CONTENTS

ABOUT …

Name: PRECLAV - **PR**operty **E**valuation by **CLA**ss **V**ariables - v. 2102 (2021, February)
called also PreclavPRO

Runs on: WindowsXP, using one processor
running under newer Windows versions was not tested

Programming language: VisualBasic5

Author: Tarko Laszlo Ph.D.
E-mail: tarko_laszlo@yahoo.com
tarkolaszlo756@gmail.com

The PRECLAV package includes:

Mydata.xls	…… starting table of the 'calculated by hand' descriptors or QSPR table
PreclavD.exe	..… program for computing the descriptors
Descript.txt	…... list of the descriptors and their conventional names
PreclavS.exe	..… program for computation of molecular similarities
PreclavQ.exe	..… program for computation of the QSPRs
PreclavL.exe	..… program for computation of the leverages and Applicability Domain
PreclavY.exe	..… program for computation of synergies

1. INTRODUCTION

Let it be a group of K molecules with known chemical structures, and P a certain property of these K molecules. The values of P for these K molecules (the prediction set) are not known.

The PreclavQ module estimates the values of P for the K molecules before their synthesis and/or before measuring the P values by physical/chemical methods. Then, PreclavQ identifies the molecules in the prediction set that are 'recommended/non-recommended for synthesis'.

To estimate the values of P for the K molecules, PreclavQ analyzes other N molecules in the same class (the calibration set) having known values of P. The 'same class' term means, 'the values of P are effect of the same mechanism, at molecular level'. For practical drug design purposes, PRECLAV should be applied only if the analyzed database (calibration set + prediction set) includes a prediction set. For theoretical purposes, PRECLAV can be applied in the absence of a prediction set.

There are many calculable molecular features, called descriptors. For the molecules in the calibration set, the user must introduce the experimental (measured by physical/chemical methods) values of P and the values of the descriptors. For the molecules in the prediction set, the user must introduce only the values of the descriptors. Using these input data the program finds the QSPR (*Quantitative Structure Property Relationship*), i.e. the mathematical formula of the studied property.

According to the QSPR, the value of P is a function of (depends on) *p* variables (predictors). The value of the predictors is effect of the molecular (chemical) structure, size, shape, lipophilicity, flexibility etc. 'The best' description of the property P is not made by only one specific predictor, but by the 'best set' of the predictors, *as a whole*. If the calibration and prediction sets include molecules in the same class, these sets are described, as a rule, by two quite similar QSPRs.

The 'dependent property' P can be 'biochemical activity' (QSPR ≡ QSAR), 'toxicity' (QSPR ≡ QSTR), 'chromatographic retention time' (QSPR ≡ QSRR), 'viscosity' (QSPR ≡ QSVR), 'molecular aromaticity' (QSPR ≡ QSArR), 'biodegradability' (QSPR ≡ QSBR), 'color' (QSPR ≡ QSCR) etc. About 90% of the QSPR studies cited in literature are QSAR studies.

From the point of view of PRECLAV, a high quality database (including a prediction set) presents some characteristics:

- the database includes just one class of molecules
- the calibration set is representative sample within database
- the number of molecules in the calibration and prediction sets is large
- the experimental values of the dependent property are correct and within large range
- there is small number of identical experimental values of the dependent property
- the analyzed molecules include a rigid common molecular fragment

IF
  the quality of the database is low
THEN
  for the molecules in the prediction set
    the **correlation** between the estimated and the (unknown) experimental values is low
    the **difference** between the estimated and the (unknown) experimental values is large
    the **order** of the molecules (ordered by the estimated values) is incorrect
END IF

In practice, the first two (most important) conditions above are not met very correctly, and are very difficult to obtain modifying the initial database. To overcome this difficulty the general work plan should include the following steps.

BLD - building of the molecules, using a molecular mechanics software
OPT - 'geometry optimization', using, mandatory, the quantum mechanics software MOPAC
DRG - get the initial table of the DRAGON descriptors
DMIN - calculating a minimum number of descriptors, using the PreclavD module
DMAX -calculating a maximum number of descriptors, using the PreclavD module and, compulsory,

the DRAGON software

CLS – identification of the new database (new QSPR table) including the 'largest chemical class in the initial database', using the PreclavD module

CLU – identification of the new database (new QSPR table) including the 'largest chemical cluster in the initial database', using the PreclavS module

NBX, NDB and NVB - identification of the new databases (new QSPR tables) obtained by the PreclavS module, by excluding molecules less similar in terms of the chemical structure, shape, size, hydrophilicity and flexibility

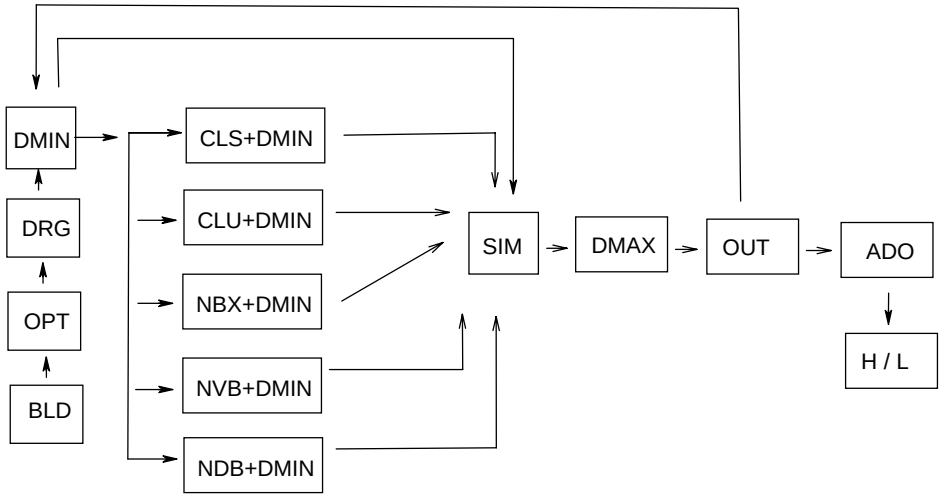
SIM – calculation of the presumed correctness of the final result, using the PreclavS module

OUT - identification and elimination of the outlier molecules from the (new) calibration set, using the PreclavQ module

ADO - identification of the Applicability Domain Outsiders (calculations with calibration set + prediction set)

H / L - labelling of the 'recommended/non-recommended for synthesis' molecules in the (new) prediction set

First step of the QSPR analysis is BLD, last step is H /L. Calculus loops involve repeated use of the PreclaD, PreclavS and PreclavQ modules (not PreclavL and PreclavY).



A simpler procedure is:

BLD → OPT → DRG → $(DMAX \rightarrow OUT)_m$ → $(DMIN \rightarrow NDB / NVB \rightarrow SIM)_n$ → DMAX → ADO → H / L

These steps are subject in the next chapters of the documentation.

Attention!

Using of the DRAGON [18] descriptors is compulsory. These descriptors are calculated based on the output files, in mol2 format, accomplished, for instance, by PCModel software [8]. Consequently, you need to know how to calculate and how to save descriptors, according to the DRAGON documentation. The format of the (only one) output table of the DRAGON descriptors must be readable by the PreclavQ module. Consequently, the 'Skip options for saving' must be 'Information on data' and 'Molecule No.'

2. DATA PREPARATION

2.1. Building of the molecules and Molecular Mechanics Analysis
(BLD step in the list in page 3)

For building the molecules and a rough optimization of the geometry, the molecular mechanics program PCMODEL program [8] is recommended, because this program is very useful to create the input file(s) for MOPAC [7], DRAGON [18] and EpiSuite [26].

The N + K molecules in the database should include a common molecular fragment and different type, number and position of the double/triple bonds and/or chemical groups grafted on the common fragment.

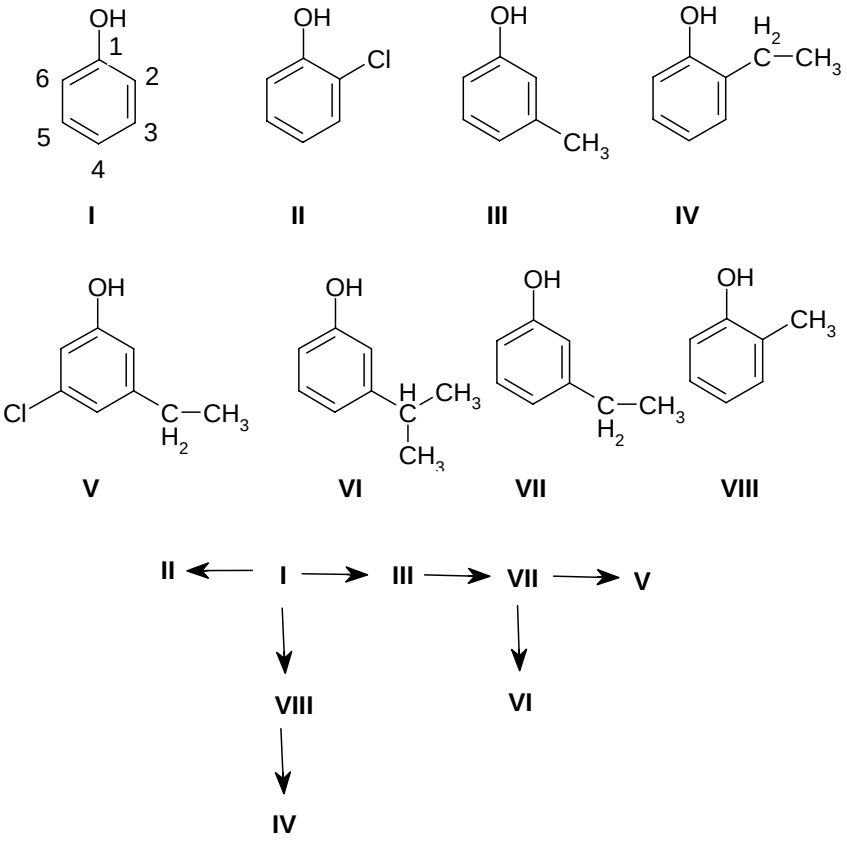
The user selects the bonded heavy (different from hydrogen) atoms having imaginary indices #1, #2 and #3 in the common skeleton. The atom having #1 index should be a carbon atom.

Using a molecular mechanics program the atoms with #1, #2 and #3 indices will be built first.

Then the whole common skeleton will be built.

The first molecule to be built, based the common fragment, should be the molecule with the smallest number of heavy atoms.

Next, each molecule is built starting **from the molecule with the most similar structure** (i.e. with the most similar number of heavy atoms, number and type of chemical groups). **Each time the smallest modification should be done**.



The molecules **I – VI** in the calibration/prediction set will be built in keeping with above sketch, although the molecules **VII** and **VIII** are not included in calibration/prediction set.

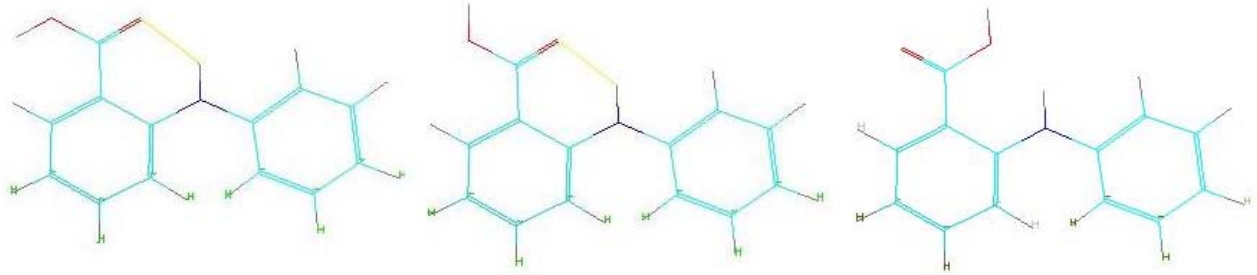
If the analyzed molecules include a common fragment, then disregarding the above instructions may result in serious problems because of erroneous computation of the 3D descriptors (see Section 4.3) and the molecular shape similarity by the 'Superposition' method (see Section 3.2 and Section 5)**.**

PRECLAV is parameterized for almost all elements. The species including lanthanides and radioactive elements cannot be analyzed.

The radicals will not be analysed.

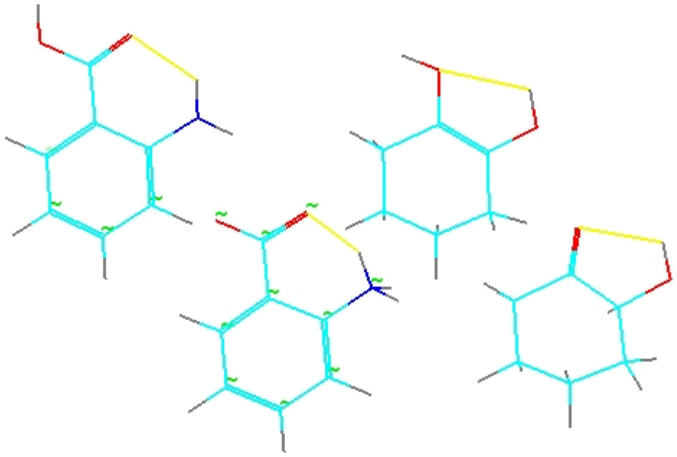
Next, the molecular mechanics program identifies, for each molecule in the calibration/prediction set, the geometry of the conformer with the minimum potential energy (geometry optimisation, i.e. the OPT step in the list in page 3). The value of the descriptors depends on the geometry, especially in flexible molecules. Therefore, **the conformers having 'false minimum energy' have to be carefully avoided**. For instance, you should find the species having the minimum energy in amino-acid/zwitter-ion and keto/enol equilibriums.

You can see below some conformers of N-phenyl anthranilic acid and MMX [22, 23] steric energies (Kcal). First and second conformer includes the intramolecular hydrogen bond N - H … O = C (yellow line).



25.05 **22.24** **22.90**

In addition, you can see below the anthranilic acid (- **1.72** Kcal), the zwitter-ion of anthranilic acid (- **23.15** Kcal), cyclohex-1-ene-1,2-diol (+ **1.22** Kcal) and 2-hydroxy-cyclohexanone (+ **11.15** Kcal).



In order to continue analysis, using the MOPAC quantum mechanics program, will create a '*NAME.mop*' file for each molecule named '*NAME*'.

The following keywords **must be** indicated:

PM6, **BONDS, VECTORS**.

Also the following (same for all molecules) extra-keywords are recommended:

GNORM=0.2, GEO-OK, PULAY

In addition, according to the MOPAC documentation,

- the keyword MMOK **must be** used in the analysis of amides and urethanes (carbamates);
  if the molecule does not include a NHCO group the use of this keyword is useless
- the keyword pKa should be used in the analysis of the molecules which include O-H bond(s);
  if the molecule does not include O-H bond(s) the use of this keyword is **forbidden**
- the keyword CHARGE=*n* **must be** used in the analysis of ions

Appendix #1 of this documentation presents a specimen of '*NAME.mop*' file.

The analysis should be repeated for each molecule. Therefore, N + K files with '*mop*' extension will be obtained. A large value for N and K is recommended, however N + K < 3000. If K = 0 (missing prediction set) then PreclavQ will perform the best multilinear model for P, making prediction only for the molecules in the calibration set.

2.2. Quantum Mechanics Analysis

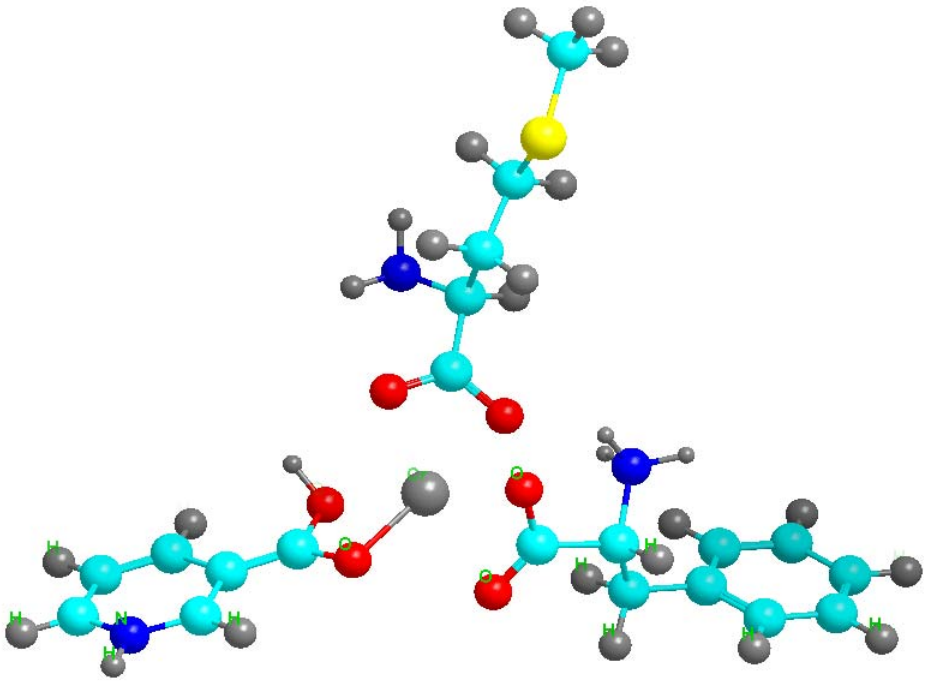
The analysis by the MOPAC program (also within OPT step), using a large number of '*NAME.mop*' files, is continued gradually, as below:

- the results of old MOPAC analysis, i.e. all *'*.mop*', *'*.out*' and '*.*arc*' files in the folder that contains MOPAC program, must be deleted (if they exist)
- all the new '*NAME.mop*' files should be moved into the folder that contains the MOPAC package
- MOPAC program is started

At the end of one molecule analysis, the MOPAC program generates the '*NAME.out*' output file.

