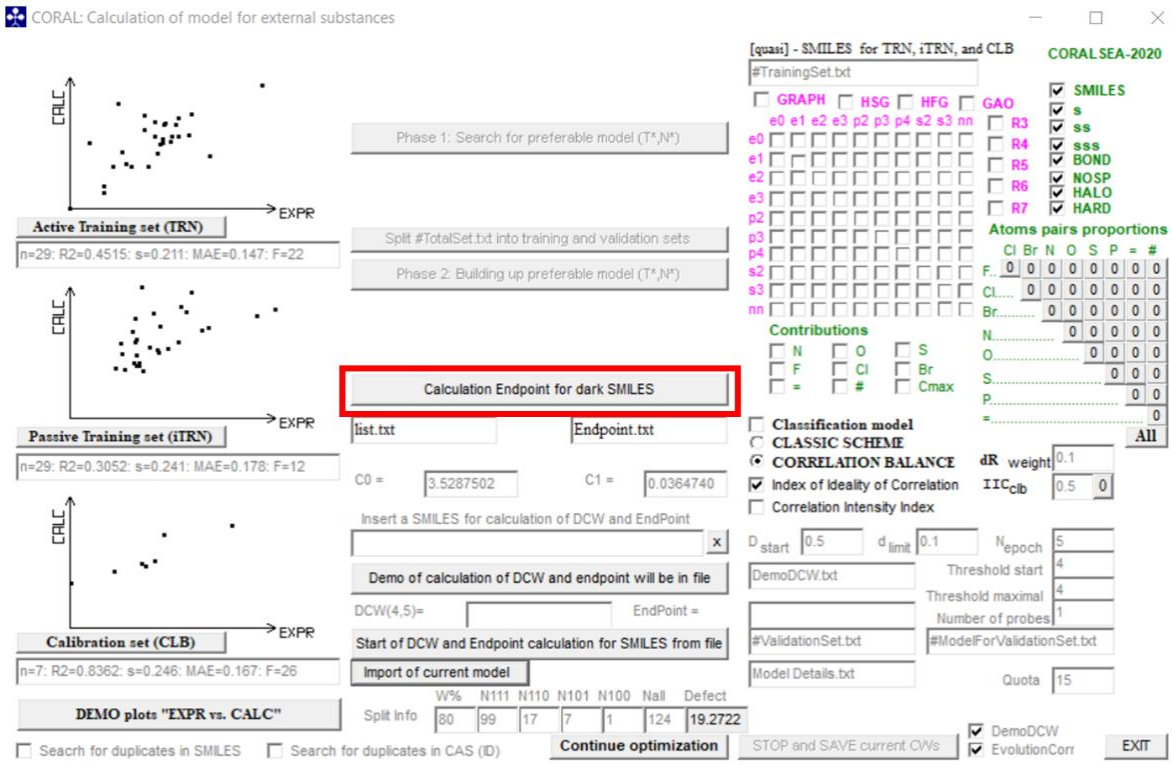
Activity prediction of unknown compounds

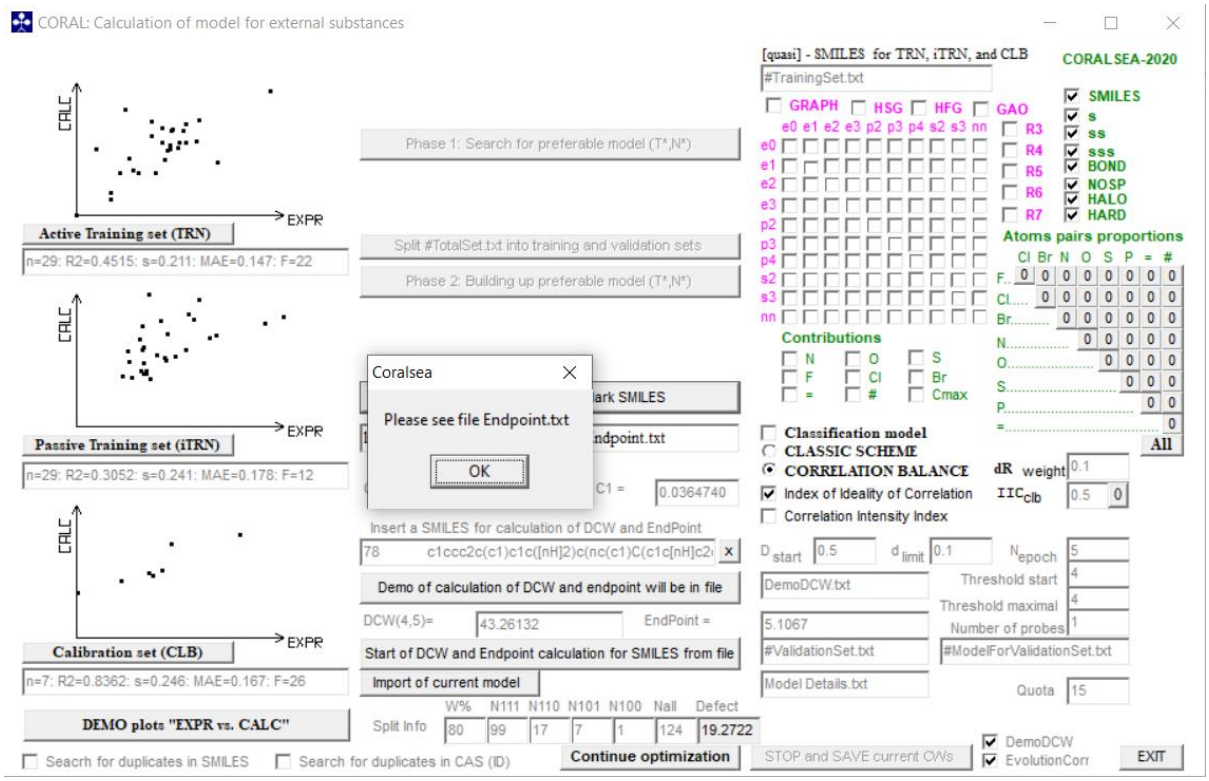
For the input we need to prepare a simple text file containing the serial number and the SMILES format.



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