QSPR MODELS ON FRAGMENT DESCRIPTORS

Vitaly Solov'ev and Alexandre Varnek

The tutorials illustrate QSPR modeling by the ISIDA_QSPR program [1-4], realizing Multiple Linear Regression (MLR) analysis on the base of ISIDA Substructure Molecular Fragment (SMF) descriptors. ISIDA SMF descriptors are counts of the occurrence of subgraphs (fragments) in a molecule, where each descriptor element is associated to one of the detected possible fragments, complying with the user-proposed fragmentation scheme (fragment type, size, etc). The program builds MLR models combining forward [4] and backward [3] stepwise variable selection techniques.

The ISIDA_QSPR program is a graphical interface piloting this workflow and supporting graphical analysis of the results linked to the compound structures. It runs under the Windows operating system (ISIDA_QSPR.exe). It is strongly recommended to use of a non-system disk for the ISIDA_QSPR directory.

The following exercises are considered in the tutorial:

1. *Individual MLR model* - multiple linear regression on a single SMF descriptor set with descriptor selection, property predictions on a test set.
2. *Fragment analysis of the individual MLR model*: fragment contributions in modeling property, a pairwise correlation matrix for fragment contributions and the similarity of molecules according to SMF.
3. *External n-fold cross-validation*.
4. *Consensus modeling* based on the ensemble of Multiple Linear Regression models involving various types of SMF descriptors.
5. *Property predictions and virtual screening*

The tutorials includes step by step instructions indicated by the vertical line on the left side

Abbreviations

AD	Applicability domain
CM	Consensus model
EdChemS	The sketcher of the MOL files
EdiSDF	SDF manager of the ISIDA_QSPR program
F	The Fischer's criterion
FIT	The Kubinyi fitness criterion
FVS	Forward variable selection
HIV	The human immunodeficiency virus
HRF	The Hamilton R-factor percentage
ISIDA	In SIlico design and data analysis
LMO	Leave-many-out
LOO	Leave-one-out
MAE	Mean absolute error
MLR	Multiple linear regression
n	The number of data points
n-fold CV	n-Fold cross-validation
Q	Leave-one-out cross–validation correlation coefficient
QSPR	Quantitative structure–property relationships
R	The Pearson's correlation coefficient
R_{det}^2	Squared coefficient of determination
RMSE	Root-mean squared error
s	Standard deviation
SDF	Structure data file
SMF	Substructural molecular fragments
SVD	Singular Value Decomposition
TC	The Tanimoto similarity coefficients
TIBO	Tetrahydroimidazobenzodiazepinone derivatives
Y_{calc}	The fitted property
Y_{exp}	The modelling property

DATA

The following Structure-Data Files (SDF) are used in this tutorial: **AHIV-TIBO.SDF** and **TEST_AHIV-TIBO.SDF.** The first file contains experimental values of anti-HIV activity log ($1/IC_{50}$) of 57 tetrahydroimidazobenzodiazepinone (TIBO) derivatives [5], where IC_{50} is the concentration (mol/L) of the TIBO compound inhibiting 50% of the HIV-1 reverse transcriptase activity. The second file **TEST_AHIV-TIBO.SDF** contains five TIBO derivatives for which experimental anti-HIV activities are available and seven virtual TIBO derivatives generated by the CombiLIB / EdChemS tool [6, 7]. Experimental values of anti-HIV activity are represented by the log_1_C_exp. The input files must be located in the ISIDA_QSPR program directory

.

EXERCISE 1. Individual MLR model: modeling setup and output analysis

*Goal:*Building an individual MLR model based on SMF descriptors. The user needs only input files in SDF forma to perform the modeling on the training set and predictions on the test set.

Click the ***Single Model*** button of the ISIDA_QSPR program (Figure 1) to open the ***Single Model Calculations*** dialog box (Figure 2). The dialog box includes the ***Data*** panel for data input setup, the ***Descriptors*** panel for selection of the SMF descriptor type, the ***Model*** panel for modeling setup, and the ***Validation*** internal and external model validation setup (Figure 2).

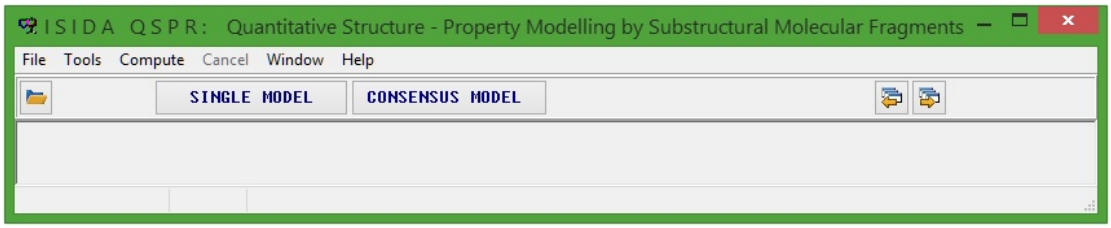


Figure 1. The ISIDA_QSPR Desktop.

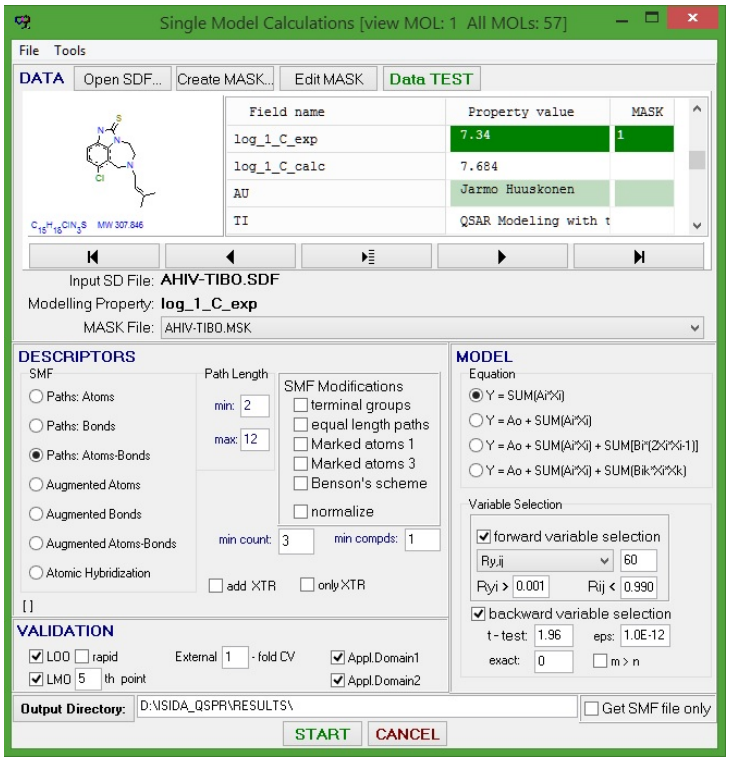


Figure 2. The single model calculations Dialog box.

1.1. ISIDA_QSPR input

Input data for ISIDA_QSPR should be prepared in Structure-Data File format [8], where the modeled property is represented by a data field. The property data field should be specified for all records in SDF, although the values of the property for the test compounds may be absent. Molecular structures may be represented as 2D- or 3D-structures. As a rule, hydrogen atoms of the structures are not specified, although molecules with explicit hydrogen atoms are supported [9, 10]. The input file must be located in the ISIDA_QSPR program directory.

From the ***Data*** panel of the single model building Dialog box (Figure 2), click on the ***Open SDF*** button and proceed to opening AHIV-TIBO.SDF file. The ***Input SD File: AHIV-TIBO.SDF*** label appears in the ***Data*** panel (Figure 2). The table in the ***Data*** panel includes the information stored in ***Field name*** and ***Property value*** fields (Figure 2) of the AHIV-TIBO.SDF file. Click on the ***log_1_C_exp*** cell of the ***Field name*** column to select the modelling property (Y_{exp}). The ***Modelling Property: log_1_C_exp*** label appears in the ***Data*** panel (Figure 2).

The SD File can be edited using EdiSDF tool. Click **Tools → SDF Editor** of the ISIDA_QSPR main menu (Figure 1) to open the EdiSDF manager (Figure 3). This versatile tool allows one to add new entries to the SD File, to add new data fields or edit existing ones, or to edit the structures (Figure 3). The current SDFs do not need any fixing, but we encourage the reader to try to use this self-explanatory tool to complete or edit corrupted SD files.

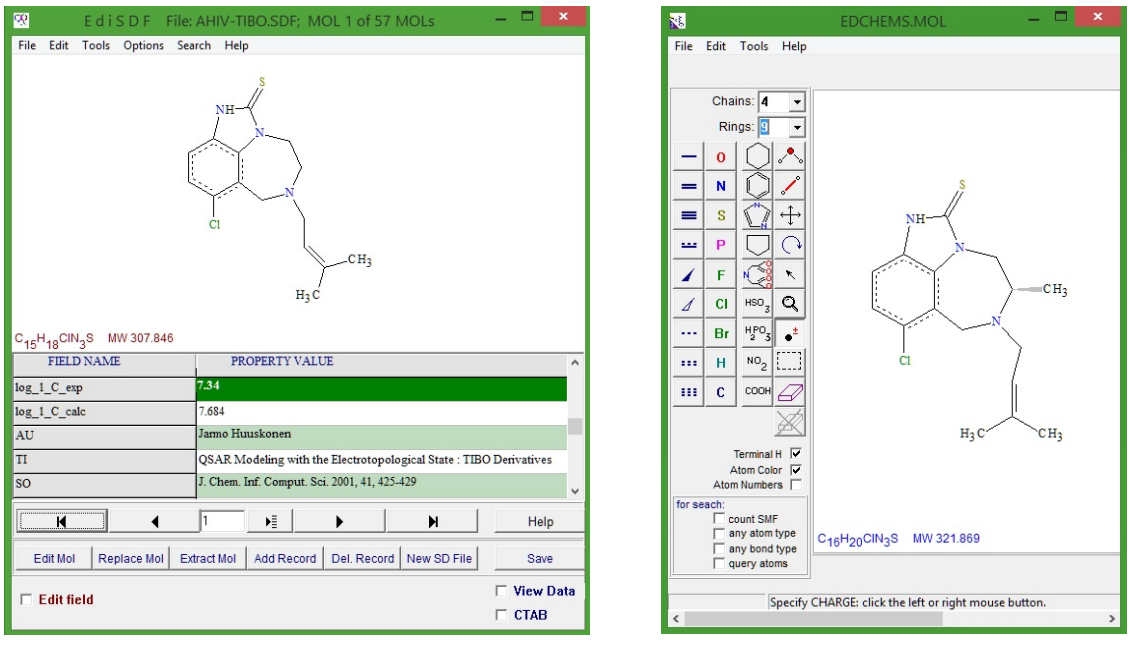


Figure 3. The EdiSDF (left) and EdChemS (right) graphical interface.