The text of this output file **must include** the 'EIGENVECTORS' and 'BOND ORDERS AND VALENCIES' chapters. In addition, the text of the output file should end with the 'MOPAC DONE' words. Moreover, the output MOPAC file must include the energy value of the HOMO molecular orbital (**not SUMO** radical orbital). **Otherwise, the analysis of the molecule should be repeated or the 'incorrect' output file has to be removed out of database. Pay attention to the MOPAC version, to check the format of its output file; the output file has to be readable by PRECLAV, without errors.** For instance, PRECLAV can read the output files of the version 11.366W.

Building and geometry optimization of organometallics is somehow difficult. The charge on the system should reflect the oxidation state of the metal. The (ionic or molecular) structure of the ligand(s) should be accurate.



In the above complex $L_1L_2L_3Cr$ the ligand L_1 is nicotinic acid CAS 59-67-6, the ligand L_2 is phenyl-alanine CAS 63-91-2, and the ligand L_3 is oxi-methionine CAS 4385-91-5. Charge on this system is null. The ligands were built virtually as zwitterions and there are six coordinative bonds Cr – O. During geometry optimisation using the PM6 method, a hydrogen atom of the NH_3^+ group of oxi-methionine is 'moved' to the COO^- group of the nicotinic acid.

If all molecules have been correctly analysed N + K files having the name '*NAME.out*' will be obtained. **Copy all these '*NAME.out*' files into the PreclavD folder.** This MOPAC database may be used for one or more QSPR studies.

2.3. Editing the QSPR table

(included in the DMIN and DMAX steps in the list on page 3)

This table includes the conventional names of the analyzed molecules and the name and the weigthed values of the dependent property.

To edit the initial QSPR table you may use the Microsoft Excel worksheet 'Mydata.xls' in the PreclavD folder and the capabilities of the Microsoft Excel.

- open the 'Mydata.xls' file
- replace the conventional name of cases (mol0001, mol0002, mol0003, …), in first column, by the names (maximum seven letters) of the '*NAME.out*' files

The names of the prediction set molecules must be placed **after** the names of the calibration set molecules.

- delete the unnecessary rows of the table
- replace the 'udz' string in the first row, of the last column, by the conventional name of the Property
- replace the null values of the Property for the calibration set molecules by real numbers

These experimental values of the dependent property should be established with the same precision and by the same method, because the statistics of the N molecules (the calibration set) should be unchanged. If you take these values from literature, a single reference source is recommended.

If the experimental values are expressed by words "YES / NO", "Active / Inactive" etc. or by several stars/crosses, these values **must be** replaced by numbers.

Sometimes, logarithming of the initial / inverse values is required to give them a distribution closer to gaussian distribution and a higher correlation with descriptors.

- replace the null values of the Property for the prediction set molecules by '?' symbol
- delete futile columns in table
- verify if the number of columns is two
- verify if the number of rows is N + K + 1
- save the table as '*NAME.txt*' file, **in 'Text Tab delimited' format**

Usually the *NAME* is $Xn_{cal}n_{pre}$, where X is a capital letter, n_{cal} is the number of calibration set molecules and n_{pre} is the number (even zero) of prediction set molecules, for instance A9348, L3799, F7800 etc.; the total number of NAME's symbols **has to be five**.

- after leaving Microsoft Excel **delete the extension *'.txt'* of the saved file** (i.e. rename '*NAME.txt*' file as '*NAME*' file, **without extension**)

Note

PreclavD and PreclavQ replace **automatically** the values V in your QSPR table by values in the range [1, 10], using the formula $V_{final} = 1 + 9 \bullet (V - minV) / (maxV - minV)$. Therefore, the 'observed values' used in calculations and printed in all output files are not experimental / logarithmed values, but values in this range.

See an example of QSPR table in Appendix #2.

2.4. Editing the 'computed by hand' descriptors table
(included in the DMAX step in the list on page 3)

This table includes the conventional names of the analyzed molecules and the names and the values of some descriptors that are not computed by PreclavD (for instance the molar refractivity of the chemical groups grafted on the common molecular fragment). The user will complete this table row by row and column by column.

Appendix #3 presents the molar refractivity of some substituents.

The name of the saved table of the molar refractivities have to be *NAMEmor.des*. The first column includes the conventional names of the analyzed molecules. The first row must include the conventional names of the refractivities, i.e. mr01, mr02, mr03 …. mr10, mr11 … Other names for table and for refractivities **are forbiden**.

Using of the 'computed by hand' descriptors is not compulsory but it is recommended. In fact, you may verify the statistical features (prediction power, correlations etc.) of any descriptor that is not computed by PreclavD. Moreover, using these descriptors you can identify some **synergistic effects** in the analyzed molecules, see Section 2.5 and 3.4. The program can use simultaneously any number of these descriptors.

To edit the these 'computed by hand' descriptors tables you may use again the 'Mydata.xls' file.

- open the 'Mydata.xls' file
- replace the conventional name of cases (mol0001, mol0002, mol0003, …), in first column, by the names (maximum seven letters) of the '*NAME.out*' files

The order of the names in first column must be the same as in the initial QSPR table

- delete the unnecessary rows in the table
- add any number of columns
- replace the conventional name of the descriptors by your desired conventional names
- replace the zero value of the descriptors by real numbers computed before, if necessary their products
- save the table as '*NAMEchd.des'* file or *NAMEmor.des* (for refractivities), **in 'Text Tab delimited' format**
- **copy** *NAMEmor.des* **table into the PreclavD folder**
- **copy the other descriptor tables into the PreclavQ folder**

2.5. Other descriptor QSPR tables

You can use other program(s) to make, in several minutes, other table(s) of descriptors, for instance DRAGON [18] within mandatory DMAX step in the list in page 3. The values of the computed descriptors must be saved as *NAMExxx.des* files, for instance *NAMEdrg.des* for table of the DRAGON specific descriptors. PreclavQ program may use any number of these tables **if these tables present a proper format**:

- number of rows is maximum 3,000
- number of columns is maximum 5,000
- the first column includes the name of the analyzed molecules

The order of the names in the first column must be the same as in the initial QSPR table or new QSPR table (see below).

- the first row includes the name of descriptors
- the column = 1, row = 1 includes the word 'Cases'
- the character 'tab' separates each field
- the missing values for descriptors are indicated by 0 (null) value

Copy these tables into the PreclavQ folder.

Therefore, the PreclavD folder includes the QSPR table, the *NAMEmor.des* table and N + K '*NAME.out*' files. The PreclavQ folder includes the QSPR table and several descriptor tables.

Attention! During QSPR computation the content of descriptor table(s) in PreclavQ folder will be changed because of elimination of the outlier rows (see Section 5). The content of QSPR table will remain unchanged.

In analysis of mixtures the QSPR table should include the value of the dependent property for all **mixtures**.

The descriptor table includes the conventional name (mix*0001*, mix*0002*, … mix*abcd*) of mixtures (in first column), the conventional name (C*1*, C*2*, …) of mixtures' components (in first row), and the chromatographically percentages of each component in each mixture. **Save this table as NAMEvfr.des** and use this table in QSPR computation by PreclavQ module.

Final QSPR includes, as predictors, some C1, … C*x* name of components. Consequently, the QSPR analysis of mixtures offers information regarding the 'best set of components', having largest influence on the dependent property of the mixtures.

Examples of mixtures' analysis:

Mixture	*Dependent property*
natural or synthetic bioactive mixture	biochemical activity
cosmetic, wine, juice	certain measurable organoleptic property
coal, mineral oil	certain physical property
reaction mixture	certain property of the final product(s)

2.6. Synergy tables

To identify the synergy of the components in a certain mixture, the columns of the table should include **the same chromatographic percentages**, see Section 2.5, but the last column of the synergy table **must include the name and the values of the bio-activity** of all mixtures.

To identify the intramolecular synergy **of the substituted positions** *i* and *j* on the common molecular fragment **(regardless of substituent)** the columns in the table should include the molar refractivity MR_i and MR_j of all substituents, see Appendix #3, and the last column must include the bioactivity values of all molecules.

To identify the intramolecular synergy **of** the grafted **PRECLAV molecular fragments (regardless of the position)**, see Section 4.1, the columns in table should include the percentages of all fragments (**included in NAMEvfr.des file**). The last column must include the values of the bioactivity of all molecules, taken from QSPR table.

These tables include only the calibration set. Hence, **in the synergy tables the rows of the prediction set molecules and the value '?' for bioactivity are banned.**

Save the synergy table under name NAME.syn.

Then, you can use the PreclavY module, see Section 3.4.

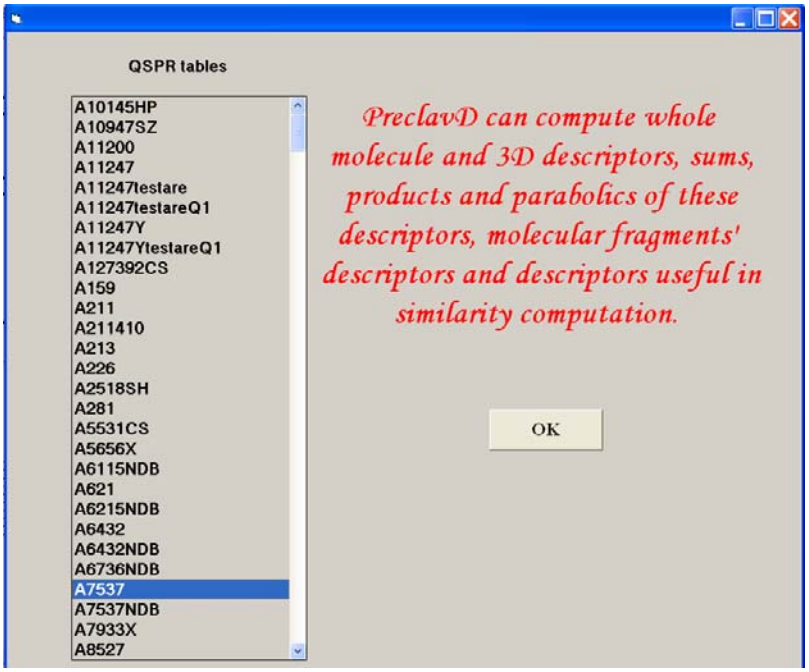
3. COMMANDS

3.1. The computation of the PRECLAV descriptors
(within DMIN and DMAX steps of the scheme on page 3)

Before turning on the PreclavD module:

- the presence of desired QSPR table within PreclavD folder will be verified
- the presence of all necessary *NAME.out*' files within PreclavD folder will be verified
- if you want to use the molar refractivities the presence of the *NAMEmor.des* file within PreclavD folder will be verified

The first image includes the QSPR table(s) that can be used by the program. User must select the needed table by clicking on it. After clicking, the image is changed as below. The '**OK**' button will take you to the next image.



3.1.1. The options in computation of the descriptors

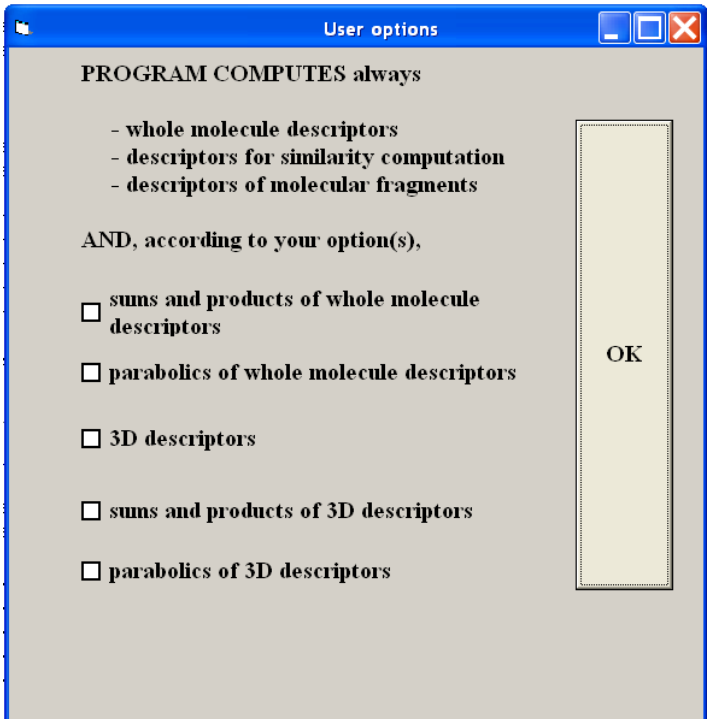
The program will compute the value of '*whole molecule*' descriptors and the molecular fragments' descriptors. Within 'descript.txt' list the descriptors until '*kic*' are considered '*whole molecule*' descriptors; the descriptor '*phi'* and the next descriptors are considered weighted descriptors. In addition, the program can compute the sums and products (**mandatory within DMAX step**) of '*whole molecule*' and 3D descriptors.

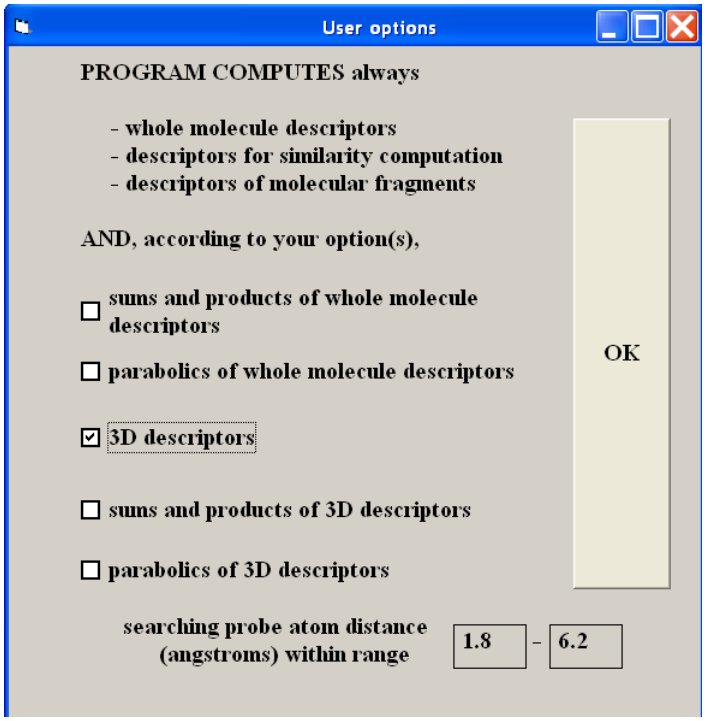
If the user option is '3D descriptors' she/he will see other version of this image, with two additional windows, i.e. the proposed value for *probe atom* distance in 3D descriptors computation (see Section 4.2)

The default minimum value is 1.8 Å and the default maximum value is 6.2 Å. PreclavD searches the optimum value within [min. value, max. value] range. The user can modify these default values by clicking within the corresponding window. If your decision is min. value = max. value the program skips this stage of the computation.

If the analysed molecules have a common skeleton the options '3D descriptors' and 'sum and products of 3D descriptors' **are mandatory in the DMAX step in the list on page 3**. On the contrary, **within DMIN step PreclavD should be used without any sums, products and 3D options, to obtain the minimum number of descriptors,** in a shorter calculation time**.**

The '**OK**' button take you to the next image.





3.1.2. Analysis of the molecules

There are no command buttons on this image.

The user will see the successive analyses of the molecules, i.e. the computation of the descriptors.

The computation time for sums and products is large because of the huge number of these descriptors.

The program refuses to analyse too small/large molecules.

After computation, you see a new image (name of saved files).

On this new image:

'**Print**' - allows the printing of the content of *rezultat.rez* file.

'**Quit**' - allows the return to the operating system.

The '*NAMEType1/2Classes*' file (table) presents the identified Type 1 and Type 2 chemical classes in the initial database/QSPR table. You can understand the value of the Class according to Appendix #4, Table 1 and Table 2. The 'largest chemical class' is considered the class having greatest value of the $N \bullet (1+K)^{0.5}$ product. This class is automatically saved as a new QSPR table, in CLS step in the list in page 3.

