MATCH
Communications in Mathematical and in Computer Chemistry

MATCH Commun. Math. Comput. Chem. **77** (2017) 245-272

ISSN 0340 - 6253

A Selection Method for Molecular Descriptors and QSPR Equations

Laszlo Tarko

Centre of Organic Chemistry – Romanian Academy, Romania, Bucharest, Sector 6, Spl. Independentei 202B, PO box 35-108, MC 060023

tarko_laszlo@yahoo.com

(Received September 19, 2016)

Abstract

The proposed method for the selection of the descriptors and QSPRs is a version of the heuristic Forward Stepwise procedure, using specific criteria to define the "significant" descriptors, the quality of QSPRs, the maximum number of predictors in QSPR, the "near constant" descriptors, the minimum value of the correlation descriptor/Property, the maximum value of the intercorrelation of the descriptors and including a specific criterion to stop the calculation. The method allows the selection of less than 1000 suitable descriptors included in a group of up to one hundred thousand significant descriptors. The method highlights the importance of the selection of the significant descriptors having a *low* correlation with the dependent property. If the number of selected significant descriptors is smaller than 1000, the heuristic Forward Stepwise procedure is a suitable method for the selection of the QSPR/QSAR equations. Two different QSARs obtained with the same database should be compared through the proposed function of leverages, calculated for the prediction set molecules. The presence or absence of "overfitting" depends on the number and type of descriptors in the initial set. The proposed method was applied with good results in QSAR studies regarding the anesthetic potency of 134 inhalation anesthetics (without validation/prediction set) and the toxicity of 50 phenol derivatives (including

validation/prediction set). The paper suggests testing some linear hydroxy-ketones, containing 6-12 carbon atoms and various distances between the hydroxyl and carbonyl groups, as new inhalation anesthetics. In analysis of λ_{max} for 66 derivatives of 9,10-anthraquinone the quality of PRECLAV prediction for the same calibration and validation sets is better than the quality of prediction made by ACO + MLR method.

1. Introduction

In QSP(A)R (*Quantitative Structure-Property(Activity) Relationship*) studies [45, 46] one uses a database which includes (mandatory) the *calibration set* (molecules having known values of the dependent property) and (optionally) the *prediction set* (molecules having unknown values of the dependent property and not used for building QSPRs). The calibration set is used to identify "the best" mathematical model, i.e. the QSPR which gives the minimum sum of square differences between the observed and the calculated values of the dependent property.

To find the best QSPR one uses descriptors (calculable molecular characteristics with the value expressed by real numbers) and some statistical procedures. The best equation includes several descriptors, named *predictors,* and it is used to calculate (predict) the value of the dependent property for all molecules in the database.

There are many programs [1–3] which calculate, before QSPR computation, thousands of descriptors. However, the best QSPR should include a small number of descriptors, for statistical reasons. Therefore, there is a huge number, usually more than 10^{18}, of *sets* of descriptors*,* which should be, as a rule, analyzed. The analysis of *all* sets would take millennia; a preliminary selection of descriptors (before QSPRs calculation) and a selection of descriptors sets (during calculation of QSPRs) are mandatory.

In addition, the selection of the descriptors should remove the "non-significant" descriptors having a very low correlation with the dependent property. Moreover, the selection of descriptor sets should remove the sets including too intercorrelated descriptors. In principle, these two rules increase the predictive power of the best QSPR.

One can quote various heuristic methods used in the selection of descriptors and QSPRs, such as Ant Colony Optimization [4, 47], Elastic Net [5], Forward/Backward Stepwise [6], Genetic Algorithm [7], Least Absolute Shrinkage and Selection Operator [8], Particle Swarm Optimization [9], Sequential Search [6], Variables Importance on PLS projections [10], Chemically Aware Model Builder [42] and review papers [11, 12]. The great number and diversity of these methods suggests that the issue of descriptors and QSPRs

selection is not yet entirely solved. Moreover, some authors [43] are skeptical regarding the utility of any selection methods: "*The results indicate that state-of-the-art learners (random forest, SVM, and neural networks) do not gain prediction accuracy from feature selection, and we found no evidence that a certain feature selection is particularly well-suited for use in combination with a certain learner.*"

Our paper presents the methods used, step by step, by the latest version (August 2016) of the PRECLAV (*PRoperty Evaluation by CLAss Variables*) software [2, 15], in the selection of descriptors and QSPR/QSAR equations. The reader can compare the presented methods and formulas with quoted procedures or other formulas and procedures in the field. The possibility of this comparison is the purpose of the text.