1.2. Data split onto training and test sets

ISIDA_QSPR can split the initial data set into two subsets: training and test sets for modelbuing and validation, respectively. The program uses a MASK file (*.msk) to indicate the train / test status for each compound. By default, ISIDA_QSPR produces mask files in which every N-th compound is kept out for testing, systematically starting from the M-th compound in the list (M≤N).

From the ***Data*** panel of the single model calculations Dialog box (Figure 2), click on the ***Create MASK*** button. The Create Mask Dialog box appears (Figure 3). Click on the *Create MASK* ***with TEST SET*** radio button; enter 5 in the ***each*** edit box and 5 in the ***starting from*** edit box. In the case shown on Figure 3, each fifth compound will be used for the test set. Click on the ***START*** button to save or overwrite the mask file AHIV-TIBO.MSK in the ISIDA_QSPR directory. The ***MASK file: AHIV-TIBO.MSK*** label appears in the ***Data*** panel (Figure 2). Click on the ***Data TEST*** button to verify the input data. The ***Information*** dialog box appears with message: "Input data files are in internal agreement". Click on the ***OK*** button to close the ***Information*** dialog box.

Figure 3. The Create Mask Dialog box.

1.3. Substructure Molecular Fragment (SMF) descriptors

The ISIDA_QSPR program includes a module for descriptor generation. ISIDA SMF, or simply SMF, descriptors [2, 5, 11-16] are counts of subgraphs (fragments) in a molecular graph. Each descriptor is associated with one of the fragments generated within the user-defined fragmentation scheme (fragment type, size, *etc*). The program can handle two main types of fragments (Figure 4): topological paths (I) and atom-centered fragments (atoms with nearest connected neighbors) (II). Either of these schemes supports indication of the atom and bond types (AB), the atom types only (A), the bond types only (B). The atom type can have different

attributes: atom symbol only, hybridization state, Benson's notation [17] and special mark [10, 15] (Figure 5).

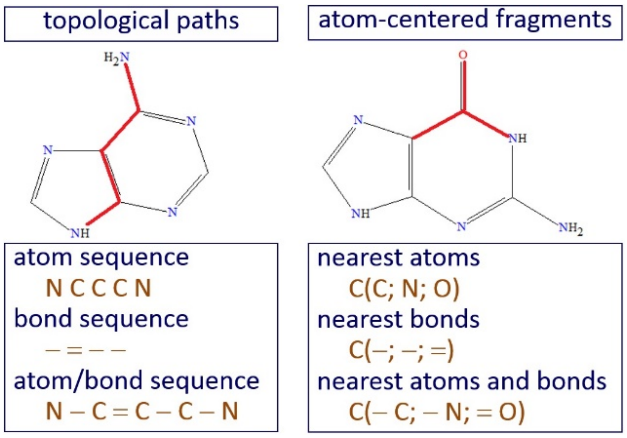


Figure 4. Two main classes of ISIDA SMF fragments: topological paths (I) and atom-centered fragments (II).

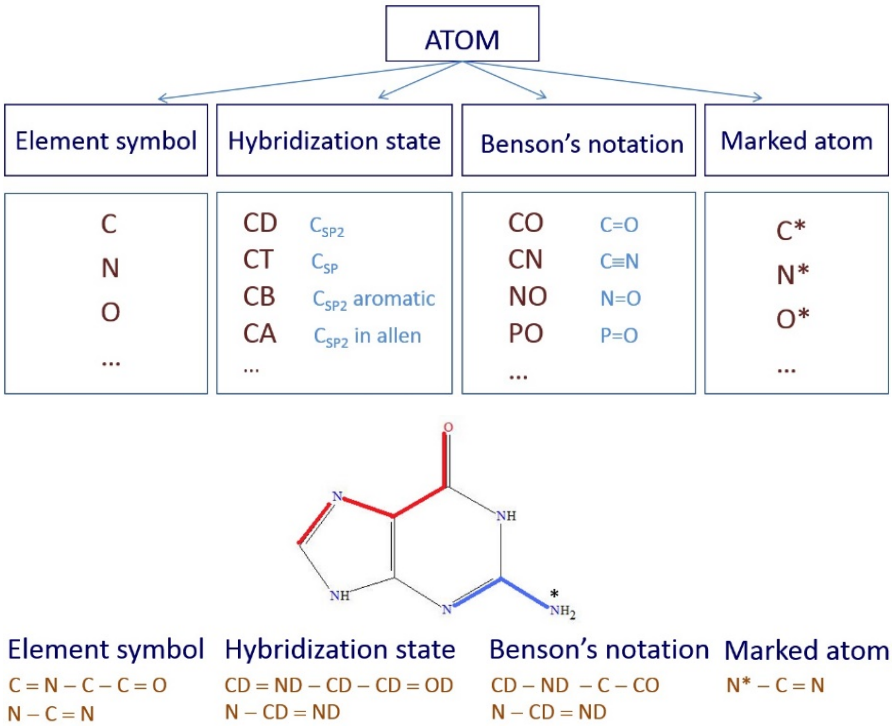


Figure 5. (*top*) Atomic attributes of substructural molecular fragments. (*bottom)* Example demonstrating different atomic labels used for the atom/bond paths containing 5 (*in red*) or 3 (*in blue*) atoms.

The bond attributes support special typing (covalent for σ-bonds, coordinating for noncovalent bonds and dynamic for reactions), order (single, double, triple, aromatic) and topology (cyclic or acyclic bond) (Figure 6). For topological paths, their optimality (shortest or

all paths), length (minimal and maximal, by defaults ranged between 2 and 15 atoms) and explicitness (all atoms are indicated or terminal atoms are indicated only) can be toggled.

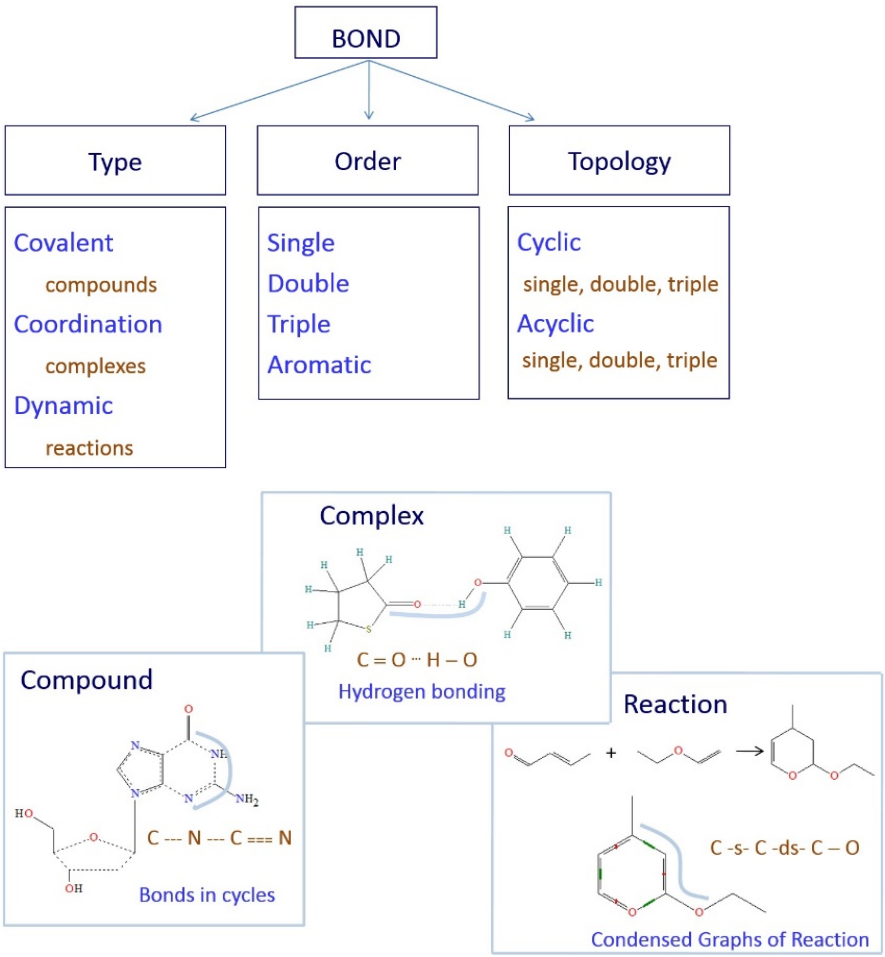


Figure 6. The bond attributes for ISIDA SMF descriptors.

From the ***Descriptors*** panel of the single model calculations Dialog box (Figure 2), click the ***Paths: Atoms-Bonds*** radio button, enter 2 in the ***min*** edit box and 12 in the ***max*** edit box for minimal and maximal path lengths, respectively. Please, verify that the *SMF Modifications* group boxes of the ***Descriptors*** panel are not checked. Use by default 3 in the ***min count*** edit box and 1 in the ***min compds*** edit box (Figure 2). For this exercise, choose shortest topological paths with explicit consideration of atom and bond types. Select the minimal (m_{min} = 2) and maximal (m_{min} = 12) numbers of atoms in the paths. Notice that the program will also generate all intermediate paths with m atoms: $m_{min} \leq m \leq m_{max}$. Please, verify that the ***Get SMF file only*** check box is not selected in the right bottom corner of the dialog box (Figure 2). This option is used only for the purpose to generate the SMF descriptors file and to store it in the working directory.