Combining classes to obtain a new database (QSPR table) is the responsibility of the user. For instance, one can aggregate the classes of the molecules with/without OH/NH chemical bonds. Other posibility: grouping all Type 1 classes which include molecules in the calibration set *and* molecules in the prediction set.

The abridged values (four decimal points)

of the fragments numbers are saved as *NAMEnfr.des* file.
of the fragments percentages are saved as *NAMEvfr.des* file.
of all sums of fragments numbers pairs are saved as *NAMEsn1,2,3....des* files.
of all sums of fragments percentages pairs are saved as *NAMEsp1,2,3....des* files.
of all products of fragments numbers pairs are saved as *NAMEpn1,2,3....des* files.
of all products of fragments percentages pairs are saved as *NAMEpp1,2,3....des* files.
of the QSPR of fragments (*frg* descriptor) are saved as *NAMEfrg.des* file.
of the QSPR of several WM descriptors (*qsc* descriptor) are saved as *NAMEqsc.des* file.
of the qsc descriptors are saved as *NAMEqsn.des* file.
of all sums of the qsc descriptors pairs are saved as *NAMEsq1,2,3....des* files.
of all products of the qsc descriptors pairs are saved as *NAMEpq1,2,3....des* files.
of the whole molecule descriptors are saved as *NAMEwho.des* file.
of all sums of the whole molecule pairs are saved as *NAMEsw1,2,3....des* files.
of all products of the whole molecule pairs are saved as *NAMEpw1,2,3....des* files.
of refractivities, all sum and all products of the refractivities are saved under the same name *NAMEmor.des* file.
of the parabolic whole molecule descriptors are saved as *NAMEpaw.des* file.
of the any type of distances are saved as *NAMEdon.des* file.
of the 3D descriptors are saved as *NAMEgrd.des* file.
of the sums of the 3D descriptors are saved as *NAMEsg1,2,3....des* files.
of the 3D descriptors products are saved as *NAMEpg1,2,3....des* files.
of the parabolic 3D descriptors are saved as *NAMEpag.des* files.

The value V_{par} of the parabolic functions is

$$V_{par} = a + b \cdot V_{desc} + c \cdot V_{desc} \cdot V_{desc}$$

where V_{desc} is the value of the descriptor and *a, b, c* are the weighting factors. The value of these weighting factors is not printed.

The null values of the descriptors are automatically replaced by the value 0.0001.

The above descriptors' tables include only the *significant descriptors*, i.e. the descriptors having quite high correlation r^2 with the dependent property P, i.e. $r^2 > \ln(N)/N$. Frequently, the total number of the *significant descriptors* is more than 10,000.

Copy these descriptors' tables into PreclavQ folder.

Also, PreclavD makes the *'NameSIM.des'* and *'NameSHP.des'* files, used in the mandatory SIM similarity computation step, according to the scheme on page 3. Copy these descriptor tables in the PreclavS folder.

As a rule, the *'NameSIM.des'* and *'NameSHP.des'* files must be obtained, by PreclavD, using all six QSPR tables, i.e.:

- the initial QSPR table
- the new QSPR table including the largest chemical class, obtained by PreclavD

- the new QSPR table including the largest chemical cluster, obtained by PreclavS, see Section 3.2
- the NBX, NVB and NDB new QSPR tables obtained by also PreclavS

For the similarity computation you must read the Section 3.2 and use the PreclavS module.

3.2. The similarity computation

(in analysis of initial database, CLU, NBX, NVB and NDB steps in the list on page 3)

As a rule, 'similar structures have similar properties'. If the 'properties' are 'biochemical activities' this statement is named 'the QSAR axiom'. Frequently, this 'axiom' is challenged because some similar structures present non-similar activities and some non-similar structures present similar (close) activities.

Chemical structure	**Bioactivity**	**Comments**
similar	similar	usual
similar	different	enantiomers?
		different pharmacophores?
		different shape/size/lipophilicity?
different	similar	intramolecular synergistic effects?
		similar shape/size/lipophilicity?
different	different	usual

A different problem is the similarity of the calibration and prediction sets (both as a whole) [32]. In fact, a high similarity of the calibration set and prediction set does not warrant fulfilment of the 'QSAR axiom' for each pair of molecules. Nevertheless, the PreclavS module offers some information regarding the similarity of the analyzed molecules regarding the molecular (chemical) structure, shape, size, hydrophilicity and flexibility:

- identifies the molecules with lowest similarity
- identifies the duplicate molecules (using of duplicates is prohibited)
- computes the 'observance of the QSPR axiom'
- identifies the 'clusters' (within clusters the similarity is high, for all molecules pairs)
- identifies the 'outliers' ('atypical') molecules
- computes the diversity of the molecules in the calibration/prediction set and database
- computes the similarity of the calibration and prediction sets
- if K > 0 (the prediction set is present) the program makes automatically (within CLU, NBX, NVB
- and NDB steps in the list on page 3) the new databases (QSPR tables) '*NameCLU*', '*NameNBX*', '*NameNVB*' and '*NameNDB*'
- calculates the presumed correctness of the final result of the next QSPR computations

Before turning on the PreclavS module:

- the presence of the necessary '*NameSIM.des*' and '*NameSHP.des*' files in the PreclavS folder will be verified
- verify if the 'Name' in '*NameSIM.des*' and '*NameSHP.des*' **includes five symbols** and correct it if necessary

The first image does not include the button '**OK**'.

First, the user should select the method for shape similarity computation. Then, the user should select the descriptors' '*NameSIM.des*' table by clicking on it. After clicking and coming out of the button '**OK**', the user can modify the minimum values of the similarity in clusters.

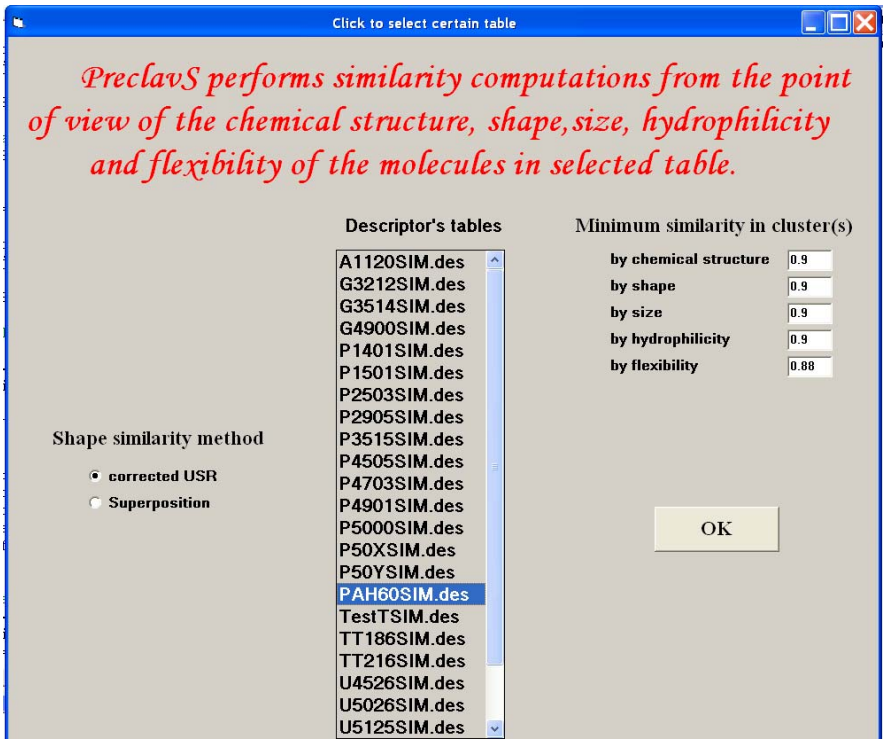
The modification of the default values is not recommended. However, if the analyzed molecules have no common skeleton and the shape similarity method is 'Superposition', the minimum value of similarity in 'shape clusters' can be much smaller, for instance 0.75.

The button '**OK**' allows the transit to the next image.

All similarity computations are made without intervention of the user.

The detailed results of the similarity computations are saved as NameSIM.res file.

The button '**Quit**' - allows the return to the operating system.



The *NameCLU* table includes the largest 'chemical cluster'.

In *NameNBX* table, the molecules in the calibration and prediction sets include the same molecular fragments.

In *NameNDB* table the calibration set includes the validation set. At the end of the output file of PreclavS you can view the number of molecules in the calibration, validation and prediction sets. The molecules in the calibration set of NDB table, which are not included in the validation set, are in the first rows in table.

The *NameNVB* table includes only the molecules in the validation set and the molecules in the prediction set.

The PreclavD and PreclavS modules must be used within loops including DMIN, NBX, NVB, and NDB. To avoid any confusion, the new saved database (QSPR table) should be renamed each time. **Applying the scheme on page 3 you should use, in the next QSPR computation, the database with highest correctness COR and N > 35.**

3.3. QSPR computation
(in the OUT step in the list on page 3)

As shown in Section 3.2, in the QSPR calculations one uses the database (QSPR table) with highest correctness COR and N > 35.

To compute some QSPRs you must use the PreclavQ.exe program.

Before turning on the PreclavQ:

- the presence of the desired QSPR table(s) in the PreclavQ folder will be verified
- the presence of some '*NAMExxx.des*' files in the PreclavQ folder will be verified

The step OUT is preceded by the step DMAX. **In the DMAX step PreclavD must be used with sums, products and 3D options** (without parabolic functions)**, to obtain the maximum number of descriptors.**

Moreover, in DMAX step the original DRAGON descriptor table must be modified according to the QSPR table used. Some rows in the DRAGON descriptor table have to be removed, other rows have to be reordered. The order of the rows must be the same as in the QSPR table. Unfortunately, the original DRAGON table needs to be changed manually. In the same situation is the table of the molar refractivities *NAMEmor.des*.

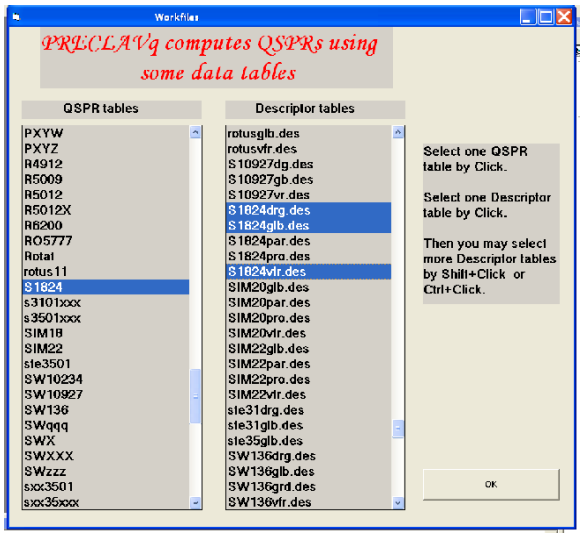
Therefore, in the OUT step, PreclavQ uses many PreclavD descriptor tables and one DRAGON descriptor table. Other descriptor tables can be used, see the 2.4 Section.

The OUT step is automatically followed by ADO and H / L steps. However, the results of the ADO and H / L steps are useful **only if there are no outliers** in the calibration set, see the Section 3.3.3. If there are outliers, the rows of the outliers must be manually removed (**including or not the last eliminated outlier,** see Section 5) from the QSPR table used.

This modified table becomes 'the initial QSPR table' and **the next step must be again DMIN,** according to the scheme on page 3.

3.3.1. List of QSPR and descriptors' table(s)

The QSPR and descriptor tables that can be used by the program are presented.



The needed file(s) will be selected by clicking on it, Shift+Click or Ctrl+Click commands.

Using different sets of descriptors the program computes different QSPRs and achieves different estimations for the prediction set molecules.

Using DRAGON descriptors **together** with the PRECLAV descriptors **is mandatory.**

The selection of the '*NAMEvfr.des*' file **is compulsory**.

If the common skeleton of analyzed molecules includes an aromatic system using the molar refractivity descriptors **is recommended.**

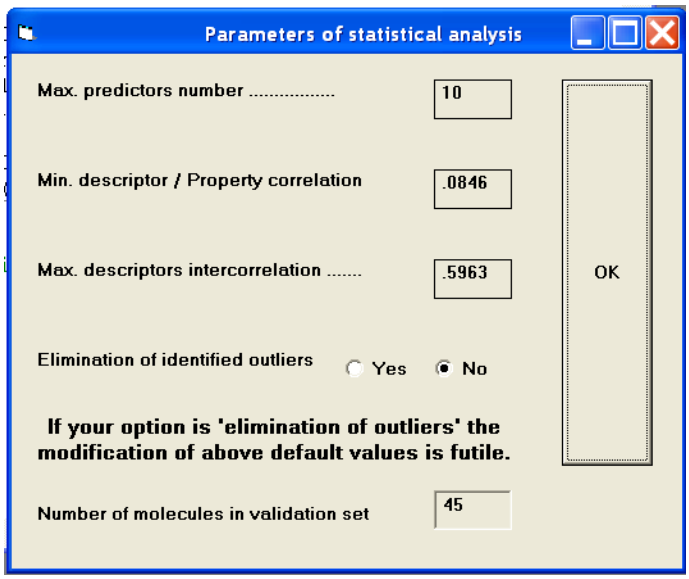
Using the parabolic functions, the '*NAMEfrg.des*' file and/or the '*NAMEqsc.des*' file is not recommended because its can produce the 'overfitting' phenomenon.

'OK' button allows the transit to the next image.

3.3.2. The parameters in the statistical analysis

The image presents the values of statistical analysis parameters.

The user can modify these default values by clicking in the corresponding window. However, the modification of the default values **is not recommended**.



- *the maximum number of variables (descriptors, predictors) in QSPR*

The default value is 20 and the accepted value is in the range [2, 19].

- *the minimum value* $minr^2$ *of square linear correlation variables / Property*

The default value is $minr^2 = 1/N \bullet Ln\ (N)$ and the accepted value is in the [0.01, 0.25] range.

- *co-linearity is the maximum value of square linear intercorrelation* r^2 *of descriptors in sets of 3 to 20 variables*

The default value is $4 \cdot N^{-0.5}$ and the accepted value is in the range [0.1, 0.64]. On the other hand, the co-linearity of descriptors in sets of two variables (orthogonality) is automatically $N^{-0.5}$ and the user cannot modify this value.

As a rule, the option regarding the elimination of outliers should be 'Yes'. However, **sometimes the elimination of the outliers reduces the predictive quality of the final equations for the molecules in the prediction set** [39]**.**

The user must indicate the number of molecules in the validation set, according to the result of the similarity calculations, see the end of the output file of PreclavS.

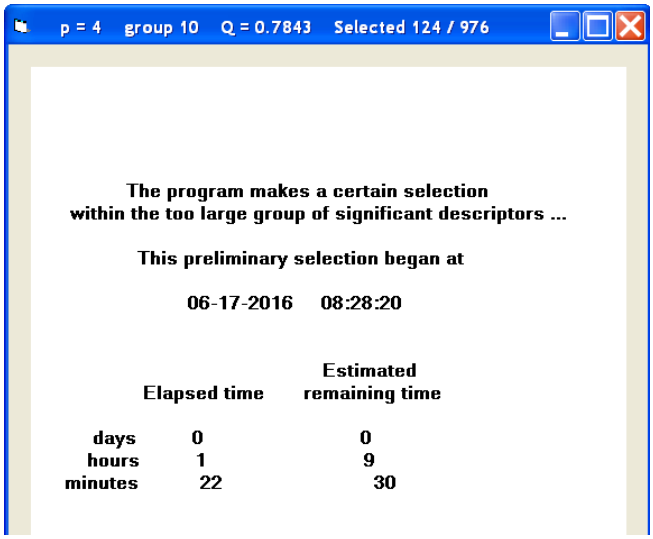
The **'OK'** button take you to the next image.

3.3.3. QSPR computation

If the number of descriptors in tables is high (> 1000) the PreclavQ procedure makes a preliminary selection of descriptors applying a certain unusual method [34]. The algorithm of the selection is very powerful. The 'significant' descriptors are selected even if they represent only 0.1% in the group of the calculated descriptors. Usually, the 'significant' descriptors represent 1-5 % in the group of the calculated descriptors.