2. Methods and Formulas

The value of many descriptors depends on the molecular geometry. Identifying the molecular geometry with minimal potential energy is called "geometry optimization". The geometry optimization, after virtual building of the molecules, was done using the programs PCModel and MOPAC [13, 14].

Based on the output files created by MOPAC, the PRECLAV software calculated, for each molecule, more than 450 WM (*whole molecule)* descriptors, specific to this program, and tens of thousands of sums and products of these descriptors. The sums are considered a measure of the synergy of descriptors [16]. If the analyzed molecules include a common skeleton, it is recommended to calculate 3D descriptors. In this case, PRECLAV calculates, for each molecule, more than 600 3D descriptors, specific to this program, and other tens of thousands of sums and products of these descriptors.

In addition, we used more than 1300 descriptors calculated by the DRAGON software [1], a few descriptors calculated by the EPISuite software [17] and 40 aromaticity descriptors calculated by the DESCRIPT software [2].

PRECLAV identifies the molecular fragments using the bond orders of the chemical bonds which link heavy atoms (different from hydrogen) [16, 18]. The percentages (in weight) of the molecular fragment are the values of a certain descriptor. The absolute size and the algebraic sign of the correlation *r* of the percentages of the fragment and the values of the dependent property highlight the influence of the fragment on the dependent property. A "plus" correlation sign means that a *high mass percentage of this fragment increases the value*

of the property. A "minus" correlation sign means that a *high mass percent of this fragment decreases the value of the property.* The molecular fragments are considered "significant" if $r^2 \geq 4/(N+3)$; N is the number of molecules in the calibration set.

The selection of the descriptors and QSPRs is made in several phases.

Phase #1 eliminates the descriptors which fulfill the empirical criterion (1).

$$r^2 \leq k \tag{1}$$

where
r^2 is the square linear correlation descriptor – dependent property
$k = 1/N \bullet \ln(N)$
N is the number of molecules in the calibration set

The value of k is corrected if necessary in the range [0.01, 0.25], i.e. if $N > 647$, the value of k is considered 0.01 and if $N < 9$ the value of k is considered 0.25.

Phase #2 includes the following steps and loops.

a) identification and recording of the non-recorded descriptors having the *minimum* value of r^2 and $r^2 \geq k$ (see phase #1)

b) calculation of the weighting factors C_i of linear QSPR (2), by Ordinary Least Square Method (first value of p is 1)

$$P = C_0 + \sum_{i=1}^{p} C_i \cdot D_i \tag{2}$$

where
P is the computed value of the dependent property
C_0 is intercept
C_i are weighting factors (coefficients)
D_i are (the values of some) descriptors which fulfill the criterion $r^2 \geq k$
p is the number of descriptors

c) calculation of the value of the function Q by formula (3); Q is a measure of the quality of the QSPR (2), where r^2 is the square linear correlation observed values / calculated values of the dependent property

$$Q = r^2 \cdot \left(1 - \frac{p}{N}\right) \tag{3}$$

By applying the Forward Stepwise procedure in the b) + c) loop, the program identifies the non-recorded descriptor which, added (i.e. p becomes $p + 1$) to the descriptor(s)

in calculated QSPR, makes the QSPR having the *greatest* value of Q; the descriptor is recorded.

When the value of p increases, the value of r^2 increases, but the value of Q increases, reaches a maximum and then decreases.

d) if the value of Q decreases or $p \geq 2 \bullet \ln(N) + 2$ the program leaves the loop b) + c) and the calculation returns to step a)

e) if the number of recorded descriptors becomes greater than 950, the program leaves the loop a) – d)

Phase #3 eliminates the "near constant descriptors".