1.4. Regression equations

Multiple linear regression analysis is applied to build relationships between the variables x_i (SMF descriptors) and a dependent variable y (modeled property). Four types of equations are considered:

$$y = \sum_i a_i x_i + \Gamma \quad (1)$$

$$y = a_o + \sum_i a_i x_i + \Gamma \quad (2)$$

$$y = a_o + \sum_i a_i x_i + \sum_i b_i (2x_i^2 - 1) + \Gamma \quad (3)$$

$$y = a_o + \sum_i a_i x_i + \sum_{i,k} b_{ik} x_i x_k + \Gamma \quad (4)$$

Here, a_i and b_i (b_{ik}) are fragment contributions, x_i is the count of the i-th type fragment. The free term a_o is fragment independent. An extra term $\Gamma = \sum c_m D_m$ can be used to describe any specific feature of the compound using external descriptors D_m; by default $\Gamma = 0$. The parameters a_i and b_i are determined using the singular value decomposition (SVD) method [18].

From the ***Model*** panel (Figure 2), click on the ***Y = SUM(Ai*Xi)*** radio button to select linear fitting equation without the free term.

1.5. F*orward and backward stepwise variable selection*

Combined forward and backward stepwise techniques have been used to select the most pertinent variables from initial pool of the generated SMF descriptors [3, 4]. Initially, the forward stepwise variable selection (FVS) algorithm is applied to pre-select the user-defined number $m_p < n$ of the most relevant variables, where n is the size of training set. The FVS employs the known equations for the correlation coefficients between the response variable y and one- two- and three variables [19] in combination with the FSMLR algorithm [20]. Accordingly, three sub-algorithms (FVS-1, FVS-2 and FVS-3) have been used. At step p, the FVS procedure defines a new response variable $y^{(p)} = y^{(p-1)} - y_{calc}$, where $y_{calc} = c_0 + c_i x_i$ (FVS-1), $y_{calc} = c_0 + c_i x_i + c_j x_j$ (FVS-2) or $y_{calc} = c_0 + c_i x_i + c_j x_j + c_k x_k$ (FVS-3), p = 1, 2, 3,… and $y^{(0)} = y_{exp}$. Thus at every

step, one (x_i), two (x_i, x_j) or three variables (x_i, x_j and x_k) are selected ensuring maximal correlation coefficients ($R_{y,i}$, $R_{y,ij}$ or $R_{y,ijk}$ correspondingly) between the variable(s) and $y^{(p)}$. The steps are repeated until the number of selected variables m_p reaches a user-defined value. Optionally, variables x_m with small correlation coefficient with $y^{(p)}$ ($|R_{y,m}| < R^0_{y,m}$), those highly correlated with other variables x_i ($|R_{i,m}| > R^0_{i,m}$) or ''rare'' fragments (i.e., found in less than q molecules, here $q < 3$) can be eliminated. Here $R^0_{y,m}$ and $R^0_{i,m}$ are the user-defined thresholds. Then backward stepwise variable selection algorithm [3] eliminates the variables with low $t_i = a_i/\Delta a_i$ values for the models (1) and (2), where Δa_i is a standard deviation for the coefficient a_i at the i-th variable in the model. First, the program selects the variable with the minimal $t_{min} < t_0$, then it builds a new model excluding this variable. This procedure is repeated until $t \geq t_0$ for all selected variables. Here t_0 is the tabulated value of Student's criterion. By default, t_0 equals 1.96.

From the ***Model*** panel (Figure 2), check the ***forward variable selection*** and ***backward variable selection*** check boxes in the ***Variable Selection*** panel. Select the $\boldsymbol{R_{y,ij}}$ item from the dropdown list of the FVS algorithm combo box. On the right of the combo box, enter 60 in the edit box for the number of pre-selected variables as the percentage of the training set size. Enter 0.001 in the $\boldsymbol{R_{yi}}$ edit box and 0.99 in the $\boldsymbol{R_{ij}}$ edit box for the correlation coefficient thresholds. Enter 1.96 in the ***t-test*** edit box, 1E-12 in the ***eps*** edit box and 0 in the ***exact*** edit box. Make sure that the ***m>n*** check box is not selected (Figure 2).

1.6. Parameters of internal model validation

One can distinguish internal and external validation. The former corresponds to the procedure - leave-one-out (LOO) or leave-many-out (LMO) cross-validation - performed after completing variables selection on the entire set. External validation in n-fold cross-validation or on a selected set is always performed on the data never used at any step of the modelbuilding.

From the ***Validation*** panel (Figure 2), check the ***LOO*** check box to calculate the leave-one-out (LOO) cross-validation correlation coefficient (Q^2). Make sure that the ***rapid*** check box is not selected. Check the ***LMO*** check box for the calculation of the leave-many-out (LMO) cross-validation correlation coefficient. Enter 5 in the ***i-th point*** edit box for the LMO calculations: each fifth data point is discarded followed by the modelbuilding on the remaining training data and to use discarded objects for the model validation. Enter 1 in the ***External n-fold CV*** edit box to perform the modelling without external n-fold cross-validation (n-fold CV).

1.7. Applicability Domain (AD) of the model

The applicability domain (AD) of the model defines an area of chemical space where the model is presumably accurate. Three types of AD definitions can be used either simultaneously or individually: fragment control, bounding box [21] and "quorum control" [22]. Bounding box approach considers AD as a multidimensional descriptor space confined by minimal and maximal occurrences of the descriptors involved in an individual model (AD1). Fragment control consists in discarding predictions for the compounds containing descriptors not occurring in the initial SMF pool generated for the training set (AD2). "Quorum control" is a threshold for the number of models accepted by AD1 and AD2. If this number is lower than a user defined threshold, the consensus prediction is ignored.

From the ***Validation*** panel (Figure 2), check the ***Appl. Domain1*** and ***Appl. Domain2*** check boxes for the fragment control and bounding box of model applicability domain, respectively.

1.8. Storage and retrieval modelling results

The output files are saved in a user-selected directory by clicking ***Output Directory*** button (Figure 2). The Open dialog box appears, where click ***Open*** button to select the directory, for instance, C:\ISIDA_QSPR\RESULTS. The output files can always be opened by clicking File → Open in the ISIDA_QSPR main menu (Figure 1). Typically, the *.out file includes the name of the input SD file name as substring and begins with the date and the time of the performed calculations.

1.9. Analysis of modelling results

To build the model, click ***Start*** button in ***Single Model Calculations*** dialog box (Figure 2). The program creates 9 output files: 4 plain text and 5 files with the graphical representation of results, see their description below.

The *.TXT file (here, *<date_time>_AHIV-TIBO.TXT*) contains the following information concerning the QSPR model:

a) initial list of the SMF descriptors,

b) setup parameters,

c) groups of concatenated fragments always occurring in the same combination in each compound of the training set,

d) statistical parameters of the multiple linear regression (MLR) including Pearson's correlation coefficient *R*, Fischer's criterion *F*, root mean squared error *RMSE*, mean absolute error *MAE*, the leave-one-out cross–validation correlation coefficient Q^2,

e) SMF descriptors involved in the MLR equation, regression equation coefficients (SMF contributions) a_i and their random errors Δa_i for the 95% confidence interval,

f) a pairwise correlation matrix for SMF contributions,

g) singular values s_i obtained in SVD calculations(see section 1.4),

h) Table of experimental (Y_{exp}) and fitted (Y_{calc}) property, residuals Y_{exp} - Y_{calc} for the training set.

The *SMF file (<date_time>_AHIV-TIBO.SMF) contains full set of generated SMF descriptors and their counts in the training set molecules:

```
    Full Set of Fragments.

     1.                                                      C*C
     2.                                                      C-N
     3.                                                      C-C
     4.                                                      C=S
     5.                                                     C-Cl
    ...

MATRIX: Compound (Line) × Fragment Count (Column).
          1    2    3    4    5    6    7    8    9   10   11   12 ...
     1    6    8    5    1    1    1    7    4    6    1    2    4 ...
     2    6    8    5    1    1    1    7    4    6    1    2    4 ...
     3    6    9    7    0    0    1    7    4    6    1    4    6 ...
     4    6    8    6    0    0    1    7    4    6    1    2    6 ...
     6    6    8    5    1    1    1    7    4    6    1    2    5 ...
     7    6    8    7    1    1    0    7    4    6    1    2    5 ...
     8    6    8    7    0    0    0    7    4    6    1    2    5 ...
     9    6    8    7    0    0    1    7    4    6    1    2    5 ...
    11    6    8    7    0    0    1    7    4    6    1    2    5 ...
    12    6    8    6    0    0    1    7    4    6    1    2    5 ...
    ...
```

The *.MF file (<date_time>_AHIV-TIBO.MF) includes a list of SMF selected for the model and related descriptor's values for the training set molecules

```
    Set of Fragments for the Model.

    12.                                                       C-C-N
    32.                                                   C*C*C-N-C
    36.                                                   C-C-N-C-N
    54.                                               Cl-C*C*C*C-N
    83.                                           C*C*C-C-N-C-C=C-C
    ...

MATRIX: Compound (Line) × Fragment Count (Column).
         12   32   36   54   83   88   99  144  167
     1    4    4    1    1    4    1    0    0    0
     2    4    4    1    1    4    1    0    0    0
     3    6    4    1    0    4    1    2    2    0
     4    6    4    1    0    0    1    1    0    0
     6    5    4    2    1    0    1    0    0    0
     7    5    4    2    1    0    0    0    0    0
     8    5    4    2    0    0    0    1    0    2
     9    5    4    2    0    0    1    1    0    2
    11    5    4    2    0    0    1    1    0    1
    12    5    4    1    0    4    1    1    0    0
    ...
```

The *.DOC file (<date_time>_AHIV-TIBO_Pred.DOC) contains the Table of predicted property (right column) for the compounds of the test set defined by the MASK file. The *Datum* column contains experimental data. If a compound is identified as being outside the AD of the model, the predicted value for this compound is excluded:

TABLE P1. Test set: Predicted property log_1_C_exp for the compounds from the AHIV-TIBO.SDF file.

cmp. no.	Datum	IAB2-120
5	4.49	2.94
10	4.48	
15	5.61	5.20
20	5.65	5.91
25	5.18	5.72
...		