The computation time in the preliminary selection can be 2-3 days, see below image, and depends on the number of descriptors and the number of molecules in the calibration set. **If the preliminary selection is interrupted quickly, you must read the contents of the 'rezultat.rez' file**

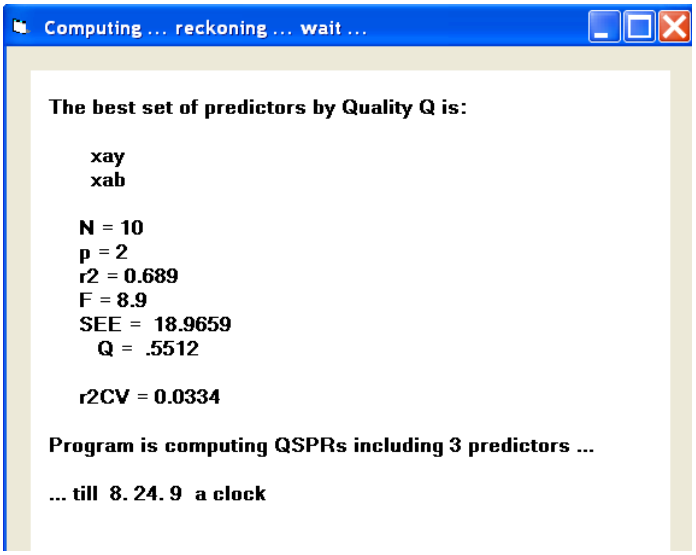
and use, when you restart the program, a smaller number of descriptor tables. In this case, you should make two selections, using half number of tables and NAMEvfr.des file each time.



The selected descriptors are saved as '*NAMEsel.des'* file. Next, in the QSPR computation phase, the program uses automatically only this file and the '*NAMEvfr.des'* file.

After the end of the preliminary selection of the descriptors the user will see:

- splitting of large descriptor tables
- other selection of the significant variables (descriptors)
- the centralization of the natural values of all descriptors
- the computations of the descriptor intercorrelations
- the build-up and selection of the descriptor sets of QSPR equations with 2, …, *p* descriptors
- the value of quality r^2, F, SEE, Q, r^2_{CV}, computed for various QSPRs
- the identification and elimination of the outlier molecules
- the estimation of the P values for the molecules with known P values (calibration set) and unknown P values (prediction set)



If the number of the analyzed molecules, the number of 'significant' descriptors and the number of outliers is large, the computation time can be several days.

The final prediction is saved as *rezultat.rez* file.

The saved file *Name*-EST includes the estimated values by the best QSPRs, (see Section 5). The ADO (Applicability Domain Outsiders) molecules in the prediction set are labelled.

The value of predictors is saved as '*NAMEprd.dat*' file(s).

3.3.4. View the results

The content of the *rezultat.rez* file is presented.

The scrolling button allows the scrolling of the text.

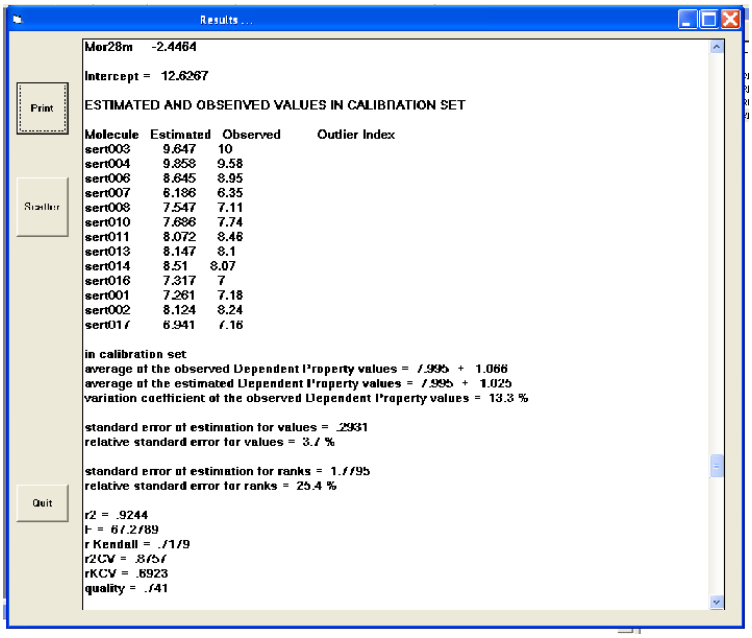
'**Print**' - allows the printing of the content of *rezultat.rez* file

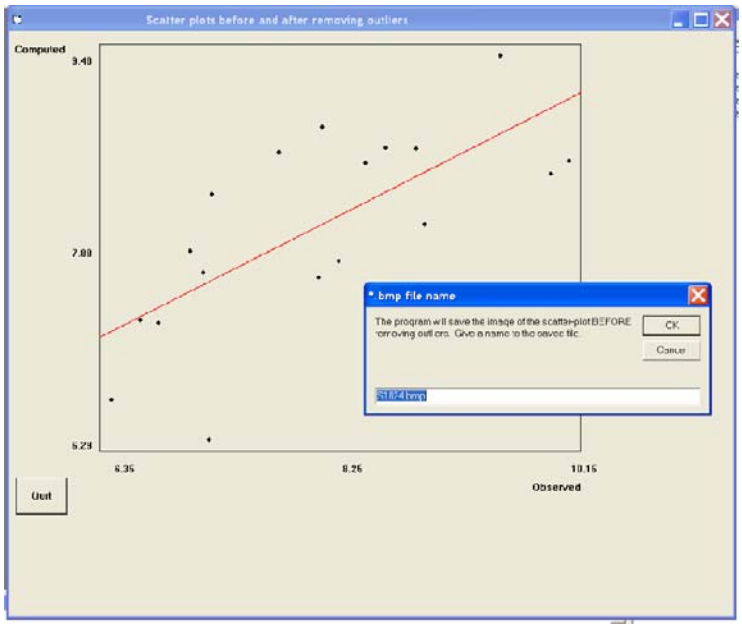
'**Scatter**' – the user may view the scatter plot of the computed/experimental values before and after the elimination of outliers. These two scatter plot images are saved under the desired name and can be used in the published papers.

'**Quit**' - allows the return to the operating system after changing of the *rezultat.rez* file name.

The ADO molecules in the prediction set are not labelled by 'high/low' word because these molecules have been removed from the list. Different QSPRs can identify different ADO molecules.

Use of the modules PreclavD, PrecalvS or PreclavQ to solve simultaneously two different problems **is not recommended** because of large combined computation time.





3.4. Synergy computations

The term 'Synergy' has been used in biochemistry, pharmacology and others fields, mathematics included. This term is derived from the Greek *syn-ergos,* meaning 'working together'. Generally speaking, 'synergy' means 'the creation of a whole that is different than sum of its parts'. If the effect of the presence of the molecular features F_i *and* F_j *is different* (higher or lower) than $E_i + E_j$ sum of their effects, there is a *synergistic* (positive or negative) effect of F_i and F_j. The negative synergistic effect is sometimes named *antagonistic* effect.

Before starting the PreclavY module one must verify the presence of necessary '*Name.syn*' file(s) within PreclavY folder, see Section 2.6.

The first image does not include the button 'OK'.

The user must select the needed table by clicking on it. After clicking, the image includes the button 'OK'.

After clicking of **'OK'** button the synergy of all pairs of descriptors is computed instantly.

The results are included in the output file NameSYN.res.

3.5. Applicability Domain computations

Before starting of PreclavL module one must verify the presence of necessary '*NAMEprd.dat*' file in the PreclavL folder.

The program is very intuitive. You should indicate the number of molecules in the calibration and prediction sets and the number of predictors related to '*NAMEprd.dat*' file.

The leverages are calculated using the classic algorithm, see Section 5.

The Applicability Domain Outsiders (ADOs) are identified using a classic procedure. However, the 'ADO by differences *and* leverages' are identified using a specific method, see Section 5.

The results of AD and leverage calculation are saved in as 'NAME.lev' file.

4. PRECLAV DESCRIPTORS

The program computes, for all molecules, the values of:

- more than 200 WM 'whole molecule' descriptors and their sums, products and parabolic functions
- almost 300 weighted variables
- other 625 3D descriptors and their sums, products and parabolics
- the 'mass percentage' of the identified molecular fragments and their sums and products
- the descriptor 'QSPR of the molecular fragments'
- the QSPR of several WM descriptors

Some descriptors' values are computed by MOPAC and read by PRECLAV from the files '*NAME.out*'. Other descriptors' values are computed by PRECLAV himself, based on the content of '*NAME.out*' files.

The reading / computation procedure is presented below.

4.1. 'Whole molecule' and 'fragmentation' descriptors

Number of atoms
is the total number of atoms of different types.

Molecular mass
is computed as a sum of atomic masses (the usual values - two decimal places - are used).

Number of C, H, N, O, halogens, S, P, As, B and Si atoms
from the MOPAC file (reading the Chapter 'NET ATOMIC CHARGES').

van der Waals radius of atoms
the tabulated values from [1].

Number of chemical bonds (by type)
is the total number of bonds of a certain type (Chapter 'BOND ORDERS AND VALENCIES').

If the bond order is 0.24 to 0.70 then the chemical bond is considered 'coordinative'.
If the bond order is 0.70 to 0.93 then the chemical bond is considered 'single (weak)'.
If the bond order is 0.93 to 1.009 then the chemical bond is considered 'single'.
If the bond order is 1.009 to k then the chemical bond is considered 'single conjugated'.
If the bond order is k to $k + 0.8$ then the chemical bond is considered 'aromatic'.
If the bond order is $k + 0.8$ to 2.5 then the chemical bond is considered 'double'.
If the bond order is greater than 2.5 the chemical bond is considered 'triple'.

The value of k limit depends on the semi-empirical MOPAC method used i.e. AM1 (k = 1.014), PM3 (k = 1.024) or PM6 (k = 1.051). 'Unsaturation index' is (lga+lgd+lgt)/nlh ratio, see descript.txt file.

Number of the chemical bonds
is the total number of chemical bonds (bond order > 0.24) i.e. coordinative, single, single conjugated, aromatic, double and triple bonds.

Cyclomatic number

$$C = L - N + 1$$

where
L is the number of chemical (not 'coordinative') bonds
N is the number of atoms

Average and maximum value of bond order for various chemical bonds
read from Chapter 'BOND ORDERS AND VALENCIES'.

Aromaticity of aromatic chemical bonds
is computed using the formula based on *B* bond order computed value of analyzed bond, according to the Topological Path Aromaticity (TPA) model [12, 13].

$$A = 1000 - 6250 \cdot (B_{lim} - B)^2$$

where B_{lim} is the bond order value of C-C bonds in benzene. The B_{lim} value depends on the semi-empirical MOPAC computation method used ($B_{AM1} = 1.41681$; $B_{PM3} = 1.42411$; $B_{PM6} = 1.44059$)
For aromatic bonds $A > 0$.

Miscellaneous distances (Å) of O, N, C, X, P atoms and OH, NH, SH, PH groups
are computed by usual geometrical formulae (the Cartesian co-ordinates of atomic nuclei from 'CARTESIAN CO-ORDINATES'); the atoms are identified by mass

Van der Waals molecular volume and area of molecular surface
are computed using a Monte-Carlo type procedure [2], Cartesian co-ordinates and atomic Van der Waals radii [1]. Also, the program reads the value of COSMO area and COSMO volume. In addition, the program reads, from 'MOLECULAR DIMENSIONS' Chapter, the molecular dimensions (considering hydrogen atoms but not the Van der Waals diameters) and computes average and coefficient of variance of the molecular dimensions.

General shape index
The general shape (1D, 2D or 3D) of the circumscribed ovoid of the analyzed molecule is the value of function *gs*i, within [1, 3] range [31].

$$gsi = (1 \bullet SIM1 + 2 \bullet SIM2 + 3 \bullet SIM3) / (SIM1 + SIM2 + SIM3)$$

$$SIM1 = [1-(1/3 \bullet \Sigma R_{i1}^2)^{0.5}]^3 \; ; \; SIM2 = [1-(1/3 \bullet \Sigma R_{i2}^2)^{0.5}]^3 \; ; \; SIM3 = [1-(1/3 \bullet \Sigma R_{i3}^2)^{0.5}]^3$$

R_{11}=smd/fmd R_{21}=tmd/fmd R_{31}=tmd/smd-1
R_{12}=smd/fmd-1 R_{22}=tmd/fmd R_{32}=tmd/smd
R_{13}=smd/fmd-1 R_{23}=tmd/fmd-1 R_{33}=tmd/smd-1

The descriptors *tmd, smd* and *fmd, tmd* $\leq$ *smd* $\leq$ *fmd,* are the molecular dimensions (see the 'MOLECULAR DIMENSIONS' Chapter and the descript.txt file). The similarities SIM1, SIM2 and SIM3 are the similarities with ideal 1D, 2D and 3D molecules. If *gsi* < 1.2 the shape of the circumscribed ovoid is very elongated (dicyane, triacetylene etc.). If 1.7 $\leq$ *gsi* $\leq$ 2.2 the shape is somehow planar (benzene, 1,3,5-trinitro-benzene, pyrene, 1,2-dichloro-etene etc.). If *gsi* $\geq$ 2.7 the shape of the circumscribed ovoid is almost spherical (CX_4, cubane, adamantane, fullerenes etc.). For biphenyl *gsi* ~ 1.78, for 2,2'-diiodo-diphenyl *gsi* ~ 2.27.

Volumes of the parallelepiped, ellipsoid and sphere, circumscribed to the molecule
are computed using the usual geometrical formula (the Cartesian co-ordinates of atomic nuclei from 'CARTESIAN CO-ORDINATES') and the atomic Van der Waals radii.

Maximum, average and RMS of distances to the geometric centre (for all atoms)
and
RMS of distances to the geometric centre (for hydrogen and halogen atoms only)
are computed using an usual geometrical formula (the Cartesian co-ordinates of atomic nuclei from 'CARTESIAN CO-ORDINATES').

Area of the circumscribed ellipsoid
is computed using co-ordinates, atomic radii and a numerical procedure [2].

The inertia moments A, B and C
are the inertial moments A, B and C with respect to the axes and are computed by the usual formulas using the molecule orientation in a standard position [2].

Maximum, average and minimum values of net atomic charges for all atoms
are computed using the net atomic charges values from the 'NET ATOMIC CHARGES' Chapter.

Maximum, average and minimum attraction force between atoms in certain chemical bonds

$$F = -100 \cdot s_1 \cdot s_2 / d^2$$

where
$s_{1,2}$ are the net atomic charges computed for the bonded atoms
d is distance between the bonded atoms

Dipole moment (and X, Y, Z components)
read directly from the MOPAC file (see 'SUM', etc.).

Molecular polarity
is the difference maximum value - minimum value of net atomic charges.

Polarity parameter
is the ratio between the molecular polarity (see above) and the distance between the atoms which posses the minimum and the maximum net atomic charge.

Heat of formation ΔH_f
is the minimum value for 'HEAT', read directly from the MOPAC output file (see 'CYCLE' Chapter, after 'INTERATOMIC DISTANCES' Chapter).

Miscellaneous functions of molecular orbital energies
are computed using the energy's value of the molecular orbitals from the Chapter 'EIGENVECTORS'.

Double occupied orbital number
read directly from the MOPAC file (see 'NO. OF FILLED LEVELS').

Maximum contribution of the bonding and antibonding molecular orbital
is computed using the values from 'BONDING CONTRIBUTION'

Free valence (minimum, maximum and average values)
is computed using the formula [25]

$$F_v = V_{max} - \Sigma B$$

where
V_{max} – is the tabulated maximum value of the oxidation state (for F atom the used value is 0.01)
ΣB – is the sum of bond orders from 'EIGENVECTORS' Chapter

Maximum and average electrophilic, nucleophilic and one-electron Fukui reactivity indices
are computed for C, N and O atoms only, using the Fukui formulas [6] and the atoms contributions to the HOMO and LUMO orbital from 'EIGENVECTORS' Chapter.

Miscellanous functions of total, electronic and core-core energies
are computed using these energy values read directly from the MOPAC output file.

Descriptors of the molecular fragments
PreclavD cuts off the fragments of the analyzed molecule [35]. The identified molecular fragments and classical chemical groups are, as a rule, identical. However, the neighbouring strong conjugated classical groups are included in the same fragment. Then, the program computes some fragments' descriptors and the percentage in weight of the fragments. These percentages are used in chemical similarity and synergy computation.

Miscellanous descriptors of cyclic topological paths (circuits)
are computed if the cyclomatic number is small enough (C < 8).
The circuits are identified using a specific PRECLAV procedure [14].
The 'hamiltonian circuits' survey all vertices of the molecular graph.
The aromaticity of circuits is computed using the TOPAZ algorithm [21].

Crowding index of circuits (cyclic topological paths) [14]
is computed if the cyclomatic number is small enough (1 < C < 8).

$$CIC = (T - C) / (2^C - C - 1)$$

where
C is the cyclomatic number
T is the number of circuits

Number of rotatable bonds
The rotatable bonds are identified using a specific PRECLAV procedure [24].

pKa descriptors
If the molecule includes O-H bonds in hydroxyl/carboxyl groups the maximum and the minimum value of pKa for hydrogen atoms attached to oxygen atoms is read directly from the MOPAC file, see the Chapter 'pKa for hydroxyl hydrogens'.