Using the minimum value V_{min} and maximum value V_{max} of a descriptor, the program divides the range $[V_{min}, V_{max}]$ in N+K intervals (classes), where K is the number of molecules in the prediction set (if K = 0 the prediction set is missing). Some classes are empty, other classes include few values. The program computes the Shannon Entropy SE_{cal} of values in the calibration set, using the Shannon's discontinuous formula (4) [19].

$$SE = -\sum_{i=1}^{C} x_i \cdot \ln x_i \qquad (4)$$

where
$x_i = n_i / N$
n_i is number of values in class i; $n_i > 0$
C is the number of non-empty classes which include values in the calibration set

The diversity D of the values in the calibration set is $SE_{cal}/\ln(N)$ ratio. The value of D is in the range [0, 1]. The descriptors having a small value of D, i.e. $D < 0.15$, are eliminated as "near constant descriptors".

Phase #4 includes the following steps.

a) building *all* linear QSPRs including two descriptors, $p = 2$ in formula (2)

The maximum square intercorrelation r^2 of descriptors in sets of two descriptors is $N^{-0.5}$.

In this step, the program computes Q quality using formula (3), and sorts the equations by Q value; the best 1000 equations are recorded.

b) by applying the Forward Stepwise procedure the program uses the recorded p descriptors and QSPRs and makes the QSPRs including $p + 1$ descriptors

The maximum square intercorrelation r^2 of descriptors in sets of $p > 2$ descriptors is $4 \bullet N^{-0.5}$, corrected if necessary in the range [0.1, 0.64].

In this step the program also computes Q quality and sorts the equations by Q value; in each loop the best 1000 equations are recorded.

c) the program leaves loop b) in three situations:

- if $p \geq 2 \bullet \ln(N) + 2$
- the value of Q of the best QSPR decreased
- no more descriptors can be added because the intercorrelations are too high, i.e. $r^2 > 4 \bullet N^{-0.5}$ for all combinations

Phase #5 calculates the "Prediction Set Leverage" function (5).

$$PSL = L_{ave} \cdot \frac{N}{p} \tag{5}$$

where L_{ave} is the average value of the leverages L_i, calculated for the prediction set molecules.

The latest version of PRECLAV computes the leverages L_i using the values of predictors for all molecules in the database (calibration set + prediction set) and formula (6).

$$L_i = x_i \cdot \left(C^T \cdot C\right)^{-1} \cdot x_i^T \tag{6}$$

In formula (6) x_i is a row vector for a particular molecule in the prediction set and **C** is the $N \bullet p$ matrix of p predictors for N calibration set molecules. If $L_i > 3 \bullet p / N$, the molecule in the prediction set is considered to be "outside of the Applicability Domain of the obtained QSAR" and the difference between the calculated value of the dependent property and the unknown observed value is, as a rule, big.

The value of PSL is calculated only if the database includes a prediction set.

3. Comments, Results and Discussions

Phase #1 is required because the "non-significant" descriptors having very low correlation with the dependent property should be eliminated before any other selection phase.

The Forward Stepwise procedure is, actually, a suitable method for the selection of predictors and QSPRs. If a certain descriptor is present, as predictor, in the best QSPR, it won the math competition with other descriptors and so one can consider the descriptor "suitable for description of the dependent property". However, the best description of the dependent property is not made by a certain predictor, but by *the best set* of predictors, *as a whole*.

Some more sophisticated methods, such as PLS (*Partial Least Square*), PCA (*Principal Component Analysis*), bootstrap, k-NN (*k-Nearest Neighbor*), SVM (*Support*

Vector Machine) or ANN (*Artificial Neural Network*), have one big disadvantage: the physical meaning of predictors is difficult to understand and thus the usefulness (for drug-design) of the equation obtained is small. In fact, these algorithms cannot be used for identification of the correlation between a *specific* molecular feature and the biochemical activity. This is why the program PRECLAV includes only the CODESSA-like BMLR (*Best Multi-Linear Regression*) procedure.

The forward procedure [20-22] yields QSPRs including *p* descriptors that *always* have a higher value of r^2 than any QSPR including a *smaller* number of descriptors. Actually, the important increase of the value of r^2, with the increase of the value of *p,* highlights a *low* intercorrelation of the descriptors, not a high correlation of the descriptors with the dependent property. Consequently, the key-step in Phase #2 is step a) and the key-word is *minimum*. If this word would be replaced by *maximum* the group of the selected descriptors would not include descriptors with low correlation with the dependent property, the maximum value of Q in Phase #4 would be achieved at a lower value of *p* and the best QSPR would be worse from the point of view of the value of r^2 and its predictive power. The results of this "worse" version of the algorithm are not presented here. If the initial number of analyzed descriptors is smaller this bias of the first selected descriptor is smaller, but the chance to identify a high quality equation (without overfitting) in Phase #4 is slim.