The *.RAC plot file (here, <date_time>_AHIV-TIBO.RAC) provides with the analysis of residuals (Y_{exp} - Y_{calc}) as a function of fitted property (Y_{calc}) for the training set (Figure 7). The red dotted line corresponds to zero deviation. To visualize the residual value and corresponding molecular structure, move the mouse pointer on the data point (small circle) and click. The structure appears on the yellow background; the internal number of the data point, the Y_{exp}**,** Y_{calc} and (Y_{exp} - Y_{calc}) values emerge in the status bar at the bottom of the program window (Figure 7).

The *PLT plot file (here, <date_time>_AHIV-TIBO.PLT) displays the correlation between Y_{exp} and Y_{calc} as well as corresponding linear equation, including the number of data points (n), correlation coefficient (R), Fischer's criterion (F), standard deviation (s), squared coefficient of determination (R_{det}^2), root-mean squared error (*RMSE*) and mean absolute error (*MAE*) (Figure 8a).

Remaining two plots (files <date_time>_AHIV-TIBO.LOO and <date_time>_AHIV-TIBO.LMO) represent the relationships between Y_{exp} and Y_{pred} as well as corresponding linear equations with their statistical parameters for the leave-one-out (Figure 8b) and leave-many-out cross-validations.

The fifth plot displays the results of linear regression analysis, including the plot Y_{pred} versus Y_{exp} for the test set defined by the MASK file. Objects identified as being outside AD of the model are excluded (Figure 8c).

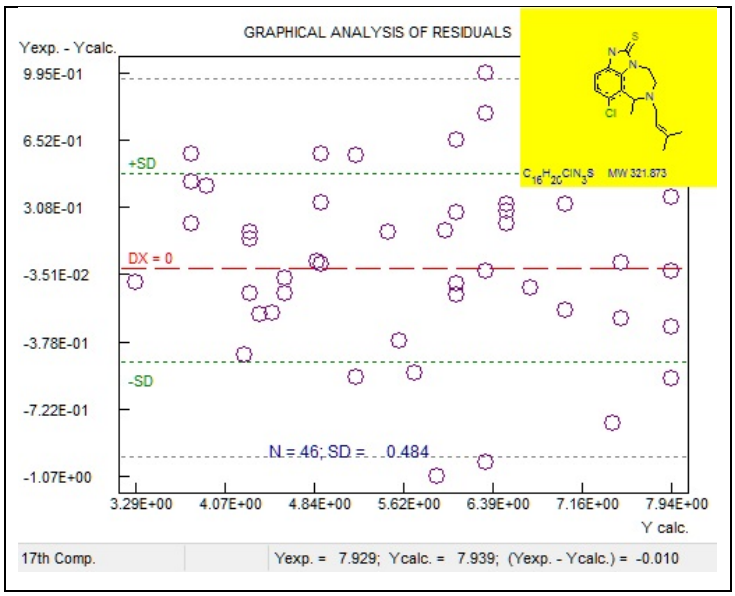


Figure 7. The graphical window of residuals' analysis.

Molecular structure corresponding to selected data point on the graphs can be visualized by mouse clicking. The structure appears on the yellow background; the internal number of the data point, the Y_{exp} and Y_{calc} (Y_{pred}) values emerge in the status bar at the bottom of the program window.

1.10. Root-Mean Squared Error (RMSE) estimation

Root-mean squared error $RMSE = [1/n \sum_{i=1}^{n}(Y_{exp,i} - Y_i)^2]^{1/2}$ characterizes the ability of the model to reproduce quantitatively the experimental data, where Y_i is the fitted $Y_{calc,i}$ or predicted $Y_{pred,i}$ value of the property for the *i*-th data point. Typically, *RMSE* values increase in the order *RMSE* (fitting) $<$ *RMSE* (LOO) $<$ *RMSE* (external test set), as demonstrated in Figure 8. This can be explained by the fact, that information about the training set compounds is used at the fitting and partially (at the variables selection step) at LOO or LMO stages whereas the test compounds are never seen at any step of the modelbuilding.

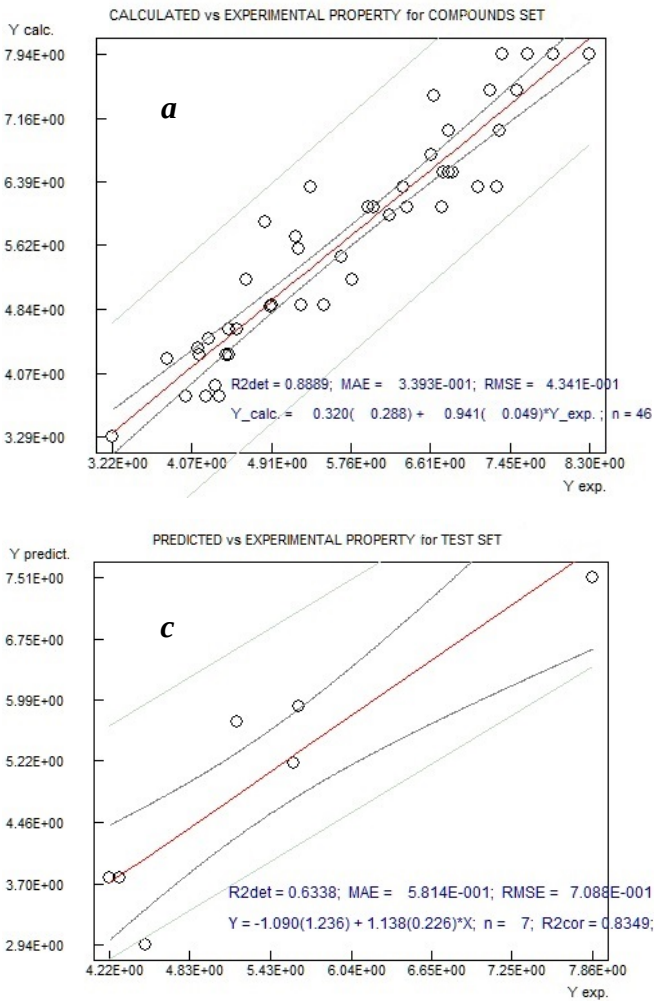


Figure 8. Modeling of anti-HIV activity (Y = $\log(1/IC_{50})$) of tetrahydroimidazobenzo-diazepinone (TIBO) derivatives. The three plots show linear correlation for calculated versus experimental activities obtained for the training set at the fitting (***a***) and LOO (***b***) stages as well as for external test set (***c***).

EXERCISE 2: Analysis of the fragment contributions for individual MLR model

One may expect that the presence of some particular structural motifs increases or decreases the compound potency. In this exercise, we demonstrate how fragment contributions can be analyzed with the help of the MolFrag module which opens the ***Statistics of Substructural Molecular Fragments*** window (Figure 9). This tool provides the user with:

a) the list of fragment descriptors and their contribution, minimal and maximal occurrence in the training set compounds (***Model Parameters*** tab),

b) an assessment of pairwise molecular similarity based on fragment descriptors involved in the model (***Similarity*** tab),

c) a pairwise correlation of these descriptors (***Corerlations*** tab),

d) the list of fragment descriptors involved in the model and their contributions for each molecule in the training set (***SMF table*** tab).

Some details of these functionalities are given below

The ***Model Parameters*** tag displays two tables (Figure 9). The upper one shows the list of SMF descriptors (molecular fragments) generated for the training set. For each descriptor it reports: identification number (*id*), name (the denomination of the associated fragment), contribution (*contrib.*) and its standard deviation (*SD*), the minimal (*min*) and maximal (*max*) fragment counts over the training set, the number of the compounds containing the given fragment (*mols*). The lower table contains the groups of fragments always occurring in the same combination in certain compounds of the training set. The “main” (longest or lexicographically high-order path) fragment in the group is indicated by the "+" sign in the ***main*** column. The navigation buttons at the bottom are used to browse the fragment groups (Figure 9).