Hydrophilicity of a certain molecular fragment
is the difference Δ between the maximum value S_{max} of the net charges for hydrogen atoms and the minimum value S_{min} of the net charges for heteroatoms in the analyzed fragment.

$$\Delta = S_{max} - S_{min}$$

If the hydrogen atoms are missing $S_{max} = 0$. If the heteroatoms are missing $S_{min} = 0$. Therefore, $\Delta = 0$ for fragments that include only carbon atoms, for instance C in carbon tetrachloride, C_2 in tetrachloroethylene, C_6 in totally substituted benzene.

The descriptors *mem, mev, mes, meb, mef* and *lco* in the descript.txt file are specific to organometallic complexes [28].

4.2. Weighted '*whole molecule*' descriptors

Within the 'descript.txt' file's list the descriptors after '*kic*' (Minimum aromaticity of aromatic circuits) are considered 'weighted descriptors' of previous 'whole molecule' descriptors. The formula of weighted descriptors can be viewed in the 'descript.txt' file.

Six *Indices of gyration* are computed using formula

$$I_{gyr} = (a / b)^{1/2}$$

where
a is the Moment of inertia A, B or C
b is the Molecular mass or the Area of molecular surface

H, ..., S, P, B, As, Si percentages
is the ratio (atomic mass · number of atoms) / molecular mass.

Molecular eccentricity
is the ratio $(M^2 - m^2)^{1/2} / M$

where
M is the maximum value of inertia moments
m is the minimum value of inertia moments

Symmetry index
is the m / M ratio

where
m is the geometric mean of axis length in the circumscribed ellipsoid
M is the arithmetic mean of axis length in the circumscribed ellipsoid

Spherical shape indices, roughness and ruggedness
are computed using the procedure described in [2]. The 'peripheral atoms' are considered the atoms having maximum two bonded neighbours. The rugosity *rug* of molecular surface is the coefficient of variance of distances of peripheral atoms to surface of the circumscribed ellipsoid. The ruggedness *rgd* of molecular surface is the average distance of peripheral atoms to the surface of the circumscribed ellipsoid.

Gravitation indices
Two gravitational indices are computed (for all atoms / for bonded atoms) using the formula:

$$I = \Sigma (m_1 \cdot m_2) / r^2$$

where
m_1 and m_2 are the atomic masses
r is the distance computed based on the Cartesian co-ordinates of the atomic nuclei

Various classical (non-weighted) topological indices
are computed using the classical procedures cited in literature [3, 10, 11].

Electronic topological indices
Two electronic indices are computed using the net atomic charges for all atoms and for the bonded atoms [5].

Positive surface, non-charged (neutral) surface and negative surface
are computed using the procedure for the surface computation [2] and the values of net atomic charges from the Chapter 'NET ATOMIC CHARGES'. The atoms having the computed net atomic charge in the range [-0.05, +0.05] are considered 'neutral'.
The values of the net charges allow the calculation of the standard deviation σ for the positive charges, for the negative charges and for all charges. Then program computes a version of the 'electrostatic balance parameter' *ebp* [27].

$$ebp = 1000 \bullet \sigma_+ \bullet \sigma_- / \sigma_{all}$$

Capability to form hydrogen bonds

is computed as the sum of the net charge of the hydrogen atoms and/or heteroatoms of the analyzed molecule [17].

Donor H - Acceptor H capacity gap

The maximum net charge A of the hydrogen atoms (A > 0) in the analyzed molecule is considered the measure of the H atoms donor capacity. The minimum net charge B of the heteroatoms (B < 0) is considered the measure of the H atoms acceptor capacity.

This descriptor is defined as 2 • A + B.

Flexibility indices #1 and #2

$$f_1 = W / (N \cdot S) \quad [15]$$
$$f_2 = f_1 / (1 + p)$$

where

W – Wiener topological index
N – number of heavy atoms
S – sum of bond orders (for bonds which bond only heavy atoms)
p – percent of aromatic bonds (for bonds which bond only heavy atoms)

Rigidity (hardness) indices #1 and #2

are the inverse of flexibility indices (see above)

Shannon entropy of atomic volumes

is computed using the Shannon formula [4] and the Van der Waals atomic radius criterion to place the atoms of analyzed molecule into five classes.

- first class < 1.30 (H atoms);
- second class 1.30 - 1.64 (N, O and F atoms);
- third class 1.65 – 1.77 (B, C and Cl atoms);
- fourth class 1.78 – 1.90 (As, Br, S and P atoms);
- fifth class >1.90 (Si and I atoms).

Shannon entropies of topological distances, atomic numbers, net charges, bond orders etc.

are computed by the same Shannon formula [4], using the topological distances, atomic numbers, net charges, bond orders etc.

Molecular lipophilicities

function #1 = [1 + Ln(M)]/(1 + *ahy*)
function #2 = [1 + Ln(M)]/(1 + $Class_{water}$)
function #3 = function #1 • function #2

where M is the molecular mass, *ahy* is the average hydrophilicity of molecular fragments and $Class_{water}$ is the similarity of net charges of peripheral atoms in the analyzed molecule and water, computed using the similarity formula in Section 5.

PRECLAV logP and color

The PreclavD module identifies the size and the Type 2 class of the molecule in the calibration/prediction sets, acording to number of C atom (≤ 9 or > 9) and chemical structure, see Appendix #4, Table 2. For each class the program calculates logP using a specific QSPR [14]. PreclavD calculates quite correctly the logP's value. However, the algorithm of the EpiSuite software [26] is more accurate. The calculated values are saved into the last column of the *NAMEwho.des* table.

PreclavD module calculates the absorbed wavelength using the below formula and estimates the color using the table in the 'NOTE about colors' section.

$$\lambda_{calc} = 7800 / (E_{LUMO} - E_{HOMO}) - 600$$

The PreclavQ module can obtain a more accurate formula for the absorbed wavelength, as dependent property**.**

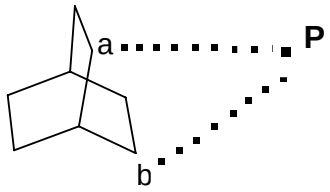
4.3 Descriptors 3D

From the viewpoint of PRECLAV's algorithm the superposition of molecules is implicit because MOPAC superposes the atoms with the same index (1, 2, and 3) in all molecules having a common skeleton as long as the atoms bearing these indices are the same in all molecules. Therefore, the superposition of molecules is the effect of the building procedure enforced in Section 2.1.

To compute the 3D descriptors the program generates a parallelepiped, that includes the group of the superposed molecules.

The sides of the box are parallel to the co-ordinate axis. Each side is divided in five equal parts. Consequently, the entire box will be divided into 125 cubic cells. In the centre of each cell, a proton (*probe atom*) is supposed to be. The probe atom distance can be at your choice (see Section 3.1.1).

Considering all atom pairs of the analysed molecule the program computes for each cell the maximum and average parallax of cell centre (i.e. maximum and average vision field for this point). Therefore, 2 • 125 = 250 geometric 3D descriptors (maximum parallaxes P and average parallaxes M) are computed for each molecule.



The parallax of the (a, b) atoms pair, seen from the point P

In addition, the program computes the attraction force, the electro-rejection force and the resultant force of all net atomic charges, for each *probe atom*. Therefore, other 3 • 125 = 375 3D descriptors (electrostatic forces A, R and F) are computed for each molecule.

The program computes the optimum value for the probe atom distance (consequently, the dimensions of parallelepiped), because the correlation between the values of 3D descriptors and the values of dependent property should be as high as possible.

If the number of overlapped heavy atoms is only two, the program 'thinks' the common skeleton of the analyzed molecules is non-existent and the 3D descriptors are not computed, regardless of the option of the user.

4.4 The conventional name of the descriptors

The pNAME / nNAME (percent in weight / number) of the molecular fragments suggests the elemental composition (e.g. *pC7H4O4, pHO, nC2H3*). Frequently, this name and the symbol of the classic chemical groups are not similar.

The symbol 'h' is used for acid hydrogen, with the calculated net charge > 0.34. Therefore, the 'ChO2' is symbol for carboxyl. Some hydroxyl groups in glucose are acid 'Oh', other are non-acid 'OH'.

…….

The name of sum fragments' numbers is (*x+y)n*, where *x* and *y* are the columns in the file NAME.vfr.des. For example, '(3+8)n' means 'the sum of numbers of the fragments in columns 3 and 8 in the file NAME.vfr.des'.

The name of sum fragments' percentages is (*x+y)p*, where *x* and *y* are the columns in NAME.vfr.des file. For example, '(3+8)p' means 'the sum of percentages of the fragments in columns 3 and 8 in NAME.vfr.des file.'

Similarly, you can understand the names (*x•y)n* and (*x•y)p* of the products.

…….

Three small letters indicate the name of a certain '*whole molecule*' descriptor (i.e. the name in the 'descript.txt' file, e.g. *bal, mas, xso*) and QSPR of the molecular fragments (*frg* descriptor). The *frg* and *qsc* descriptors are not included in the descript.txt file.

The name of 'whole molecule' sums is indicated by six small letters (i.e. name#1 + name#2, for instance *bal+xso, mas+bal*).

The name of 'whole molecule' products is indicated by six small letters (i.e. name#1 * name#2, for instance *bal*xso, mas*bal*).

…….

The name of parabolic 'whole molecule' functions is indicated by three capital letters (for instance *BAL, MAS, and XSO*).

…….

The name of the distances in Namedon.des file is 'num', 'sum', 'min', 'max' and 'ave' followed by the name of the linked groups X (halogen), P, O, S, N, C (sp3 C atom), OH, SH, NH or PH. For instance, 'sumOOH' means 'sum of the distances between O (non-linked to H) and O (linked to H) atoms', 'maxNN' means 'the maximum distance between N (non-linked to H) atoms', 'avePO' means 'the average distance between P (non-linked to H) and O (non-linked to H) atoms', 'numCO' means 'number of pairs C sp3 and O (non-linked to H) atoms', etc.

…….

The 3D descriptors name is indicated by one capital letter followed by the cell number (such as *A4, P119, R15*). The cell no. 1 is the cell having the smallest X, Y, Z Cartesian co-ordinates of centre, while the cell no. 125 is the cell having the largest X, Y, Z Cartesian co-ordinates of centre.

The name of 3D sums is indicated by name#1 + name#2, for instance *P119+R15*.

The name of 3D products is indicated by name#1 + name#2, for instance *P119*R15*.

The name of parabolic 3D functions is indicated by name + 'p' (for instance *P119p, R15p*).

…….

The name of the variables in *qsc* descriptor (31 types of atoms and 81 types of chemical bonds):

the symbol of the atoms is xZy

where

x is the type of the chemical bond (of the atom Z) with the lowest calculated bond order, i.e. *s* (single), *a* (aromatic) or *d* (double)

Z is the type of atom, i.e. H (hydrogen), A (acid hydrogen), B (boron), C (carbon), N (nitrogen), O (oxygen), F (fluorine), X (silicon), P (phosphorus), S (sufur), L (chlorine), R (bromine) or I (iodine)

y is the type of the chemical bond (of the atom Z) with the greatest calculated bond order, i.e. *s* (single), *a* (aromatic), *d* (double) or *t* (triple)

the symbol of the chemical bonds is A_1xA_2

where

A_1 is the type of the atom with the lowest mass

x is the type of the chemical bond between atoms, i.e. *s* (single), *a* (aromatic), *d* (double) or *t* (triple)

A_2 is the type of atom with the greatest mass

For instance,
sLs is the chlorine atom,
aCa is the carbon atom (non-bonded with H) in any aromatic cycle,
aNa is the nitrogen atom (non-bonded with H) in any aromatic cycle,
sCt is the carbon atom in the cyano / alkinyl group,
sAs is the acid hydrogen atom,
sCd is the carbon atom in any non-conjugated C=O group,
PaS is the aromatic bond P-S in thiotepa (RN 52-24-4),
NsP are the single bonds N-P in thiotepa,
CtN is the triple bond in the cyano/cyanate group,
CdO is the double bond in various C=O group (carbonyl, carboxyl, amide, carbamate,

isocyanate) etc.

…….

The name of the DRAGON descriptors is unchanged (e.g. *nBM, SCBO, ARR*).

…….

The names of the descriptors 'calculated by hand' and the names of the descriptors in the synergy tables are user's descriptors' names.

…….

The names of the molar refractivities are *mr01, mr02*, etc, the names of the refractivities' sums are *mr(01+02), ..., mr(03+07)*, etc. and the names of the refractivities' products are *mr(01*02), ..., mr(02*06)*, etc.

5. FORMULAS, PROCEDURES

- *linear Pearson correlation r*

$$r = S_1 / (S_2 \cdot S_3)$$

where
$S_1 = \Sigma [(X_i - X_m) \cdot (Y_i - Y_m)]$
$S_2 = \Sigma [(X_i - X_m)^2]^{1/2}$
$S_3 = \Sigma [(Y_i - Y_m)^2]^{1/2}$
X_i and Y_i are values of two variables
X_m and Y_m are the average values of X_i and Y_i

The value of the linear correlation is in the range [- 1, + 1]. If the value of the correlation between the values of (any) *X* variable and the values of (any) *Y* variable is far from zero, i.e. $r^2 \sim 1$, there are three possibilities:

- *X* is the cause of *Y*
- *Y* is the cause of *X*
- *X* and *Y* are effects of the cause *Z*

The value and the sign of *r* cannot identify the cause *X, Y* or *Z* of a certain phenomenon, but they can measure the level of the cause-effect relationship.

- *standard deviation* σ

$$\sigma = [\Sigma (X_i - X_m)^2 / (N - 1)]^{1/2}$$

If all considered values are positive the ratio $100 \bullet \sigma / X_m$ is the 'Coefficient of Variation CV%'.

- *RMS, standard error of equation (or SEE) quality function*

$$RMS = [\Sigma (V_{obs} - V_{calc})^2 / N]^{1/2}$$
$$SEE = [\Sigma (V_{obs} - V_{calc})^2 / (N - p)]^{1/2}$$
$$s_{rank} = [\Sigma (R_{obs} - R_{calc})^2 / (N - 1)]^{1/2}$$

where V are property values, *p* is the number of predictors and R are ranks of property values

The values of *standard error of equation* for values and ranks are posted as supplementary information for user.

In addition, the value of SEE is used for the identification of 'statistical outlier' molecules in the calibration set, using a classical outlier index in non-classical manner (see below). Frequently, the 'statistical outliers' are not 'chemical', 'shape', 'size', 'hydrophilicity' or 'flexibility' outliers.

The outlier index OI is the ratio $|V_{obs} - V_{calc}| / RMS$.

The molecules which present a high value of the outlier index (OI > 1.5) are marked by printing of the OI value.

The biochemical active high outliers are possible good starting points for *lead hopping* [19, 20] because they are both active as well different from other calibration set molecules. For 'lead hopping molecules' the value of OLHI (Outlier Lead Hopping Index) is large enough (OLHI > 6) [39]

$$OLHI = OI \bullet |V_i^{obs} - V_{average}^{obs}| / \sigma^{obs}$$

The molecule with the highest value of OI is automatically eliminated. The computations are repeated using the new descriptor table(s), i.e. without the row of the eliminated outlier.

The elimination of outliers is stopped if

- RMS $\leq 0.08 \bullet V_m^{obs}$ (reason #1; V_m^{obs} is the average of the observed values)
- the number of the eliminated outliers $\geq (2 \bullet N)^{1/2}$ (reason #2)

The rows of the outliers must be manually removed from the initial QSPR table used. The program displays the cause of the outliers elimination stopping. If it has been the reason #1 the new QSPR table **should not include** the last eliminated outlier. If it has been the reason #2 the new QSPR table **should include** the last eliminated outlier.

- *the average relative residue RR%*

$$RR\% = 100 \bullet \Sigma |V_{obs} - V_{calc}| / \Sigma |V_{obs}|$$

- *the average relative rank KK%*

$$KK\% = 100 \bullet \Sigma |K_{obs} - K_{calc}| / \Sigma K_{obs}$$

The denominator ΣK_{obs} is equal to $N \bullet (N+1) / 2$.

The RR% and KK% ratios are calculated only for the validation set. **The calibration set and the validation set can be identical.**

- *the square of the linear correlation r^2*

$$r^2 = r \cdot r$$

In PRECLAV, this function has a major role. It is used for:

- selection of the descriptors using their correlation with the P property values
- building the descriptors sets (as an orthogonality / co-linearity criterion for descriptors)
- QSPR selection (in quality criterion of equations)

The function Q, calculated for the molecules in the calibration set, is the main quality criterion for QSPRs [39]. The value of Q is usually in the range [1, 4].

$$Q = [r^2 / (1 - r^2)]^{1/2} \bullet [(N - p) / (N + p)]^2$$

where
p - number of predictors
N - number of molecules in the calibration set

- *the F function*

$$F = r^2 / (1 - r^2) \cdot (N - p) / p$$

- *the centralisation function*

The centralised value V_c of a certain descriptor is:

$$V_c = V_n - V_m$$

where
V_n – is the natural value of the descriptor
V_m – is the average of the natural values of the descriptor

The program computes multilinear QSPRs:

$$P = C_0 + \sum_{i=1}^{p} C_i \cdot D_i$$

where
P is the computed value of the dependent property
C_0 is intercept
C_i are weighting factors
D_i are the (values of the) significant' descriptors
p is the number of QSPR's descriptors/predictors $1 < p < 21$

In the QSPR of the molecular fragments (*frg* descriptor) P is the computed value of the *frg* descriptor, D_i are the percentages of the molecular fragments and *p* is the number of molecular fragments. As a rule, this QSPR includes all identified fragments included in the molecules of the calibration *and* prediction sets, however $p < 101$ and $p < N - N^{1/2}$.