The value of r^2 in formulas (3) and (8), the value of Standard Error of Equation SEE in formula (7) and the value of Fisher function F in formula (8) are good measures of the quality of QSPRs, at least for moderate values of N. In practice, there is a high (inverse) correlation between r^2 and SEE.

$$SEE = \left[\frac{\sum (V_{obs} - V_{calc})^2}{N - p} \right]^{1/2} \tag{7}$$

$$F = \frac{r^2}{1 - r^2} \cdot \frac{N - p}{p} \tag{8}$$

However, r^2 and the statistical functions (7) and (8) cannot be used as criteria to stop the calculation because the value of r^2 continues to increase with the increasing value of *p*, the value of SEE continues to decrease with the increasing value of *p* and the value of F increases and decreases irregularly with the increasing value of *p*. Consequently, PRECLAV uses the specific function Q in formula (3) as a quality criterion *and* a criterion to stop the calculation, because when the value of *p* increases the value of Q increases, reaches a maximum and then it decreases. Sometimes, Q reaches the maximum value at a very high value of *p* and it is

necessary to impose an empirical maximum value for *p*. Therefore, according to the PRECLAV criteria, "the best" QSPR is the multilinear equation having the highest value of Q and *p* in the range [p_{min}, p_{max}], where $p_{min} = 2$ and $p_{max} = 2 \bullet \ln (N) + 2$.

The elimination of the "near constant descriptors" in Phase #3 is necessary because the presence of these descriptors causes some arithmetic problems. The use of Shannon Entropy, as a measure of the diversity of values and as a criterion for elimination, is specific for PRECLAV. This particular criterion is chosen because other module of the program uses the Shannon Entropy for various calculations regarding molecular diversity and similarity, without regard to selection of descriptors.

Phase #2 and Phase #4 are similar enough because they both use the Forward Stepwise procedure. However, the first step of Phase #4 is missing in Phase #2. In addition, Phase #4 includes the calculation of the intercorrelation of descriptors as criterion to build the sets of descriptors. In Phase #2 each descriptor is selected once. In Phase #4 the selected sets may include common descriptors.

There is a fair similarity of the Phase #4 and the CODESSA Best Multi-Linear Regression (BMLR) procedure [3]. If Phase #2 is missing, it should be replaced by another selection procedure, because the heuristic algorithms of PRECLAV and CODESSA do not work well in presence of > 1000 descriptors and the computation time is too high. This is why we added Phase #2 to the other selection procedures #1, #3 and #4 that have been previously presented briefly, without comments [2, 15, 16, 18].

In brief, the values of the statistical parameters in Phase #4 are

the maximum number of descriptors in QSARs $p = 2 \bullet \ln (N) + 2$
the minimum correlation r^2 descriptor-dependent property $r^2 = 1/N \bullet \ln (N)$
the maximum intercorrelation r^2 of descriptors in sets of 2 descriptors $r^2 = N^{-0.5}$
the maximum intercorrelation r^2 of descriptors in sets of > 2 descriptors $r^2 = 4 \bullet N^{-0.5}$

The values of these statistical parameters have not been theoretically defined and were determined empirically by the author. The value of these parameters influences the final result. PRECLAV offers to the user the possibility to change the default values, but the "human action" in defining of these parameters seems to be dangerous.

Formula (5) allows the comparison of different predictions for the *same* prediction set, made by different QSPRs that used the *same* calibration set but *different* sets of descriptors, selected in Phases #1 - #4. The best QSAR is the equation which has the lowest value for PSL.

For each QSAR, PRECLAV calculates r^2_{CV} using the Leave Half Out (LHO) method. However, this internal cross-validation method is applied *after ordering* the molecules in the calibration set according to the observed values of the dependent property. The function r^2_{CV} is viewed as a measure of the homogeneity of the calibration set from the viewpoint of QSAR, not as a result of a very drastic internal validation [39]. If $r^2_{CV} > 0.4$ the calibration set can be considered 'homogeneous enough'. The value of r^2_{CV} is not a criterion for the selection of descriptors and QSARs.

Further we present the result of four QSPR/QSAR studies which used the presented selection algorithm. We used a single processor / Pentium 4 / 3.2 GHz / 2 GB RAM computer.