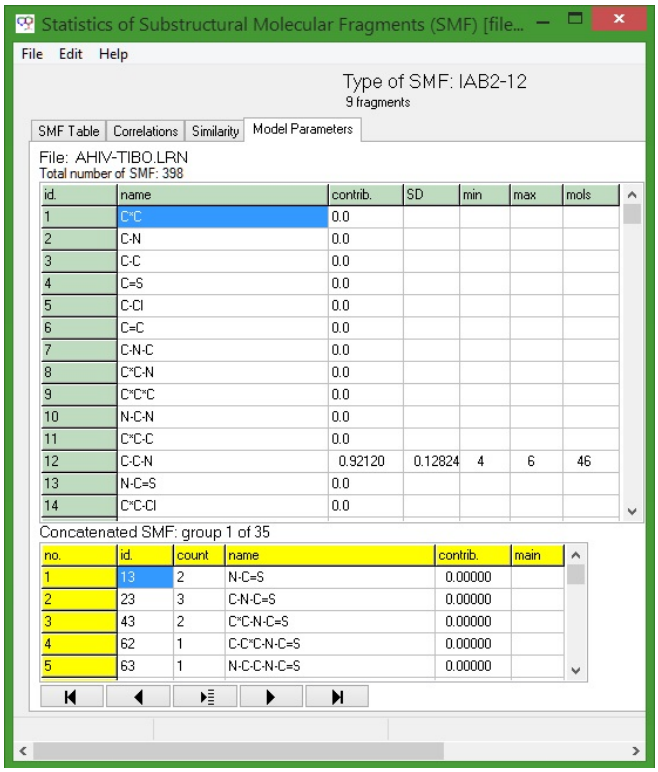


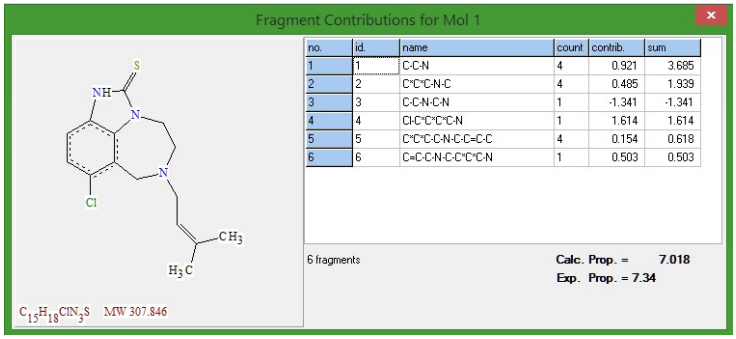
Figure 19. MolFrag graphical interface

The ***Similarity*** tab reports paitwise Tanimoto coefficients (*TC*) calculated with a help of fragment descriptors involved in the model. The user can enter a *TC* threshold value in the ***TC*** edit box then click on the ***Mark*** button to highlight the Tanimoto coefficients exceeding this threshold.

The ***Correlations*** tab reports a pairwise correlation of the descriptors included in the model. The user can enter a threshold value of the squared correlation coefficient R^2 in the ***R*** edit box then click on the ***Mark*** button to highlight in the ***Correlated fragments*** window all R^2

exceeding this threshold. Clicking on any highlighted cell opens the ***Y versus X*** window which reports corresponding linear equation and related statistical parameters squared correlation coefficient (*R2cor*), Fischer's criterion (*F*) and standard deviation (*s*).

The ***SMF Table*** tab summarizes occurrences of molecular fragments involved in the model. Click on a cell corresponding to a particular molecule (e.g., ***mol 1*** cell) opens the ***Fragment Contributions*** window describing its 2D structure, constituting fragments and their contributions into the modeled property (Figure 11).



no.	id.	name	count	contrib.	sum
1	1	C-C-N	4	0.921	3.685
2	2	C*C*C-N-C	4	0.485	1.939
3	3	C-C-N-C-N	1	-1.341	-1.341
4	4	Cl-C*C*C*C-N	1	1.614	1.614
5	5	C*C*C-C-N-C-C=C-C	4	0.154	0.618
6	6	C=C-C-N-C-C*C*C-N	1	0.503	0.503

Figure 11. Selected training set compound, its constituting fragments and their contributions into the modeled property

EXERCISE 3: External n-fold cross-validation

The external *n*-fold cross-validation procedure is often used [16, 23, 24] as a standard protocol for the estimation of the predictive performance of the model. According to this procedure, an entire dataset is split into *n* non-overlapping pairs of training and test sets. On each fold, a training set covers $(n - 1)/n$ of the data points while related test set covers the remaining $1/n$ of the data points. The model developed on the training set is applied to the corresponding test set. Finally, predictions for all test sets are concatenated and, in such a way, all data points in the entire data set are predicted. Note that the bigger *n*, the larger the training set, meaning that the information available at model training stage – and hence, implicitly, the chance to encounter, at training stage, compounds that are similar to test molecules – is increased. Thus, the bigger *n*, the more "optimistic" cross-validation results become. The most aggressive cross-validation, at $n = 2$, challenges a model trained on half of the original set to predict the other half, and therefore may be too pessimistic – unless very large data sets are used. At the other extreme of the spectrum, LOO cross-validation (which is nothing else but N-fold cross-validation, with N =

number of compounds in the entire set) is definitely too optimistic, but generates equations that are closest to the one that could be obtained on hand of the entire set.

3.1. Setting the parameters

Specify the parameters for the modelling as stated above in sections 1.1 – 1.8 of the Exercise 1. Then click on the ***Create MASK*** button of the single model calculations Dialog box (Figure 2). The Create Mask Dialog box appears (Figure 6). Click on *Create MASK**, No test set*** button. Click on the ***START*** button to save or overwrite the mask file **AHIV-TIBO.MSK** using the Save mask file Dialog. Click on the ***Data TEST*** button to verify if the input data are suitable for the modelbuilding (Figure 2). If this is a case, the ***Information*** dialog box displays a message: "Input data files are in internal agreement". Close the ***Information*** dialog box.

From the ***Validation*** panel (Figure 2), enter **5** in the ***External n-fold CV*** edit box to perform the modelling with external 5-fold cross-validation (5-fold CV).

3.2. Analysis of n-fold cross-validation results

Click on the ***Start*** button of the ***Single Model Calculations*** dialog box (Figure 2) to perform the calculations. The program creates 6 output files: 4 plain text files (*.TOM, *.DOC, *.AVE and *.ECV) and 2 files of the graphical presentation of results.

The ***.TOM** file contains statistical parameters of the individual MLR model built on every fold of 5-fold CV calculations:

```
5-Fold External Cross-Validation Procedure.

...
File of Mol Structures: AHIV-TIBO.SDF;  57 compounds in training set.
Modeling Property Name: log_1_C_exp
Mask File:              AHIV-TIBO.MSK
Exter.Descriptors File: -
...
 no fragment  fitting     n    k   R2          F        FIT     s           HRF    Q2
    type      equation
   1 IAB2-12      0       45   13  0.943843    44.82    2.846   3.69E-01    5.410  0.901341
   2 IAB2-12      0       45   15  0.964574    58.35    3.389   2.81E-01    4.109  0.905603
   3 IAB2-12      0       46   10  0.913880    42.45    3.008   4.30E-01    6.700  0.862763
   4 IAB2-12      0       46   20  0.982215    75.57    3.528   2.18E-01    2.847  0.937505
   5 IAB2-12      0       46    9  0.888879    37.00    2.691   4.84E-01    7.448  0.789069
```

For each fold, the following statistical parameters of related individual model are given:

- the number of the data point (*n*) in the training set of 5-fold CV,
- the number of fitted parameters (fragment contributions) (*k*),
- squared Pearson correlation coefficient (R^2),
- the Fischer criterion (*F*),

- the Kubinyi fitness criterion [19] (*FIT*),
- standard deviation (*s*),
- the Hamilton R-factor percentage [25] (*HRF*) and
- squared LOO cross–validation correlation coefficient (Q^2).

The ***.DOC** file reports the fitted property values for every fold of 5-fold CV:

```
...
TABLE L1. Training set: Calculated property log_1_C_exp for the compounds from the AHIV-TIBO.SDF
file.
                              1          2          3          4          5
cmp. no.       Exp.     IAB2-120   IAB2-120   IAB2-120   IAB2-120   IAB2-120
       1        7.34                   7.39       6.98       7.33       7.02
       2        6.80        7.06                  6.98       7.03       7.02
       3        5.20        5.49       5.26                  5.27       5.57
       4        4.64        4.72       4.51       4.37                  5.20
       5        4.49        4.33       4.51       4.37       4.39
```

The ***.AVE** file contains the average fitted property and its standard deviation calculated from the data given in the described above ***.DOC** file

In the ***.ECV** file, predicted in 5-CV property values (right column) are compared with the experimental ones given in the *Datum* column. If a compound is identified as being outside AD of the model, the predicted value for this compound is excluded (e.g., the case of compound 11):

```
TABLE P1. Test set: Predicted property log_1_C_exp for the compounds from the AHIV-TIBO.SDF file.

cmp. no.       Datum    IAB2-120
       1        7.34        7.27
       6        6.17        5.68
      11        4.32
      16        7.11        6.75
      21        4.84        6.12
```

The plots display the relationship between observed Y_{exp} and predicted Y_{pred} (or fitted Y_{calc}) property as well as corresponding linear equation and its statistical parameters, including data points for all folds cross-validation. Clicking on selected data point visualizes corresponding molecular structure and the Y_{exp} and Y_{pred} (Y_{calc}) values.

EXERCISE 4. Consensus model: obtaining and validation

The ISIDA_QSPR program may generate many different linear models, each involving particular set of SMF descriptors and /or variable selection technique. The individual models are recruited into the consensus model according to two criteria: the LOO cross–validation correlation coefficient Q^2 should be larger than a user defined threshold Q^2_{lim} and a residual (R^2 – Q^2) between the squared correlation coefficient (R^2) and Q^2 should also be larger than (R^2 –

$Q^2)_{lim}$ threshold (Figure 12). The program then applies this consensus model (CM) to every test compound, i.e., predicts the target property as an arithmetic average of the values estimated by selected individual models. A given individual model doesn't contribute into consensus calculations if it produces the outliers according to Tompson's rule [26] or it can't be applied to a given compound due to applicability domain (AD) problem (Figure 12). Three types of AD criteria can be used simultaneously or individually: fragment control, bounding box [21] and "quorum control" [22].

In this tutorial, we use only a few descriptor types and one fitting equation type leading to generation of 144 individual models. It ensures short time of calculations and demonstrates ensemble learning and predictions by consensus model.

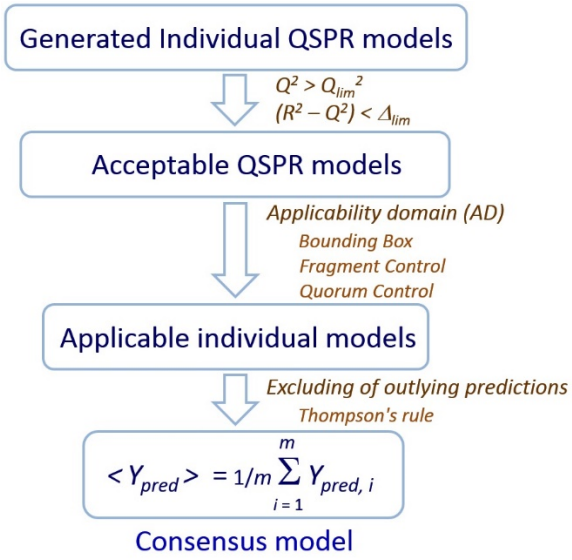


Figure 12. Consensus calculations based on ensemble of selected individual models.