In the QSPR of several WM descriptors (*qsc* descriptor) P is the computed value of the *qsc* descriptor, D_i are the values of variables (natural numbers, i.e. the number of various atoms/chemical bonds), $p > 1$, $p < 100$ and $p < N - N^{1/2}$. See the name of variables in 4.4 Section.

The program uses only the *p* variables which are not constant. For instance, the number of phosphorus atoms or the number of triple bonds can be zero in all molecules, the number of OH bonds can be the same in all molecules etc.

- *the method in computation of the coefficients (weighting factors)*

The Ordinary Least Square Method applied to the centralised values of predictors gives the coefficients ('weighting factors') values of the computed QSPR and parabolic functions. Solving the systems of the linear equations is done by Gauss-Jordan elimination [40].

- *relative utility of predictors* to describe the variation of dependent property values

If K = 0 the utility U value is computed by the formula:

$$U = (R^2 - r^2) / (1 - r^2)$$

where
R^2 is the square of the Pearson correlation between the experimental P values and the computed P values (using the *p* predictors QSPR)
r^2 is the square of Pearson correlation between the experimental P values and the calculated P values (using the *p-1* predictors QSPR, i.e. the equation *without* the analysed predictor)

After the computation of U for each predictor of the final QSPR, the values of U are normalised with the highest value (the highest value for U becomes 1000). The predictors with high enough value of U (U > 600) can be considered 'with high relative utility'. These predictors are useful because they correlate well with the P_{obs} values and present low correlations with other predictors. Each 'useful' predictor explains (quite) a lot of the P_{obs} variation and, at the same time, a different thing than other predictors.

- *cross-validated Pearson square linear correlation*

PRECLAV calculates r^2_{CV} using LHO (*Leave Half Out*) method **after ordering of molecules in the calibration set according to the experimental value of the dependent property**. The cross-

validated function r^2_{CV} is a measure of homogeneity of the calibration set from the point of view of the predictors' set, i.e. from the point of view of the structure-property relationship. A low value (< 0.4) of r^2_{CV} means '*the QSAR equation for molecules having high values of bio-activity and the QSAR equation for molecules having low values of bio-activity (including the same descriptors),havet very different weighting factors*'.

In addition, r^2_{CV} can be viewed as a measure of the extrapolation capacity of the obtained QSPR. The r^2_{CV} function is not used, but the value is posted as supplementary information for the user.

- chemical similarity calculations

The chemical similarity calculations use the result of molecular fragments identification, according to the PRECLAV algorithm (see Section 4.1 and [35]). The identified fragments are classified according to the below criteria #1 and #2.

If the number of heavy atoms included is < 4 the value of the criterion #1 is the number of heavy atoms included. If the number of heavy atoms included is > 3 the value of criterion #1 is 4.

The criterion #2 is the string of symbols of included elements, in alphabetical order.

If the value of criterion #1 and criterion #2 is the same the analyzed fragments are considered 'in the same class'.

Exceptionally, the fragment C is considered 'in the same class' with the fragments CH, CH_2 and CH_3. The fragments F, Cl/Br, I are identified as different fragments and are included in different classes. Also, the fragments B, N, O, S, P, As, Si, Se and Te are identified as different fragments and are included in different classes. The fragments NO and SO are 'different'. The fragments NH, OH and SH are also 'different'. The fragments in pairs NO/NO_2, SO/SO_2 and NH/NH_2 are included in the same class. The fragments NCO (di-substituted amide), OCN (cyanate) and NCO (*iso*-cyanate) are included in the same class. The fragments NHCO (substituted amide) and N_2H_nCO (substituted urea) are different because of the different value of criterion #1. All aromatic fragments C_nH_m are included in the same class (however, if m=0 the fragment is considered different). An error of this algorithm puts the fragments $–CONH_2$ and –CH=N-OH in the same class.

The program computes the percents (in weight) of each class of fragments and uses these percents in the computation of Shannon Entropy SE of the analyzed molecule. If the molecule includes just one fragment or just one class of fragments the value of SE is null.

The Chemical Structure similarity SIM_{CS} of two molecules is

$$SIM_{CS} = (SE_1+0.05) / (SE_{12}+0.05) \bullet (SE_2+0.05) / (SE_{12}+0.05)$$

If the value of SIM_{CS} is high enough the molecules can be inserted into the same 'chemical cluster'. However, the program includes a certain molecule in a certain already existent cluster only if the similarity with the molecules in the cluster is high enough and the similarity with other molecules is low enough, according to a *MiniMax* type procedure [29].

The molecules included in small enough 'chemical clusters', including maximum $0.5 \bullet (N + K)^{1/2}$ molecules, are considered 'chemical outliers'.

The 'chemical diversity' is Shannon Entropy of calibration/prediction/database weighted by Log(n) where n is the number of included molecules (n = N, n = K or n = N + K).

After 'clusterization' of the calibration set molecules, prediction set molecules and the entire Database, program computes, in the same manner, the 'chemical similarity' of calibration and prediction sets. In this final computation 'the classes' are the identified chemical clusters.

- shape similarity calculations

To compute similarity SIM_{SH} of the shape for two molecules the program uses a specific version [31] of the USR (*Ultrafast Shape Recognition*) method [30] or a method based on the superposition of molecules.

According to the corrected USR method, if the molecules are elongated or quite different from the point of view of symmetry, program uses the ratio of *gsi* descriptors' values, otherwise it uses the

USR method. The descriptors u01, u02, …, u12 in *NameSIM.des* file are the twelve 'moments' specific to USR method. The shape similarity SIM_{SH} of molecules M_1 and M_2 is $SIM_{SH} = 1/(1+m)$, where *m* is the average of Manhattan distances of moments.

The PreclavD module superposes the molecules having a common skeleton, because all molecules are placed, by translations/rotations, in a standard position:

- heavy atom #1 (see Section 2.1, page 3) in origin of Cartesian coordinates system
- heavy atom #2 on OX axis (positive value of X)
- the farthest heavy atom from OX axis in the OXY plane (positive value of Y)

The *NameSHP.des* file includes the X, Y and Z coordinates of heavy atoms in superposed molecules. For each heavy atom A_i in molecule M_1 PreclavS identifies the closest heavy atom A_j in the superposed molecule M_2 and computes the Euclidian distance d_{ij}. According to the Superposition method the shape similarity SIM_{SH} of molecules M_1 and M_2 is $SIM_{SH} = 2/(2+M)$, where M is the average of distances d_{ij}. The reliability of the Superposition method is higher if the common skeleton of the analyzed molecules is rigid.

If the value of SIM_{SH} is high enough the molecules can be inserted into the same 'shape cluster'.

- *size similarity calculations*

To compute the size similarity SIM_{SZ} of two molecules the program uses the value of the descriptor *cvo* 'COSMO volume' (see descript.txt file).

$$SIM_{SZ} = cvo_1/cvo_2$$

If $SIM_{SZ} > 1$ the program uses the inverse of this ratio.

If the value of SIM_{SZ} is high enough, the molecules can be inserted in the same 'size cluster'.

- *hydrophilicity similarity calculations*

To compute the hydrophilicity similarity SIM_{HP} of two molecules the program uses the value of the descriptors 'average hydrophilicity of fragments' *ahy* and 'maximum hydrophilicity of fragments' *xhy* (see descript.txt file).

$$SIM_{HP} = \min (r_1, r_2)$$

where $r_1 = ahy_1/ahy_2$ $r_2 = xhy_1/xhy_2$

If $r_1 > 1$ and/or $r_2 > 1$ the program uses the inverse of these ratios. Therefore, two molecules are similar from the point of view of hydrophilicity if the values of *ahy* AND *xhy* descriptors are similar. Dodecane and dodecanol present close values of the *ahy* descriptor but very different values of the *xhy* descriptor. On the contrary, methanol and dodecanol present different values of the *ahy* descriptor but very close values of the *xhy* descriptor.

If the value of SIM_{HP} is high enough the molecules can be inserted in the same 'hydrophilicity cluster'.

- *flexibility similarity calculations*

To compute the flexibility similarity SIM_{FL} of two molecules the program uses the value of the descriptor *pro* 'percentage of rotatable bonds' (see descript.txt file).

$$SIM_{FL} = pro_1/pro_2$$

If $SIM_{FL} > 1$ the program uses the inverse of this ratio. Therefore, two molecules are similar from the point of view of flexibility if the values of *pro* descriptor are similar.

If the value of SIM_{FL} is high enough the molecules can be inserted in the same 'flexibility cluster'.

The value of similarities SIM_{CS}, SIM_{SH}, SIM_{SZ}, SIM_{HP} and SIM_{FL} is in the range [0, 1].

The program prints the similarity of the calibration and prediction sets for each these five points of view.

- *the observance of the QSPR axiom*

The similarity SIM_P of P_1 and P_2 Property's values (for instance biochemical activities), for two molecules in calibration set, is, in calculation of the observance

$$SIM_P = 1 - \Delta/D$$

where
Δ is difference $|P_1 - P_2|$
D is difference $P_{max} - P_{min}$
P_{max} is the maximum value of P in the calibration set
P_{min} is the minimum value of P in the calibration set

As a rule, the program computes SIM_P and similarities SIM_{CS}, SIM_{SH}, SIM_{SZ}, SIM_{HP} and SIM_{FL} for all pairs of molecules in the calibration set. However, if the number of pairs is huge (> 10,000) the program computes the similarities using a representative sample of these pairs.

The observances (compliances) of the QSPR axiom are the Kendall rank correlations [9] between the values of SIM_P and the values of similarities, in the range [- 1, + 1], computed from the point of view of chemical structure, shape, size, hydrophilicity and flexibility.

The weighted similarity by the observances and diversities in the calibration/prediction sets is an empirical many variables mathematical function [14]. The program display the value of this COR function, i.e. the estimated corectness of the final labeling 'low / high'.

- *obtaining the new databases NDB and NVB*

In each of the above similarity calculation a certain molecule of the initial calibration set is included or not in a cluster including (some) molecules of the initial prediction set. Therefore, a certain molecule of the initial calibration set is included 0, 1, … 4, 5 times in clusters that include (some) molecules of the initial prediction set. In the same manner, a certain molecule of the initial prediction set is included 0, 1, … 4, 5 times in clusters that include (some) molecules of the initial calibration set.

The molecules of the initial calibration set included 0 times in such clusters are eliminated.

The molecules of the initial calibration set included > 0 times in such clusters are included in the new calibration set.

The molecules of the initial calibration set included > 2 times in such clusters are included in the validation set (therefore included in the new calibration set).

The molecules of the initial prediction set included > 2 times in such clusters are included in the new prediction set.

In brief [38]:

- the molecules of the initial calibration set that are very non-similar with the molecules in the initial prediction set are eliminated
- the molecules of the initial calibration set that are similar enough with the molecules in the initial prediction set are included in the validation set
- the molecules of the initial prediction set that are similar enough with the molecules in the initial calibration set are included in the new prediction set

The new database NameNDB includes the new calibration set, the included validation set and the new prediction set. The new database NameNVB includes only the validation set and the new prediction set.

- *estimated correctness from the point of view of the QSPR obtained*

A value of COR ≥ 0.75, calculated by PreclavS, does not guarantee a correct 'low/high' labeling because, sometimes, an appropriate combination of similarities, diversities and observances is not sufficient. Below **each** list of the 'low/high' labeled molecules in the prediction set, PreclavQ prints the value of two corrected COR functions [14].

If $COR_{PreclavS} < 0.75$

$$COR = - 0.1829 + 0.0021 \bullet RR\% + 0.1647 \bullet (0.5 + Rap1) / (0.5 + Rap2) + 0.0053 \bullet (0.5 + Rap1) \bullet (0.5 + Rap2) + 0.7499 \bullet COR_{PreclavS}$$

If $COR_{PreclavS} \geq 0.75$

$$COR = - 0.8563 - 0.0981 \bullet RR\% + 1.3067 \bullet (0.5 + Rap1) / (0.5 + Rap2) - 0.7220 \bullet (0.5 + Rap1) \bullet (0.5 + Rap2) + 0.7220 \bullet COR_{PreclavS}$$

where

$Rap1 = \sigma^{pre} / V_m^{pre}$

$Rap2 = |V_m^{cal} - V_m^{pre}| / V_m^{cal}$

$RR\% = 100 \bullet \Sigma |V_{obs} - V_{calc}| / \Sigma |V_{obs}|$

σ^{pre} is standard deviation of the calculated values in the prediction set

V_m^{pre} is average of the calculated values in the prediction set

V_m^{cal} is average of the calculated values in the calibration set

V_{obs} are the observed values in the calibration set

V_{calc} are the the calculated values in the calibration set

If, for all equations, one calculates COR < 0.75, there is nothing to do. There are several possible causes for this unpleasant situation:

- some values of the dependent property are wrong
- some structures used are incorrect
- the similarity of the initial calibration and prediction sets is too low
- the calibration and prediction set include two or more different classes of molecules, despite applying the scheme on page 3
- the diversity of the analyzed molecules / the values of the dependent property is too small or to great

However, a quite low value of COR can be associated with a small value of RR% for the molecules in the prediction set.

- *Property's cliffs*

Some pairs of molecules present highest/lowest value of SIM_P/SIM_{xx} ratio, where *xx* is CS, SH, SZ, HP or FL. These pairs present 'Property's cliffs', i.e. they include molecules having high *xx* similarity compared to low Property similarity or high Property similarity compared to low *xx* similarity. Actually, the presence of 'cliffs' emphasizes the violation of the QSPR axiom for some molecules.

The program calculates the 'property cliff' P_{cliff} for all pairs of molecules in the calibration set,

using the fuve similarities SIM_{xx} and (other formula for) the similarity SIM_{DP} of the studied property's values P_i and P_j.

$$SIM_{DP} = 2 \bullet \min(P_i, P_j) / [\min(P_i, P_j) + \max(P_i, P_j)]$$

$$P_{cliff} = (SIM_{xx})^3 / SIM_{DP}$$

The value of SIM_{DP} is small if the difference $|P_i - P_j|$ is large. The value of P_{cliff} is large if the value of SIM_{DP} is small and the value of SIM_{xx} is large (the value of P_{cliff} is incorrect large for bioactivity of the enantiomers).

- *Applicability Domain (AD)*

This version of PRECLAV computes the leverages h_i, using the file '*NAMEprd.dat*' (see section 3.3.3) and the formula

$$h_i = x_i \bullet (M^T \bullet M)^{-1} \bullet x_i^T$$

where x_i is a row vector for a particular molecule in the prediction set and M is the N • *p* matrix of *p* predictors for N calibration set molecules. If $h_i > 3 \bullet p / N$ the molecule in the prediction set is considered, in a classic manner, 'outside of AD of the obtained QSPR'.

The calculation of the Applicability Domain (AD) and the calculation of the leverages **cannot provide reliable information regarding the real values of the dependent property for the molecules in the prediction set.** The values h_i of the leverages for the prediction set molecules are quite well correlated only with the differences between the average P_{cal} (calculated for molecules in the calibration set) and the values of P_{pre} (calculated for molecules in the prediction set). For molecules in the prediction set 'outside of AD' the difference between the calculated value P_{pre} and average of P_{cal} is, as a rule, great.

To identify the ADOs (Applicability Domain Outsiders) the program uses a method not yet published, as below.

Using the calculated values of the dependent property for molecules in the calibration set the program calculates the average P_{ave} of these values.

Using the calculated values P_i of the dependent property for molecules in the prediction set program calculates the differences $D_i = |P_{ave} - P_i|$, the average D_{ave} of these differences and standard deviation S_D of these differences.

The ADO molecules *by differences* are the molecules having $D_i > D_{ave} + S_D$.

Using the leverages l_i for the molecules in the prediction set the program calculates the average L_{ave} of these values, the differences $L_i = |L_{ave} - l_i|$, the average D_{Lave} of these differences and the standard deviation S_L of these differences.

The ADO molecules *by leverages* are the molecules having $L_i > D_{Lave} + S_L$.

The ADO molecules are the molecules which are ADO 'by differences' *and* ADO 'by leverages'.