Study #1

In this QSAR study the calibration set includes 134 organic inhalation anesthetics in Table 1. The dependent property is the Anesthetic Potency AP, defined as AP = log (1/MAC), where MAC is the *Minimum Alveolar Concentration*, i.e. the partial pressure of the vapors in lungs that prevents movement in 50% of subjects in response to pain (surgical) stimulus [23]. The experimental values of MAC and AP (for rats), see Table 1, are from literature [24-29].

Table 1 The observed/calculated values of Anesthetic Potency

MolID	Name	AP_{exp}	AP_{calc}
a001	Chloropentafluoroethane	-0.89	-0.431
a002	F2CH-O-CF2-CF3	-0.75	-0.186
a003	1,1,1,2,2-pentafluoropropane	-0.74	0.178
a004	1-bromoheptafluoropropane	-0.59	-0.043
a005	2-bromoheptafluoropropane	-0.57	-0.214
a006	1,1,1,2,2,3,3,4,4-nonafluorobutane	-0.53	-0.285
a007	1,1,1,2,2,3-hexafluoropropane	-0.38	-0.281
a008	Bromopentafluoroethane	-0.34	-0.224
a009	F3C-O-CHF-CF3	-0.29	-0.360
a010	1,1,1-trifluoroethane	-0.25	-0.049
a011	Pentafluoroethane	-0.18	-0.170
a012	1,2-dichlorotetrafluoroethane	0.01	0.088
a013	1,1,1,2,3,3,3-heptafluoropropane	0.02	-0.332
a014	Propane	0.03	0.28
a015	F2CH-O-CF2-CClF2	0.22	0.446
a016	1,1,1,2-tetrafluoroethane	0.25	0.118
a017	1,1,1,3,3,3-hexafluoropropane	0.25	-0.112
a018	1,1-difluroethane	0.48	0.668
a019	F2CH-O-CClF-CF3	0.49	0.682
a020	2,2-difluoropropane	0.52	0.623
a021	ClF2C-O-CH2-CF3	0.54	0.883

a022	ClF2C-O-CHF-CF3	0.54	0.495
a023	Butane	0.54	0.710
a024	Fluoroethane	0.61	0.539
a025	1,1,2,2-tetrafluoroethane	0.62	0.795
a026	2-fluoropropane	0.69	0.874
a027	*trans*-2-butene	0.70	0.745
a028	ClF2C-O-CF2-CCl2F	0.73	0.821
a029	Cl2FC-O-CF2-CClF2	0.74	0.582
a030	*cis*-2-butene	0.77	0.746
a031	Cyclopropane	0.80	0.356
a032	ClF2C-O-CCl2-CF3	0.88	1.045
a033	1,1,2,2,3,3-hexafluoropropane	0.89	0.362
a034	*n*-pentane	0.90	1.128
a035	1,1,2-trifluoroethane	0.94	1.371
a036	1,1,2,2,3,3,4,4,5,5,6,6-dodecafluorohexane	0.95	0.984
a037	F2CH-O-CH2-CF3	0.96	1.240
a038	F2CH-O-CCl2-CF3	1.01	1.330
a039	F2CH-O-CF2-CFCl2	1.05	1.035
a040	1,4-pentadiene	1.06	1.015
a041	F2CH-O-CHF-CF3	1.14	1.069
a042	Sevoflurane RN: 28523-86-6	1.22	1.529
a043	1,1,2,2,3,3,4,4-octafluorobutane	1.23	0.596
a044	*trans*-1,3-pentadiene	1.27	1.229
a045	Cyclopentane	1.28	1.299
a046	ClF2C-O-CHCl-CF3	1.31	1.337
a047	FCH2-CHF-CF3	1.33	0.116
a048	FCH2-O-CF2-CHF2	1.38	1.540
a049	Cyclohexane	1.38	1.717
a050	*cis*-1,3-pentadiene	1.41	1.229
a051	1,5-hexadiene	1.46	1.443
a052	Hexane	1.46	1.588
a053	ClF2C-CF2-CHClF	1.52	1.196
a054	*trans*-1,2-dichloroethylene	1.64	1.330
a055	1,3,5-trifluorobenzene	1.65	1.765
a056	Enflurane RN: 13838-16-9	1.66	1.443
a057	2,4-*trans,trans*-hexadiene	1.66	1.788
a058	Heptane	1.70	2.002
a059	1,1,2,4,4-pentafluorobutane	1.71	1.113
a060	(S)-isoflurane RN: 26675-46-7	1.77	1.836
a061	octane	1.77	2.388
a062	Hexafluorobenzene	1.79	1.365
a063	H3C-O-CF2-CHClF	1.80	1.761
a064	F2CH-O-CBrCl-CF3	1.82	2.048
a065	F2CH-O-CF2-CBrClF	1.82	1.755
a066	(R)-isoflurane RN: 26675-46-7	1.84	1.837
a067	Cycloheptane	1.86	2.246
a068	1,1,2,3,4,4-hexafluorobutane	1.89	0.961
a069	Pentafluorobenzene	1.90	1.631
a070	Halothane RN: 151-67-7	1.90	1.290