4.1. Loading Structure-Data file

Restart ISIDA_QSPR, upload input SD file and select the modelling property as shown in Exercise 1. Click on the ***Create MASK*** button. The Create Mask Dialog box appears (Figure 6). Click on the Create MASK Dialog box, ***No test set*** radio button. Click on ***START*** to save or overwrite the mask file **AHIV-TIBO.MSK** in the ISIDA_QSPR directory using the Save mask file Dialog. The ***MASK file: AHIV-TIBO.MSK*** label appears in the ***Data*** panel. Click on ***Data TEST*** to verify the consistency of input data (Figure 2). The ***Information*** dialog box appears with message: "Input data files are in internal agreement". Close the ***Information*** dialog box and then the ***Single Model Calculations*** dialog box by clicking on the ***CANCEL*** button (Figure 2).

Figure 13. The consensus model calculations Dialog box.

4.2. Descriptors and fitting equation

Click on the ***Consensus Model*** button of the ISIDA_QSPR program (Figure 1) to open the ***Consensus Modelling Calculations*** Dialog box (Figure 13) used to enter parameters for consensus model. The dialog box includes the ***Data*** panel for data input, the ***Descriptors*** panel for selection of the SMF descriptor types, the ***Model*** panel for options of individual MLR equation types and forward and backward stepwise variable selection techniques, and the ***Validation*** panel for parameters of internal and external individual model validation (see details about validations in sections 1.6 and 3).

From the ***Descriptors*** panel (Figure 15), check the ***Paths: Atoms-Bonds*** check box and then the ***equal length paths*** check box, enter 2-3 in the ***min*** edit box and 7-12 in the ***max*** edit box for minimal and maximal path lengths. Please, verify that there are no additional check marks in the check boxes of the ***Descriptors*** panel. Use by default 3 in the ***min count*** edit box and 1 in the ***min compds*** edit box (Figure 13). This setup leads to generation of two descriptor classes: *a*) shortest topological paths with explicit atoms and bonds and *b*) similar shortest paths including paths of equal length. For every class of the sequences, the minimal ($2 \le n_{min} \le 3$) and maximal ($7 \le n_{max} \le 12$) numbers of constituent atoms (n) are defined. The sequences include all

intermediate shortest paths with *n* atoms: $n_{min} \leq n \leq n_{max}$, thus leading to generation of 24 types of fragment descriptors.

From the ***Model*** panel (Figure 15), click on the ***Y = SUM(Ai*Xi)*** check box for the selection of one linear fitting equation type only.

4.3. Variables selection

From the ***Model*** panel (Figure 13), check the ***FVS*** (forward variable selection) ***methods*** and ***backward variable selection*** check boxes in the ***Variable Selection*** subpanel. On the right of the FVS methods' combo box, enter 2,3 in the first edit box for the selection of the $\boldsymbol{R_{y,i}}$ and $\boldsymbol{R_{y,ij}}$ variable selection algorithms [4]. In the second edit box, enter 50-70 for the scalable numbers of pre-selected variables presented as the percentage of the training set size (*m*). In this case, 0.5*m*, 0.6*m* and 0.7*m* variables will be pre-selected by the $\boldsymbol{R_{y,i}}$ ($\boldsymbol{R_{y,ij}}$) algorithm and sequentially applied to individual model preparations. Enter 0.001 in the $\boldsymbol{R_{yi}}$ edit box and 0.99 in the $\boldsymbol{R_{ij}}$ edit box for the correlation coefficient thresholds. Enter 1.96 in the ***t-test*** edit box, 1E-12 in the ***eps*** edit box and 0 in the ***exact*** edit box. Make sure that the ***m>n*** check box is not selected (Figure 13).

4.4. Consensus model

From the ***Model*** panel (Figure 13), enter 0.7 in the ***Q2 lim*** edit box for the threshold Q^2_{lim} of minimal LOO cross–validation correlation coefficient (Q^2) of acceptable individual models. Enter 0.1 in the ***R2 – Q2*** edit box for the threshold of maximal residual between the squared correlation coefficient (R^2) and Q^2 of acceptable individual models.

4.5. Model applicability domain

From the ***Validation*** panel (Figure 13), user can select three AD approaches: fragment control (AD1), bounding box (AD2) and "quorum control" (AD3). Here, check the ***Appl. Domain 1*** check box and uncheck the ***Appl. Domain 2*** and ***Appl. Domain 3*** check boxes.

4.6 n-Fold external cross-validation

From the ***Validation*** panel (Figure 13), check the ***LOO*** and ***rapid*** check boxes for fast calculation of *Q,* enter 5 in the ***External n-fold CV*** edit box for the execution of the external 5-fold cross validation.

4.7. Saving and loading of the consensus modeling results

The program saves the output files in user-defined directory by clicking on the ***Select Directory*** button (Figure 13). The Open dialog box appears, where click on ***Open*** to select a directory, for instance, C:\ISIDA_QSPR\RESULTS.

The output files can always be opened by clicking File → Open in the ISIDA_QSPR main menu (Figure 1). Typically, the *.out file includes the name of the input SD file name as substring and begins with the date and the time of the performed calculations. Verify that the ***Save Models*** check box is not selected in the right corner of the dialog box (Figure 13).

4.8. Statistical parameters of the consensus model

In order to obtain an ensemble of individual models forming CM, click on ***START*** in the ***Consensus Modelling Calculations*** Dialog box (Figure 13). The program creates 8 output files: 6 plain text files and 2 files of the graphical presentation of results, see their description below.

The ***.TOM** file contains statistical parameters of the individual MLR models for every fold of CV:

```
5-Fold External Cross-Validation Procedure. Table of models.

=>  Subset 1/5

File of Mol Structures: AHIV-TIBO.SDF;  45 compounds in training set.
Modeling Property Name: log_1_C_exp
Mask File:              AHIV-TIBO.MSK


  no fragment  fitting    n    k   R2           F       FIT     s          HRF      Q2
     type      equation
   1 IAB3-837     0       45   13  0.951211     51.99    3.301   3.44E-01    5.043  0.904357
   2 IAB2-1036    0       45   13  0.943843     44.82    2.846   3.69E-01    5.410  0.901341
   3 IAB2-1136    0       45   13  0.943843     44.82    2.846   3.69E-01    5.410  0.901341
   4 IAB2-1236    0       45   13  0.943843     44.82    2.846   3.69E-01    5.410  0.901341
   5 IAB2-10a26   0       45   10  0.933738     54.80    3.914   3.83E-01    5.877  0.896698
...
```

For each individual model, the following parameters are reported: the number of the data point (*n*) in the training set at a given fold CV, the number of fitted variables (*k*), squared Pearson correlation coefficient (R^2), the Fischer criterion (*F*), the Kubinyi fitness criterion [19] (*FIT*), standard deviation (*s*), the Hamilton R-factor percentage [25] (*HRF*) and squared LOO cross–

validation correlation coefficient (Q^2). The models are sorted according to Q^2 in descending order.

The ***_TST_5fCV_.AVE** file reports for the compounds of the test set at the given fold the following parameters: average predicted property values (*Average*) and their standard deviation (*STDEV*) estimated by consensus models and the number of the individual models (*Nm*) used for the *Average* value calculation. If a compound is identified as being outside AD of any individual model, no predictions are reported (e.g., see compound 11):

```
TABLE PA. Test set: Average predicted property log_1_C_exp

 cmp. no.       Datum   Average         STDEV          Nm    Dat.- Ave.
       1        7.34  7.23103E+000    4.080E-001       97    1.090E-001
       6        6.17  6.09287E+000    3.144E-001       80    7.713E-002
      11        4.32                                    0
      16        7.11  6.42392E+000    3.050E-001      102    6.861E-001
...
```

The ***.TSP** file contains property values predicted by individual MLR models for every fold of cross-validation:

```
5-Fold External Cross-Validation =>  Subset 1/5
File of Mol Structures: AHIV-TIBO.SDF
Property Name:          log_1_C_exp
Mask File:              AHIV-TIBO.MSK

TABLE P1. Test set: Predicted property log_1_C_exp
          102 Selected MODELs, Q2 >= 0.7
 cmp. no.       Datum   IAB3-8370  IAB2-10360  IAB2-11360  ...
       1        7.34        6.67        7.27        7.27   ...
       6        6.17        5.80        5.68        5.68   ...
      11        4.32                                       ...
      16        7.11        6.34        6.75        6.75   ...
      21        4.84        5.50        6.12        6.12   ...
...
```

Remaining three text files, similarly to described above *_TST_5fCV_.AVE and *.TSP files, contain information about fitted property values for the compounds of training sets at each fold.

The plots display the relationship between observed Y_{exp} and predicted Y_{pred} (or fitted Y_{calc}) property as well as corresponding linear equation and its statistical parameters, including data points for all folds cross-validation. Clicking on selected data point visualizes corresponding molecular structure and the Y_{exp} and Y_{pred} (Y_{calc}) values.

4.9. Consensus model performance as a function of individual models acceptance threshold

The consensus model predictive performance depends on recruited individual models which selection, in turn, depends on user-defined threshold Q^2_{lim} for determination coefficient of LOO calculations (see introduction to Exercise 4). Treating the output of consensus modeling the user

can build a plot of the dependence of determination coefficient R^2_{det} in *n*-fold CV as a function of Q^2_{lim} which may help to determine an optimal Q^2_{lim} value providing with reasonable R^2_{det}.