- *Synergy* of two descriptors *i* and *j* is computed [33] using the V_i and V_j values of descriptors and the values of the dependent property P. Before computation of synergy S all these values are normalized within [0, 1] range. The value of synergy is in the range [0, 1].

$$S = (A^2 - B^2 - C^2 + D^2 \bullet E)^{1/3}$$

where
A is linear correlation of the values of $V_i + V_j$ sum and the values of P
B is linear correlation of the values of V_i and the values of P
C is linear correlation of the values of V_j and the values of P
D is linear correlation of the values of V_j and the values of V_j
E is the minimum value in pair (B^2, C^2)

- *Drug Design Hope* index DDH, based on results of synergy computation, is the sum B + C + S, see above. The value of DDH index is in the range [-3, + 3].

The descriptors in synergy computation can be the molar refractivities mr_i and mr_j of the substituents grafted in positions *i* and *j* on the common skeleton or the percentages p_i and p_j of molecular fragments, see Section 2.6.

The position pairs (*i*, *j*) or the molecular fragment pairs (p_i, p_j) having DDH very different from zero are thought to be interesting in drug design.

NOTE about colors

If the dependent property in your QSCR (*Quantitative Structure Color Relationship*) study is 'color', the QSCR table must include the values of λ (certain wavelength in the visible spectrum, presumed to be correlated with the color). Finally, you can roughly estimate the color of the substances in the prediction set using the calculated values λ_{calc} and the below empirical table.

λ_{calc}	Color	λ_{calc}	Color
< 380	colorless	543 – 565	violet
380 – 416	greenish – yellow	566 – 616	violet/purple/blue
417 – 434	yellow	617 - 638	blue
435 – 470	yellow/brown/orange	639 – 682	blue/cyan/green
471 – 488	orange	683 – 704	green
489 – 506	red	705 – 740	yellowish – green
507 – 542	red/magenta/violet	> 740	colorless

The wavelength called λ_{max} is the wavelength in the UV-VIS spectrum, i.e. in the range [200, 800] nm, at which the absorption is maximum. The value of λ_{max} is quoted in literature for many substances.

Attention! There are many colored substances with λ_{max} *outside* visible range [380, 740]. These substances are colored because there are some absorbances *inside* this range. Therefore, the values of λ used in QSCR table can be different from the cited λ_{max} and you must choose it very carefully in the range [380, 740].

If the UV-VIS spectra for the molecules in the calibration set are not available, you should use, in the QSCR table, the average above value of λ, for each visually observed color.

For instance

for yellow you should use ~ 426, for magenta ~ 525, for green ~ 694 etc.

for colorless (without conjugated/condensed aromatic cycles) you should use ~ 200 – 290

for colorless (2 - 3 conjugated/condensed aromatic cycles) you should use ~ 290 – 380

for colorless (> 3 conjugated/condensed cycles) you should use ~ 750 – 850

6. ALGORITHM IN THE COMPUTATION OF QSPRs

The algorithm of QSPR computation includes the following steps.

Selection of the 'significant' variables

The program eliminates as 'non-significant' the 'constant' and 'near-constant' descriptors, i.e. the descriptors having low diversity of values (DIV < 0.15) for the calibration set molecules. The diversity is the Shannon Entropy SE weighted by logarithm of N. This criterion is not applied in the elimination of variables included in *frg* and *qsc* descriptors (QSPR of the molecular fragments and QSPR of several WM descriptors).

The variables having high enough diversity of values are considered 'significant' only if their quality *q* is high enough.

$$q > 1$$

where

$q = (1 - Z \cdot \text{minr}^2) \cdot / (1 - r^2)$

Actually, q > 1 if $r^2 > \text{minr}^2$. In above formula Z = 2 for 3D descriptors and Z = 1 for other descriptors. Therefore, the selection of 3D descriptors is more drastic. This criterion is applied also in the elimination of the *frg* descriptor itself.

Then, the selected descriptors, see Section 3.3.3, and the descriptors in NAMEvfr.des file are used in the QSPR computation phase.

Build-up of the two descriptors sets

All pairs of 'significant' descriptors are built using descriptors that have a low intercorrelation $r^2 < 4 \bullet N^{-0.5}$ (see Section 3.3.2.). For each pair, a QSPR is computed. Then, the quality criterion Q is applied and the pairs are ordered. The best 1000 pairs are used further. However, the program does not use further the two descriptor sets having 'too low' quality ($r^2 < 2 \bullet \text{minr}^2$).

The construction of the *p* descriptors sets

By adding one descriptor to a set of *p* descriptors, PreclavQ obtains a set with *p* + 1 descriptors. For each set, the program computes a QSPR and applies the quality criterion. Then, it orders the sets (*forward stepwise* procedure [36]) and identifies the best 1000 sets (by Q, see Section 5). The best 1000 sets (by Q) are used further.

The process of build-up / selection of the sets is interrupted in three situations:

- the value of Q decreased before the number of added descriptors becomes 20 (the most common situation)
- the number of added descriptors becomes 20 (very rare situation)
- no more orthogonal descriptors can be added (almost impossible)

Identification of the XY type outlier molecules

Using the best QSPR (by Q) equation the program computes the value of the dependent property for the calibration set molecules and identifies the outliers (see Section 5). After elimination of all 'high outliers' [39] the program estimates the value of the dependent property for the prediction set molecules.

Estimation of the P value for the K molecules in the prediction set

In the final group of the best (ordered by Q) 1000 equations the program identifies four equations:

- the equation with the highest value of Q (the first equation)
- the equation with the lowest value of RR%
- the equation with the lowest value of KK%
- the equation with the highest value of COR

Some above equations can be identical. In the prediction for the molecules in the prediction set the program uses only these equations. The estimated values are centralized, transformed in natural estimated values and saved as *Name*-EST file (without extension).

The program computes the average value P_{calc}^{m} of the estimated values and the standard deviation σ of the estimated values.

The program considers 'high' the values fulfilling the criterion $P_{calc} > P_{calc}^{m} + \sigma/2$ considering only the molecules in the prediction set and 'HIGH' the values fulfilling the same criterion considering all molecules in the database (calibration set+prediction set). Similarly, the program considers 'low' the values fulfilling the criterion $P_{calc} < P_{calc}^{m} - \sigma/2$ considering only the molecules in the prediction set and 'LOW' the values fulfilling the same criterion considering all molecules in the database. Therefore, the estimated values are 'high', 'low', 'HIGH' or 'LOW' **not in absolute manner, but relative to the other values** in the prediction set or database.

The program presents the list of the K prediction set molecules. There are some molecules labelled using 'high', 'low', 'HIGH' or 'LOW' words. If the user wishes to synthesize molecules having, for instance, high biochemical activity, the molecules labelled 'high' and/or 'HIGH' are 'recommended for synthesis', while the ones labelled 'low' and/or 'LOW' are 'not recommended for synthesis'. If the user wishes to synthesize molecules having low toxicity, the molecules labelled 'low' and/or 'LOW' are 'recommended for synthesis', while the ones labelled 'high' and/or 'HIGH' are 'not recommended for synthesis'.

Finally, the program offers a final *combined* 'low' / 'high' (not LOW / HIGH) labelling, from the point of view of all equations used for prediction.

Synthesis of the ADO molecules, marked in the *Name*-EST file, is the responsibility of the user and can be a useful option.

From the point of view of some practical purposes, the labelling of the prediction set molecules is the main result of all PRECLAV computations. However, the user can utilize other data in the *rezultat.rez* file to … write a scientific paper.

7. FINAL COMMENTS

In brief, PRECLAV offers some major information:

In the absence of the prediction set

- the identification of the *outliers for lead hopping* molecules
  (a possible good starting point for the development of a new bioactive class of molecules)
- the identification of the molecular features having the largest influence on the dependent property
  (useful in *drug design*)

In the presence of the prediction set

- the identification of the molecules 'recommended for synthesis' in the prediction set
  (useful to decrease the cost of the research)

According to the scheme on page 3, the database used in ADO step has two characteristics:

- the calibration set does not include outliers
- the similarity of the calibration and prediction sets is high enough

Beside the QSPR studies the program offers information, possible catchy in drug design, concerning the synergistic/antagonistic intramolecular effects

- of the substituents grafted in certain positions on the common skeleton
- of the included molecular fragments

The abilities of the program in synergy computation are also useful in analysing mixtures.

8. REFERENCES

1. BONDI A., *J. Phys. Chem.,* **80**, 1966, p. 3006
2. TARKO L., CALAFETEANU S., *Rev. Chim.,* **49**, 1998, p.169
3. RANDIC M., *J. Am. Chem. Soc*., **97**, 1975, p. 6609
4. C. E. SHANNON, *Bell System Technical Journal*, **27**, 379 (1948)
5. OSMIALOWSKI K., HALKIEWICZ J., KALISZAN R., *J. Chromatogr.,* **63**, 1986, p. 361
6. FUKUI K., *Theory of Orientation and Stereoselection*, Springer-Verlag, Berlin, 1975
7. MOPAC program is available from Stewart, J.J.P., 15210 Paddington Circle, Colorado Springs, CO 80921; E-mail: MrMOPAC@OpenMOPAC.net; http://www.openmopac.net/
8. GILBERT K., GAJEWSKI J.J., PCModel v. 9.0, SerenaSoftware, Box 3076, Bloomington, IN, USA
9. ANDREI T., STANCU S., *Statistica*, Ed. ALL, Bucuresti, 1995, p. 341
10. BALABAN A.T., *Chem. Phys .Lett.,* **89**, 1982, p. 399
11. DIUDEA M.V., IVANCIUC O., *Topologie moleculara*, Ed. Comprex, Cluj, 1995
12. TARKO L., BRUCKNER A., FILIP P., *Rev. Chim.,* **53**, 2002, p. 619
13. TARKO L., FILIP P., *Rev. Roum. Chim.*, **48**, 2003, p. 745
14. **not yet published**
15. TARKO L., *Rev. Chim.*, **55**, 2004, p. 169
16. TARKO L., *Rev. Chim.,* **55**, 2004, p. 539
17. TARKO L., *Rev. Roum. Chim.*, **51**, 2006, p. 463
18. DRAGON program is available from *Talete srl*., via V. Pisani, 13-20124, Milano, Italy; http://www.talete.mi.it
19. CRAMER R. D., JILEK R. J., GUESSREGEN S., CLARK S. J., WENDT B., CLARK R. D., *J. Med. Chem.* **47**, 2004, p. 6777
20. SAEH J. C., LYNE P. D., TAKASAKI B. K., COSGROVE D. A., *J. Chem. Inf. Comput. Sci.,* **45**, 2005, p. 1122
21. TARKO L., *ARKIVOC*, **2008**, Part xi, p. 24
22. MMX is the up-dated version of MM2 algorithm
23. ALLINGER N. L., YUH Y. H., LII J.-J., *J. Am. Chem. Soc.*, **99**, 1977, p. 8127
24. TARKO L., *Rev. Chim.,* **62**, 2011, p. 135
25. SANNIGRAHI A. B., *Adv. Quantum Chem.*, **23,** 1992, p. 301
26. EPISuite®; http://www.epa.gov/oppt/exposure/pubs/episuitedl.htm
27. MURRAY J. S., LANE P., BRINCK T., POLITZER P., *J. Phys. Chem.,* **97**, 1993, p. 5144
28. J. D. WALKER, M. C. NEWMAN, M. ENACHE *et al*., *Fundamental QSARs for Metal Ions*, CRC Press, Boca Raton, London, New York, 2012, p. 97 – 159
29. TARKO L., *J. Math. Chem.*, **52**, 2014, p. 948
30. P. J. BALLESTER, W. G. RICHARDS, *Proc. R. Soc. A,* **463**, 2007, p. 1307
31. TARKO L., *J. Math. Chem.*, **53**, 2015**,** p. 1576
32. TARKO L., *MATCH*, **75**, 2016**,** p. 511
33. TARKO L., *MATCH*, **75**, 2016**,** p. 533
34. TARKO L., *MATCH*, **77**, 2017**,** p. 245
35. TARKO L., *MATCH*, **78**, 2017**,** p. 565
36. DRAPER, N., SMITH, H. (1981) *Applied Regression Analysis, 2d Edition,* New York: John Wiley & Sons, Inc.
37. TARKO L. STECOZA C. E., ILIE C., CHIFIRIUC M. C. Chifiriuc *Rev. Chim.,*, **60**, 2009, p. 476
38. HRUBARU M., TARKO L., *Rev. Chim. (Bucharest)*, **70**, 2019, p. 887
39. TARKO L., *J. Math. Chem.*, **57**, 2019, p. 1770
40. GHEORGHIU M., FILIP P., *Aplicatiile calculatoarelor în chimie*, Ed. Ştiinţifică Bucureşti, 1973, p. 99

Appendix #1 specimen of MOPAC input file

```
pm6 gnorm=0.2 geo-ok pulay bonds vectors mmok
N-methyl-4-fluoro-phenyl-carbamate

 C    0.000000  0    0.000000  0    0.000000  0   0   0   0
 C    1.401512  1    0.000000  0    0.000000  0   1   0   0
 C    1.400523  1  121.445152  1    0.000000  0   2   1   0
 C    1.396071  1  120.040779  1   -0.320206  1   3   2   1
 C    1.466279  1  119.179604  1   -0.223812  1   4   3   2
 C    1.392272  1  119.440834  1    0.184517  1   5   4   3
 O    1.353815  1  117.849220  1  176.613722  1   1   2   3
 C    1.358180  1  122.700073  1  156.711903  1   7   1   2
 O    1.230559  1  122.674034  1  -26.422155  1   8   7   1
 N    1.370438  1  120.531723  1  155.169674  1   8   7   1
 C    1.453791  1  122.038651  1    1.561726  1  10   8   7
 F    1.351048  1  120.498581  1 -179.960433  1   4   3   2
 H    0.959138  1  118.300819  1  176.448186  1  10   8   7
 H    1.102992  1  119.286385  1  179.734600  1   2   1   3
 H    1.102145  1  120.119095  1  179.684097  1   3   2   1
 H    1.101813  1  119.977875  1 -179.642269  1   5   4   3
 H    1.101790  1  118.705589  1 -179.253495  1   6   5   4
 H    1.113954  1  110.742233  1  -48.379849  1  11  10   8
 H    1.113229  1  109.125313  1 -167.698946  1  11  10   8
 H    1.113674  1  109.793533  1   73.049400  1  11  10   8
```

Appendix #2 specimen of QSPR table

Cases	Toxic
pirid1	1
pirid2	1.017
pirid3	1.062
pirid4	2.545
pirid5	3.681
pirid6	4.009
pirid7	5.842
pirid8	1.118
pirid9	2.473
pirid10	7.708
pirid11	8.734
pirid12	6.094
pirid13	10
pirid14	?
pirid15	?
pirid16	?

Appendix #3 Molar Refractivity for some substituents

Substituent	grafted on non-aromatic system*	grafted on aromatic system**	Substituent	grafted on non-aromatic system*	grafted on aromatic system**
H	0.09	0.11	N_3	0.92	1.03
C	0.27	0.24	NO	0.51	0.61
CH	0.37	0.35	NO_2	0.69	0.74
CH_2	0.47	0.46	SO	0.86	0.92
CH_3	0.57	0.57	SO_2	1.09	1.01
ethyl	1.04	1.03	SO_3	1.33	1.08
propyl	1.51	1.49	SO_3H	1.43	1.19
iso-propyl	1.51	1.49	C = N	0.76	0.75
butyl	1.98	1.97	CH = NH	0.96	0.97
tert-butyl	1.98	1.95	C ≡ N	0.55	0.59
cyclopentyl	2.25	2.21	C = O	0.44	0.52
cyclohexyl	2.72	2.67	CHO	0.54	0.63
CH = CH	0.91	1.07	C = S	1.17	1.24
$CH = CH_2$	1.01	1.18	CH = S	1.27	1.35
allyl	1.47	1.46	PO	0.55	0.59
C ≡ C	0.72	0.73	PO_2	0.73	0.77
C ≡ CH	0.82	0.84	PO_3	0.91	0.94
CH = C = CH	1.39	1.53	PO_3H	1.01	1.05
$CH = C = CH_2$	1.49	1.64	COO	0.62	0.69
phenyl	2.54	2.54	COOH	0.72	0.81
naphthalenyl	4.32	4.33	NHCO	0.81	0.88
azulenyl	4.32	4.33	NHCOO	0.98	1.08
anthracenyl	6.11	6.11	NCO	0.97	1.08
F	0.11	0.08	OCN	0.72	0.74
Cl	0.58	0.57	NCS	1.62	1.72
Br	0.86	0.85	SCN	1.35	1.39
I	1.39	1.38	2-tetrahydro-furanyl	1.89	1.91
CF_3	0.61	0.58	1-piperidinyl	2.56	2.57
CCl_3	2.01	1.99	2-piperidinyl	2.53	2.55
N	0.24	0.29	dioxanyl	2.06	2.06
NH	0.34	0.41	1-morpholinyl	2.25	2.27
NH_2	0.44	0.51	2-morpholinyl	2.23	2.24
$N(CH_3)_2$	1.38	1.43	1-piperazinyl	2.43	2.44
O	0.14	0.16	2-piperazinyl	2.39	2.42
OH	0.24	0.27	2-furanyl	1.77	1.77
$O\text{-}CH_3$	0.71	0.73	1-pyrrolyl	2.11	2.21
O-allyl	1.61	1.63	2-pyrrolyl	1.98	1.99
S	0.78	0.79	2-thiophenyl	2.38	2.38
SH	0.88	0.91	2-oxazolyl	1.58	1.58
$S\text{-}CH_3$	1.35	1.36	2-thiazolyl	2.19	2.19
S-allyl	2.25	2.26	1-imidazolyl	1.95	2.05
Si	0.64	0.69	2-imidazolyl	1.79	1.81
SiH_3	0.94	1.02	2-pyridinyl	2.35	2.35
$Si(CH_3)_3$	2.35	2.41	1-indolyl	3.69	3.81
N = N	0.59	0.71	2-indolyl	3.77	3.77

* for instance in sterols

** for instance on indole

You can compute molar refractivities MR for other substituents using additiveness. For instance, MR of 4-cyclohexenyl, grafted on non-aromatic system, should be

$$MR_{4\text{-cyclohexenyl}} = MR_{CH} + 3 \bullet MR_{CH2} + MR_{CH=CH} = 0.37 + 3 \bullet 0.47 + 0.91 = 2.69$$

If the last digit of the sum is 0, replace the last digit by 1. Thus PreclavD will correctly read the table *NAMEmor.des*.