a071	1-hexyne	1.94	1.784
a072	Fluorobenzene	1.95	2.088
a073	*n*-nonane	1.95	2.736
a074	*cis*-1,2-dichloroethylene	1.98	1.332
a075	1-propanethiol	1.99	1.904
a076	Benzene	2.00	1.876
a077	1,2,4-trifluorobenzene	2.01	2.051
a078	H3C-O-CF2-CHBrF	2.16	2.117
a079	2,3,4,5,6-pentafluorotoluene	2.19	2.375
a080	1,4-difluorobenzene	2.19	2.152
a081	1,2-difluorobenzene	2.21	2.371
a082	3-hexyne	2.24	1.884
a083	F2CH-O-CHBr-CF3	2.28	1.797
a084	Toluene	2.35	2.315
a085	1-butanethiol	2.38	2.374
a086	2-butanone	2.55	2.491
a087	Methoxyflurane RN: 76-38-0	2.57	2.651
a088	1-pentanethiol	2.75	2.806
a089	2-pentanone	2.80	2.819
a090	1,4-dimethylbenzene	2.82	2.704
a091	1,3-dimethylbenzene	2.86	2.714
a092	Ethylbenzene	2.92	2.700
a093	1,2-dimethylbenzene	2.93	2.707
a094	Ethanol	3.00	3.018
a095	3-hexanone	3.07	3.152
a096	2,2,2-trifluoroethanol	3.16	3.398
a097	2-hexanone	3.21	3.186
a098	1-hexanethiol	3.23	3.166
a099	1,1,1-trifluro-2-propanol	3.28	3.907
a100	2,2,3,3,3-pentafluoro-1-propanol	3.28	3.363
a101	3,3,4,4,5,5,5-heptafluoro-2-pentanol	3.30	3.402
a102	2,2,3,3,4,4,5,5,6,6,6-undecafluoro-1-hexanol	3.34	3.313
a103	4-heptanone	3.35	3.442
a104	S-(+)-2-butanol	3.37	3.636
a105	2,2,3,3,4,4,5,5,6,6,7,7,7-tridecafluoro-1-heptanol	3.37	3.450
a106	1-propanol	3.39	3.431
a107	2,2,3,3,4,4,4-heptafluoro-1-butanol	3.4	3.389
a108	R-(-)-2-butanol	3.44	3.636
a109	S-(+)-2-pentanol	3.62	4.062
a110	1,1,1,3,3,3-hexafluoro-2-methyl-2-propanol	3.64	3.265
a111	3,3,4,4,5,5,6,6,7,7,8,8,8-tridecafluoro-1-octanol	3.65	3.646
a112	2-heptanone	3.68	3.512
a113	4-octanone	3.71	3.693
a114	R-(-)-2-pentanol	3.76	4.062
a115	1-butanol	3.88	3.803
a116	R-(-)-2-hexanol	4.06	4.328
a117	2,2,3,3,4,4,5,5,6,6,7,7-dodecafluoro-1-	4.10	4.189

	heptanol		
a118	S-(+)-2-hexanol	4.12	4.328
a119	2,2,3,3-tetrafluoro-1-propanol	4.24	4.243
a120	4-heptanol	4.34	4.606
a121	2,2,3,4,4,4-hexafluoro-1-butanol	4.35	3.944
a122	1,1,1,3,3,3-hexafluoro-2-propanol	4.35	3.938
a123	S-(+)-2-heptanol	4.41	4.631
a124	R-(-)-2-heptanol	4.43	4.631
a125	1-pentanol	4.58	4.138
a126	1-hexanol	4.67	4.391
a127	2,2,3,3,4,4,5,5-octafluoro-1-pentanol	4.70	4.559
a128	1-octanol	5.93	4.622
a129	tetrafluoromethane	-1.82	-1.392
a130	Trifluoromethane	-0.20	-0.127
a131	Dichlorodifluoromethane	-0.05	0.003
a132	Difluoromethane	0.22	0.404
a133	ClF2C-O-CF2-CF3	0.54	-0.351
a134	Difluorodibromomethane	0.96	0.663