Click ***Tools*** → ***R2det vs Q2*** of the ISIDA_QSPR main menu (Figure 1) to open the **Averaging** tool (Figure 14), which enables to explore *n*-fold CV determination coefficient R^2_{det} as a function of Q^2_{lim}. In this window check ***use Q2 threshold*** and ***do R2det vs Q2*** boxes, enter 0.7 in the ***use Q2 threshold*** edit box, enter the increment value 0.05 in the ***Step*** edit box. In order to load consensus model information, click ***File*** → ***Open*** of the **Averaging** window. In the Open dialog box select the <date_time_>**AHIV-TIBO_5fCV.TSP** file from the list of the *.TSP files in the directory, where the consensus modelling results are saved (e.g., C:\ISIDA_QSPR\RESULTS directory) and then click ***Open***. The Open dialog box appears again. Select the proper <date_time_> **AHIV-TIBO_5fCV.TOM** file from the list of the *.TOM files in the same directory, click the ***Open*** button. After the **Averaging** tool performed calculations, click the ***Graph*** tab (Figure 14). The plot displays the relationship between R^2_{det} and Q^2_{lim}. One can see that R^2_{det} insignificantly increases with Q^2_{lim}: R^2_{det} = 0.881 at Q^2_{lim} = 0.70 and R^2_{det} = 0.883 at Q^2_{lim} = 0.85. The complementary textual information related to this plot is available in the ***Table*** tab (Figure 14).

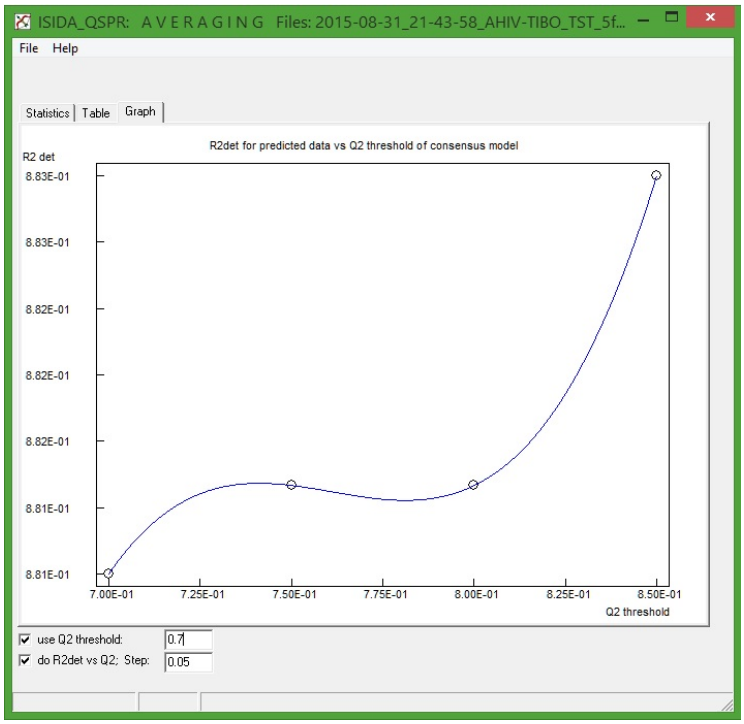


Figure 14. The Averaging tool graphic interface.

4.10. Building consensus model on the entire data set

This section explains how to build and save the CM on the entire data set. Click the ***Consensus Model*** button of the ISIDA_QSPR program (Figure 1) to open the ***Consensus Modelling Calculations*** Dialog box (Figure 13). Keep all previously used settings (Figure 13) except the number of folds and the name of directory for output files. Enter 1 in the ***External n-fold CV*** edit box to deactivate n-fold CV. Check the ***Save Models*** check box, click on the ***Select Directory*** button. In the Open dialog box select a directory (e.g., C:\ISIDA_QSPR\AHIV-TIBO_MODELS. Click ***START*** to prepare and save a set of individual MLR models.

The program creates and opens 5 output files: 4 plain text files and 1 file of the graphical presentation of results which are similar to the files for training subsets in *n*-fold CV described in Section 4.8. The consensus model is described in AHIV-TIBO.TSC and AHIV-TIBO.TOM files containing information about 117 constituting individual models (the *.SPE files).

EXERCISE 5. Property predictions and virtual screening using consensus models

This exercise demonstrates the Consensus Predictor program tool [27] which applies previously obtained consensus models to an external data set. As an input, Consensus Predictor uses chemical structures in SDF format [8]. The input can also include experimental or estimated property values in a data field named as a property aimed to be predicted with the help of stored QSPR model (see Section 3.1). In this case, this input value will be compared with the predicted one followed by the assessment of prediction performance.

5.1. Loading input data

Click **Tools** → **Property Prediction** of the ISIDA_QSPR main menu (Figure 1) to open the Consensus Predictor graphical interface (Figure 15). Click ***LOAD***, then select in the Open dialog box the TEST_AHIV-TIBO.SDF file from the list. The <...>\ISIDA_QSPR\TEST_AHIV-TIBO.SDF string appears under the upper LOAD button indicating that the selected file is downloaded. At the same time, the output file name <...>\ISIDA_QSPR\TEST_AHIV-TIBO_FMF.TSP appears under the SAVE button.

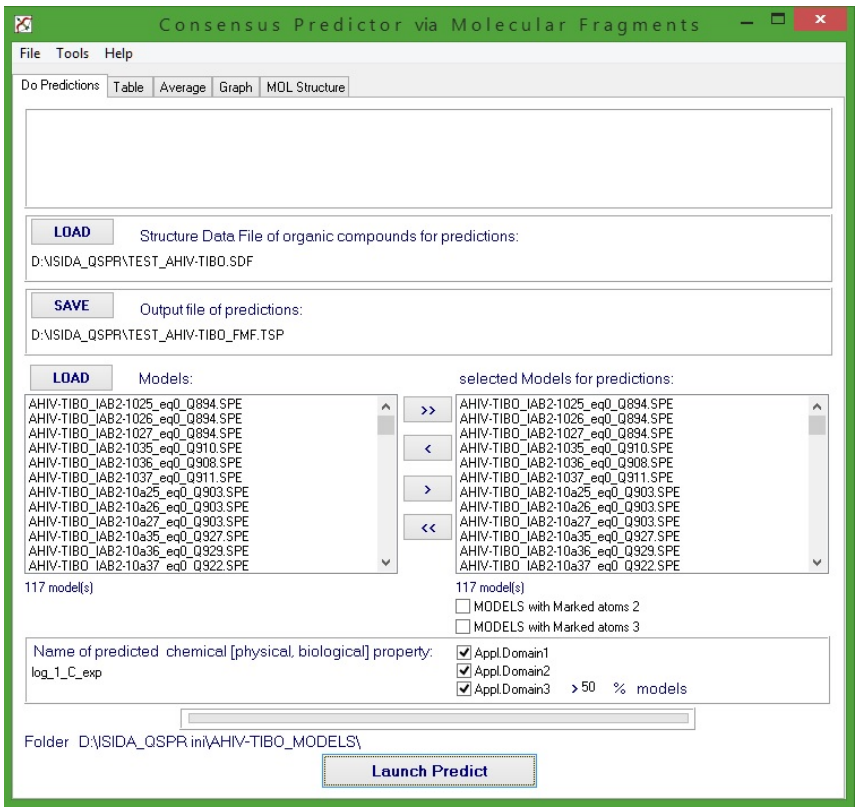


Figure 15. The Consensus Predictor program tool.

5.2. Loading selected models and choosing their applicability domain

Click the lower ***LOAD*** button on the Consensus Predictor window (Figure 15). In the Open dialog box open the AHIV-TIBO_MODELS directory containing stored MLR models (see Section 3.11), select any *.SPE file from the list. The *.SPE file names appear in the left list box of Consensus Predictor. Click the >> button to select all *.SPE files. The list of the 117 models appears in the right part of the box. Make sure that the ***MODELS with Marked atoms 2*** and ***MODELS with Marked atoms 3*** check boxes are not selected.

The program uses three types of AD definitions to ensure reliable predictions. Check the ***Appl. Domain1*** (AD1), ***Appl. Domain2*** (AD2) and ***Appl. Domain3*** check boxes for the strict AD control. Enter ***50*** in the edit box of percentage of applicable individual models (i.e. the models for which AD1 and AD2 do not discard the given molecule). If this number is lower than a threshold, the overall CM prediction is ignored.

5.3. Reporting predicted values

Click the ***Launch Predict*** button of Consensus Predictor (Figure 15) and agree to overwrite the output *.TSP file if it does exist. Results of the calculations are given in four tabs of the Consensus Predictor window. The ***Average*** tab reports for each molecule the average predicted property value (*Average*), its standard deviation (*STDEV*) estimated by CM and the number of the individual models (*Nm*) used for the *Average* value calculation. If for some compounds, experimental or somehow estimated property values are known (e.g., compounds 1 – 5); they are displayed in the *Datum* column. If a compound is identified as being outside AD, the predicted value for this compound is not given (e.g., for compound 11):

TABLE PA. Average predicted property log_1_C_exp

cmp. no.	Datum	Average	STDEV	Nm	Dat.- Ave.
1	7.92	7.96224E+000	5.925E-002	74	-4.224E-002
2	7.64	7.59814E+000	1.160E-001	99	4.186E-002
3	8.30	8.31903E+000	6.017E-002	74	-1.903E-002
4	7.86	7.50816E+000	3.707E-002	81	3.518E-001
5	7.53	7.50816E+000	3.707E-002	81	2.184E-002
6	-	6.67168E+000	4.470E-001	117	
7	-	7.07149E+000	1.500E-001	95	
8	-	7.53993E+000	1.078E-001	87	
9	-	8.98958E+000	1.957E-001	91	
10	-	7.48592E+000	4.201E-001	111	
11	-			0	
12	-	8.47588E+000	2.735E-001	98	

Click the ***Graph*** tab. For the compounds 1 – 5 with known activity, a plot displays a linear correlation between observed Y_{exp} and predicted Y_{pred} property and related statistical parameters.. Click on selected data point, and then click the ***MOL Structure*** tab in order to visualize a corresponding molecular structure.