Appendix #4 Chemical classes

Some PRECLAV descriptors, see descript.txt file, are used to define 'the chemical classes'.

ohs: Number of OH bonds
nhs: Number of NH bonds
noa: Number of aromatic N-O bonds
nod: Number of double N-O bonds
soa: Number of aromatic S-O bonds
sod: Number of double S-O bonds
cnd: Number of double C-N bonds
cnt: Number of triple C-N bonds
nrp: Number of P atoms
nrf: Number of F atoms
nrl: Number of Cl atoms
nrb: Number of Br atoms
nri: Number of I atoms
ata: Number of atoms in aromatic circuits
hta: Number of heteroatoms in aromatic circuits
nro: number of O atoms
nrn: number of N atoms
nnh: number of N atoms linked to H

See seven 'indices of Type 1 classes' in the head of the below Table 1. These indices are descriptors presumed to be significant for the biochemical activity. There, sgn (*x*) is the algebraic sign of (*x*); here sgn(x) is 0 or 1. Each index has two or three values. Consequently, there are 3 • 2 • 2 • 2 • 2 • 2 • 3 = 288 combinations and so, there are 288 'Type 1 chemical classes'.

Table 2 includes only 16 'Type 2 chemical classes', which are defined according to the presence / absence of the O/N atoms (non-linked to H) and O/N atoms (linked to H).

Actually:

ohs + sgn (nhs) = 0 means 'there are no OH or NH bonds'
ohs + sgn (nhs) = 1 means 'there is 1 OH (maybe in COOH) or many NH bond (there are no NH_2 groups)'
ohs + sgn (nhs) > 1 means 'there are 2 or more OH or NH bonds, possible 1 NH_2 group'
sgn (noa+nod) = 0 means 'there are no N=O or NO_2 groups'
sgn (noa+nod) = 1 means 'there are N=O or NO_2 groups'
sgn (soa+sod) = 0 means 'there are no S=O or SO_2 groups'
sgn (soa+sod) = 1 means 'there are S=O or SO_2 groups'
sgn (cnd+cnt) = 0 means 'there are no C=N or C≡N bonds'
sgn (cnd+cnt) = 1 means 'there are C=N or C≡N bonds'
sgn (nrp) = 0 means 'there are no P atoms'
sgn (nrp) = 1 means 'there are P atoms'
sgn (nrf+nrl+nrb+nri) = 0 means 'there are no halogen atoms'
sgn (nrf+nrl+nrb+nri) = 1 means 'there are some (any) halogen atoms'
sgn (ata) + sgn (hta) = 0 means 'there are no aromatic cycles'
sgn (ata) + sgn (hta) = 1 means 'there are aromatic cycles, but not aromatic heterocycles '
sgn (ata) + sgn (hta) = 2 means 'there are aromatic heterocycles'

Table 1 **Type 1 chemical classes**

Class	ohs + sgn (nhs)	sgn (noa+ nod)	sgn (soa+ sod)	sgn (cnd+ cnt)	sgn (nrp)	sgn (nrf+ nrl+ nrb+ nri)	sgn(ata) + sgn(hta)
1	0	0	0	0	0	0	0
2	0	1	0	0	0	0	0
3	0	0	0	1	0	0	0
4	0	0	0	0	1	0	0
5	0	0	0	0	0	1	0
6	0	1	0	1	0	0	0
7	0	1	0	0	1	0	0
8	0	1	0	0	0	1	0
9	0	0	0	1	1	0	0
10	0	0	0	1	0	1	0
11	0	0	0	0	1	1	0
12	0	1	0	1	1	0	0
13	0	1	0	1	0	1	0
14	0	1	0	0	1	1	0
15	0	0	0	1	1	1	0
16	0	1	0	1	1	1	0
17	0	0	0	0	0	0	1
18	0	1	0	0	0	0	1
19	0	0	0	1	0	0	1
20	0	0	0	0	1	0	1
21	0	0	0	0	0	1	1
22	0	1	0	1	0	0	1
23	0	1	0	0	1	0	1
24	0	1	0	0	0	1	1
25	0	0	0	1	1	0	1
26	0	0	0	1	0	1	1
27	0	0	0	0	1	1	1
28	0	1	0	1	1	0	1
29	0	1	0	1	0	1	1
30	0	1	0	0	1	1	1
31	0	0	0	1	1	1	1
32	0	1	0	1	1	1	1
33	0	0	0	0	0	0	2
34	0	1	0	0	0	0	2
35	0	0	0	1	0	0	2
36	0	0	0	0	1	0	2
37	0	0	0	0	0	1	2
38	0	1	0	1	0	0	2
39	0	1	0	0	1	0	2
40	0	1	0	0	0	1	2
41	0	0	0	1	1	0	2
42	0	0	0	1	0	1	2
43	0	0	0	0	1	1	2
44	0	1	0	1	1	0	2
45	0	1	0	1	0	1	2

46	0	1	0	0	1	1	2
47	0	0	0	1	1	1	2
48	0	1	0	1	1	1	2
49	1	0	0	0	0	0	0
50	1	1	0	0	0	0	0
51	1	0	0	1	0	0	0
52	1	0	0	0	1	0	0
53	1	0	0	0	0	1	0
54	1	1	0	1	0	0	0
55	1	1	0	0	1	0	0
56	1	1	0	0	0	1	0
57	1	0	0	1	1	0	0
58	1	0	0	1	0	1	0
59	1	0	0	0	1	1	0
60	1	1	0	1	1	0	0
61	1	1	0	1	0	1	0
62	1	1	0	0	1	1	0
63	1	0	0	1	1	1	0
64	1	1	0	1	1	1	0
65	1	0	0	0	0	0	1
66	1	1	0	0	0	0	1
67	1	0	0	1	0	0	1
68	1	0	0	0	1	0	1
69	1	0	0	0	0	1	1
70	1	1	0	1	0	0	1
71	1	1	0	0	1	0	1
72	1	1	0	0	0	1	1
73	1	0	0	1	1	0	1
74	1	0	0	1	0	1	1
75	1	0	0	0	1	1	1
76	1	1	0	1	1	0	1
77	1	1	0	1	0	1	1
78	1	1	0	0	1	1	1
79	1	0	0	1	1	1	1
80	1	1	0	1	1	1	1
81	1	0	0	0	0	0	2
82	1	1	0	0	0	0	2
83	1	0	0	1	0	0	2
84	1	0	0	0	1	0	2
85	1	0	0	0	0	1	2
86	1	1	0	1	0	0	2
87	1	1	0	0	1	0	2
88	1	1	0	0	0	1	2
89	1	0	0	1	1	0	2
90	1	0	0	1	0	1	2
91	1	0	0	0	1	1	2
92	1	1	0	1	1	0	2
93	1	1	0	1	0	1	2
94	1	1	0	0	1	1	2
95	1	0	0	1	1	1	2
96	1	1	0	1	1	1	2
97	> 1	0	0	0	0	0	0
98	> 1	1	0	0	0	0	0

99	> 1	0	0	1	0	0	0
100	> 1	0	0	0	1	0	0
101	> 1	0	0	0	0	1	0
102	> 1	1	0	1	0	0	0
103	> 1	1	0	0	1	0	0
104	> 1	1	0	0	0	1	0
105	> 1	0	0	1	1	0	0
106	> 1	0	0	1	0	1	0
107	> 1	0	0	0	1	1	0
108	> 1	1	0	1	1	0	0
109	> 1	1	0	1	0	1	0
110	> 1	1	0	0	1	1	0
111	> 1	0	0	1	1	1	0
112	> 1	1	0	1	1	1	0
113	> 1	0	0	0	0	0	1
114	> 1	1	0	0	0	0	1
115	> 1	0	0	1	0	0	1
116	> 1	0	0	0	1	0	1
117	> 1	0	0	0	0	1	1
118	> 1	1	0	1	0	0	1
119	> 1	1	0	0	1	0	1
120	> 1	1	0	0	0	1	1
121	> 1	0	0	1	1	0	1
122	> 1	0	0	1	0	1	1
123	> 1	0	0	0	1	1	1
124	> 1	1	0	1	1	0	1
125	> 1	1	0	1	0	1	1
126	> 1	1	0	0	1	1	1
127	> 1	0	0	1	1	1	1
128	> 1	1	0	1	1	1	1
129	> 1	0	0	0	0	0	2
130	> 1	1	0	0	0	0	2
131	> 1	0	0	1	0	0	2
132	> 1	0	0	0	1	0	2
133	> 1	0	0	0	0	1	2
134	> 1	1	0	1	0	0	2
135	> 1	1	0	0	1	0	2
136	> 1	1	0	0	0	1	2
137	> 1	0	0	1	1	0	2
138	> 1	0	0	1	0	1	2
139	> 1	0	0	0	1	1	2
140	> 1	1	0	1	1	0	2
141	> 1	1	0	1	0	1	2
142	> 1	1	0	0	1	1	2
143	> 1	0	0	1	1	1	2
144	> 1	1	0	1	1	1	2
145	0	0	1	0	0	0	0
146	0	1	1	0	0	0	0
147	0	0	1	1	0	0	0
148	0	0	1	0	1	0	0
149	0	0	1	0	0	1	0
150	0	1	1	1	0	0	0
151	0	1	1	0	1	0	0

152	0	1	1	0	0	1	0
153	0	0	1	1	1	0	0
154	0	0	1	1	0	1	0
155	0	0	1	0	1	1	0
156	0	1	1	1	1	0	0
157	0	1	1	1	0	1	0
158	0	1	1	0	1	1	0
159	0	0	1	1	1	1	0
160	0	1	1	1	1	1	0
161	0	0	1	0	0	0	1
162	0	1	1	0	0	0	1
163	0	0	1	1	0	0	1
164	0	0	1	0	1	0	1
165	0	0	1	0	0	1	1
166	0	1	1	1	0	0	1
167	0	1	1	0	1	0	1
168	0	1	1	0	0	1	1
169	0	0	1	1	1	0	1
170	0	0	1	1	0	1	1
171	0	0	1	0	1	1	1
172	0	1	1	1	1	0	1
173	0	1	1	1	0	1	1
174	0	1	1	0	1	1	1
175	0	0	1	1	1	1	1
176	0	1	1	1	1	1	1
177	0	0	1	0	0	0	2
178	0	1	1	0	0	0	2
179	0	0	1	1	0	0	2
180	0	0	1	0	1	0	2
181	0	0	1	0	0	1	2
182	0	1	1	1	0	0	2
183	0	1	1	0	1	0	2
184	0	1	1	0	0	1	2
185	0	0	1	1	1	0	2
186	0	0	1	1	0	1	2
187	0	0	1	0	1	1	2
188	0	1	1	1	1	0	2
189	0	1	1	1	0	1	2
190	0	1	1	0	1	1	2
191	0	0	1	1	1	1	2
192	0	1	1	1	1	1	2
193	1	0	1	0	0	0	0
194	1	1	1	0	0	0	0
195	1	0	1	1	0	0	0
196	1	0	1	0	1	0	0
197	1	0	1	0	0	1	0
198	1	1	1	1	0	0	0
199	1	1	1	0	1	0	0
200	1	1	1	0	0	1	0
201	1	0	1	1	1	0	0
202	1	0	1	1	0	1	0
203	1	0	1	0	1	1	0
204	1	1	1	1	1	0	0

205	1	1	1	1	0	1	0
206	1	1	1	0	1	1	0
207	1	0	1	1	1	1	0
208	1	1	1	1	1	1	0
209	1	0	1	0	0	0	1
210	1	1	1	0	0	0	1
211	1	0	1	1	0	0	1
212	1	0	1	0	1	0	1
213	1	0	1	0	0	1	1
214	1	1	1	1	0	0	1
215	1	1	1	0	1	0	1
216	1	1	1	0	0	1	1
217	1	0	1	1	1	0	1
218	1	0	1	1	0	1	1
219	1	0	1	0	1	1	1
220	1	1	1	1	1	0	1
221	1	1	1	1	0	1	1
222	1	1	1	0	1	1	1
223	1	0	1	1	1	1	1
224	1	1	1	1	1	1	1
225	1	0	1	0	0	0	2
226	1	1	1	0	0	0	2
227	1	0	1	1	0	0	2
228	1	0	1	0	1	0	2
229	1	0	1	0	0	1	2
230	1	1	1	1	0	0	2
231	1	1	1	0	1	0	2
232	1	1	1	0	0	1	2
233	1	0	1	1	1	0	2
234	1	0	1	1	0	1	2
235	1	0	1	0	1	1	2
236	1	1	1	1	1	0	2
237	1	1	1	1	0	1	2
238	1	1	1	0	1	1	2
239	1	0	1	1	1	1	2
240	1	1	1	1	1	1	2
241	> 1	0	1	0	0	0	0
242	> 1	1	1	0	0	0	0
243	> 1	0	1	1	0	0	0
244	> 1	0	1	0	1	0	0
245	> 1	0	1	0	0	1	0
246	> 1	1	1	1	0	0	0
247	> 1	1	1	0	1	0	0
248	> 1	1	1	0	0	1	0
249	> 1	0	1	1	1	0	0
250	> 1	0	1	1	0	1	0
251	> 1	0	1	0	1	1	0
252	> 1	1	1	1	1	0	0
253	> 1	1	1	1	0	1	0
254	> 1	1	1	0	1	1	0
255	> 1	0	1	1	1	1	0
256	> 1	1	1	1	1	1	0
257	> 1	0	1	0	0	0	1

258	> 1	1	1	0	0	0	1
259	> 1	0	1	1	0	0	1
260	> 1	0	1	0	1	0	1
261	> 1	0	1	0	0	1	1
262	> 1	1	1	1	0	0	1
263	> 1	1	1	0	1	0	1
264	> 1	1	1	0	0	1	1
265	> 1	0	1	1	1	0	1
266	> 1	0	1	1	0	1	1
267	> 1	0	1	0	1	1	1
268	> 1	1	1	1	1	0	1
269	> 1	1	1	1	0	1	1
270	> 1	1	1	0	1	1	1
271	> 1	0	1	1	1	1	1
272	> 1	1	1	1	1	1	1
273	> 1	0	1	0	0	0	2
274	> 1	1	1	0	0	0	2
275	> 1	0	1	1	0	0	2
276	> 1	0	1	0	1	0	2
277	> 1	0	1	0	0	1	2
278	> 1	1	1	1	0	0	2
279	> 1	1	1	0	1	0	2
280	> 1	1	1	0	0	1	2
281	> 1	0	1	1	1	0	2
282	> 1	0	1	1	0	1	2
283	> 1	0	1	0	1	1	2
284	> 1	1	1	1	1	0	2
285	> 1	1	1	1	0	1	2
286	> 1	1	1	0	1	1	2
287	> 1	0	1	1	1	1	2
288	> 1	1	1	1	1	1	2

Table 2 Type 2 chemical classes

Class	nro-ohs	nrn -nnh	ohs	nnh
1	0	0	0	0
2	> 0	0	0	0
3	0	> 0	0	0
4	> 0	> 0	0	0
5	0	0	> 0	0
6	> 0	0	> 0	0
7	0	> 0	> 0	0
8	> 0	> 0	> 0	0
9	0	0	0	> 0
10	> 0	0	0	> 0
11	0	> 0	0	> 0
12	> 0	> 0	0	> 0
13	0	0	> 0	> 0
14	> 0	0	> 0	> 0
15	0	> 0	> 0	> 0
16	> 0	> 0	> 0	> 0