5.4. Analysis of the fragments contributions

Click **Tools → Fragment Contributions** of the Consensus Predictor main menu (Figure 15) to open the *Fragment Contributions for Molecule* window (Figure 16). This window includes 2D structure , its constituting fragments, their occurrence and contributions in the context of selected individual MLR model. Enter **9** in the ***Mol Number*** edit box to examine the compound predicted as the most active, select any individual model by the dropdown ***Selected Model***.

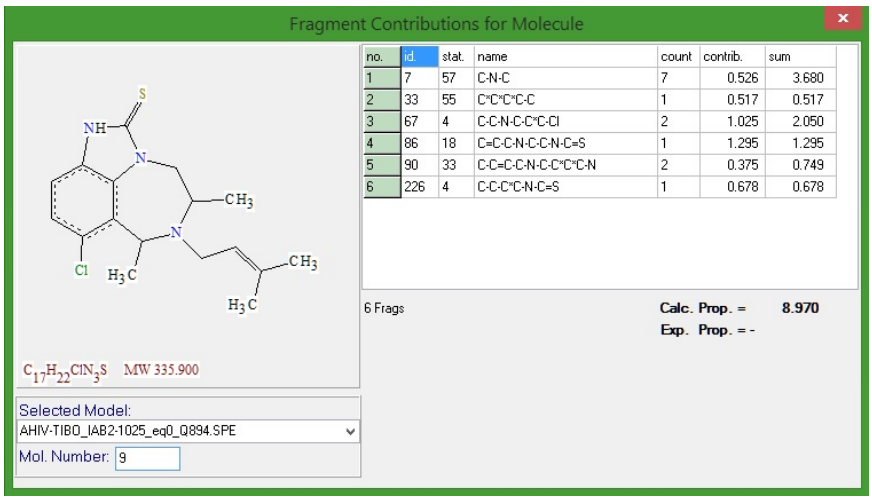


Figure 16. The graphic window of Fragment Contributions for Molecule.

References

1. Solov'ev, V. P.; Varnek, A. A. *ISIDA (In Silico Design and Data Analysis) program;* version 5.79; 2008 - 2014. http://infochim.u-strasbg.fr/spip.php?rubrique53 http://vpsolovev.ru/programs/ (accessed 21 August 2014).
2. Solov'ev, V. P.; Varnek, A.; Wipff, G., Modeling of ion complexation and extraction using substructural molecular fragments. *Journal of Chemical Information and Computer Sciences* **2000,** *40*, (3), 847-858.
3. Solov'ev, V. P.; Varnek, A. A., Structure-Property Modeling of Metal Binders Using Molecular Fragments. *Rus. Chem. Bull.* **2004,** *53*, (7), 1434-1445.
4. Solov'ev, V. P.; Kireeva, N.; Tsivadze, A. Y.; Varnek, A., QSPR ensemble modelling of alkaline-earth metal complexation. *J. Incl. Phenom. Macrocycl. Chem.* **2013,** *76*, (1-2), 159-171.
5. Solov'ev, V. P.; Varnek, A., Anti-HIV Activity of HEPT, TIBO, and Cyclic Urea Derivatives: Structure-Property Studies, Focused Combinatorial Library Generation, and Hits Selection Using Substructural Molecular Fragments Method. *J. Chem. Inf. Comp. Sci.* **2003,** *43*, (5), 1703-1719.
6. Varnek, A.; Solov'ev, V. P., "In Silico" Design of Potential Anti-HIV Actives Using Fragment Descriptors. *Comb. Chem. High Throughput Screening* **2005,** *8*, (5), 403-416.
7. Solov'ev, V. P.; Varnek, A. A. *EdChemS (Editor of Chemical Structures);* version 2.6; 2008 - 2014. http://infochim.u-strasbg.fr/spip.php?rubrique51 http://vpsolovev.ru/programs/ (accessed 21 August 2014).
8. Dalby, A.; Nourse, J. G.; Hounshell, W. D.; Gushurst, A. K. I.; Grier, D. L.; Leland, B. A.; Laufer, J., Description of several chemical structure file formats used by computer programs developed at Molecular Design Limited. *J. Chem. Inf. Comput. Sci.* **1992,** *32*, (3), 244-255.
9. Solov'ev, V.; Oprisiu, I.; Marcou, G.; Varnek, A., Quantitative Structure_Property Relationship (QSPR) Modeling of Normal Boiling Point Temperature and Composition of Binary Azeotropes. *Ind. Eng. Chem. Res.* **2011,** *50*, (24), 14162-14167.
10. Ruggiu, F.; Solov'ev, V.; Marcou, G.; Horvath, D.; Graton, J.; Le Questel, J.-Y.; Varnek, A., Individual Hydrogen-Bond Strength QSPR Modelling with ISIDA Local Descriptors: a Step Towards Polyfunctional Molecules. *Mol. Inf.* **2014,** *33*, (6-7), 477-487.
11. Varnek, A. A.; Wipff, G.; Solov'ev, V. P., Towards an Information System on Solvent Extraction. *Solv. Extr. Ion. Exch.* **2001,** *19*, (5), 791-837.

12. Varnek, A. A.; Wipff, G.; Solov'ev, V. P.; Solotnov, A. F., Assessment of the Macrocyclic Effect for the Complexation of Crown-Ethers with Alkali Cations Using the Substructural Molecular Fragments Method. *J. Chem. Inf. Comput. Sci.* **2002,** *42*, (4), 812-829.

13. Varnek, A.; Fourches, D.; Solov'ev, V. P.; Baulin, V. E.; Turanov, A. N.; Karandashev, V. K.; Fara, D.; Katritzky, A. R., "In Silico" Design of New Uranyl Extractants Based on Phosphoryl-Containing Podands: QSPR Studies, Generation and Screening of Virtual Combinatorial Library, and Experimental Tests. *J. Chem. Inf. Comp. Sci.* **2004,** *44*, (4), 1365-1382.

14. Katritzky, A. R.; Fara, D. C.; Yang, H.; Karelson, M.; Suzuki, T.; Solov'ev, V. P.; Varnek, A., Quantitative Structure-Property Relationship Modeling of beta-Cyclodextrin Complexation Free Energies. *J. Chem. Inf. Comp. Sci.* **2004,** *44*, (2), 529-541.

15. Varnek, A.; Fourches, D.; Hoonakker, F.; Solov'ev, V. P., Substructural Fragments: An Universal Language to Encode Reactions, Molecular and Supramolecular Structures. *J. Comput. Aid. Mol. Des.* **2005,** *19*, (9-10), 693-703.

16. Solov'ev, V.; Varnek, A.; Tsivadze, A., QSPR Ensemble Modelling of the 1:1 and 1:2 Complexation of Co2+, Ni2+, and Cu2+ with Organic Ligands. Relationships between Stability Constants. *J. Comput. Aided Mol. Des.* **2014,** *28*, (5), 549-564.

17. Reid, R. C.; Prausnitz, J. M.; Sherwood, T. K., *The Properties of Gases and Liquids*. McGraw-Hill Book Co: New York, 1977.

18. Lawson, C. L.; Hanson, R. J., *Solving Least Squares Problems*. Prentice Hall, Englewood Cliffs: New Jersey, 1974.

19. Kubinyi, H., Variable Selection in QSAR Studies. II. A Highly Efficient Combination of Systematic Search and Evolution. *Quant. Struct.-Act. Relat.* **1994,** *13*, (4), 393-401.

20. Zhokhova, N. I.; Baskin, I. I.; Palyulin, V. A.; Zefirov, A. N.; Zefirov, N. S., Fragmental Descriptors with Labeled Atoms and Their Application in QSAR/QSPR Studies. *Doklady Chemistry* **2007,** *417*, (2), 282-284.

21. Solov'ev, V. P.; Oprisiu, I.; Marcou, G.; Varnek, A., Quantitative Structure-Property Relationship (QSPR) Modeling of Normal Boiling Point Temperature and Composition of Binary Azeotropes. *Industrial & Engineering Chemistry Research* **2011,** *50*, (24), 14162-14167.

22. Solov'ev, V. P.; Tsivadze, A. Y.; Varnek, A. A., New Approach for Accurate QSPR Modeling of Metal Complexation: Application to Stability Constants of Complexes of Lanthanide Ions Ln3+, Ag+, Zn2+, Cd2+ and Hg2+ with Organic Ligands in Water. *Macroheterocycles* **2012,** *5*, (4-5), 404-410.

23. Varnek, A.; Kireeva, N.; Tetko, I. V.; Baskin, I. I.; Solov'ev, V. P., Exhaustive QSPR Studies of a Large Diverse Set of Ionic Liquids: How Accurately Can We Predict Melting Points? *J. Chem. Inf. Model.* **2007,** *47*, (3), 1111-1122.

24. Tetko, I. V.; Solov'ev, V. P.; Antonov, A. V.; Yao, X. J.; Fan, B. T.; Hoonakker, F.; Fourches, D.; Lachiche, N.; Varnek, A., Benchmarking of Linear and Non-Linear Approaches for Quantitative Structure-Property Relationship Studies of Metal Complexation with Organic Ligands. *J. Chem. Inf. Model.* **2006,** *46*, (2), 808-819.

25. Hartley, F. R.; Burgess, C.; Alcock, R. M., *Solution Equilibria*. John Wiley: Chichester, 1980; p 361.

26. Muller, P. H.; Neumann, P.; Storm, R., *Tafeln der mathematischen Statistik*. VEB Fachbuchverlag: Leipzip, 1979; p 280.

27. Solov'ev, V. P. *FMF (Forecast by Molecular Fragments). Predictions of metal-ligand stability constants.;* version 2.5; 2014. http://vpsolovev.ru/programs/ (accessed 21 August 2014).