If the best QSAR is obtained using *all* available molecules the equation cannot be validated by external validation because other molecules are not available to make a validation set. On the contrary, if the best QSAR is obtained using a training set *extracted* from the initial set of molecules, the equation can be validated (non-extracted molecules are included in the validation set), but the validated equation is *very different* from the point of view of predictors and weighting factors. The usefulness of the validated equation for identification of the significant molecular features and of the *outliers for lead hopping* [40, 41] is smaller. This is why, in this QSAR study, the validation set does not exist and the validation is not applied.

Number of calculated descriptors is 9,380

Number of selected descriptors after Phase #1 is 8,555

Number of selected descriptors after Phase #2 is 978

Computation time in Phase #2 of selection was 50 hours 10 minutes.

Number of selected descriptors after Phase #3 is 928

The list of significant molecular fragments and their correlation with AP is

HO	r = 0.6165	F	r = - 0.4154
CH_2	r = 0.3413	C	r = - 0.2667
O	r = - 0.2290	CH_3	r = 0.2177
Cl	r = - 0.2004	CO	r = 0.1989

According to the sign and the absolute values of *r,* a high percentage in weight of OH (hydroxyl), CH_2 (methylene), CH_3 (methyl) and CO (carbonyl) molecular fragments increases the anesthetic potency.

A high percent in weight of F (fluorine), O (ether), Cl (chlorine) and C (tetra-substituted C atom) fragments decreases the anesthetic potency. This is an unexpected statistical result because many usual inhalation anesthetics are fluorinated/chlorinated ethers (see Table 1).

The best type (2) QSAR:

$C_0 = 3.2270$
$C_1 = - 0.3386$
D_1 is the sum d_1+d_2 where
  d_1 is logP in octanol-water system [17]
  d_2 is ln (1+VP) where VP is the Selected Vapor Pressure [17]
$C_2 = 45.1427$
D_2 is the product $d_3 \bullet d_4$ where
  d_3 is Number of OH bonds
  d_4 is Minimum free valence of H atoms
$C_3 = 0.0680$
D_3 is Total information index of atomic composition [30]
$C_4 = 0.0055$
D_4 is the product $d_5 \bullet d_6$ where
  d_5 is Maximum net charge of atoms
  d_6 is Heat of formation (Kcal/mol)

The minimum square correlation predictor / AP is $r^2 = 0.0390$ for D_4 predictor.
The maximum intercorrelation of predictors is $r^2 = 0.2634$ for D_1/D_2 pair.

The physical/chemical meaning of the descriptors and the algebraic sign of the coefficients emphasize the favorable influence on AP of the presence of OH and CO fragments, high enough molecular mass and a diversity of atoms. According to the list of significant molecular fragments and QSAR's structure, we suggest testing some *linear* hydroxy-ketones, not included in Table 1, containing 6-12 carbon atoms and various distances between the OH and CO groups, as new inhalation anesthetics.

The computation time in Phase #4 of selection was 2 hours.

The quality of (the prediction by) the best QSAR:

$p = 4$ $N = 134$ $r^2 = 0.9489$ $Q = 0.9206$ $F = 603.3$ $SEE = 0.340$

According to the value of r^2_{CV} the homogeneity of the calibration set from the viewpoint of QSAR #1 is low, $r^2_{CV} = 0.2939$.

The last column in Table 1 presents the calculated values of AP. According to the PRECLAV criterion Q • SEE / $|AP_{calc} - AP_{exp}| < 0.28$ the molecules **a047** and **a128** are atypical (outliers) from the point of view of the obtained QSAR.

Figure 1 presents the scatter-plot of the Observed/Calculated values of AP. The big black points indicate the two outlier values.

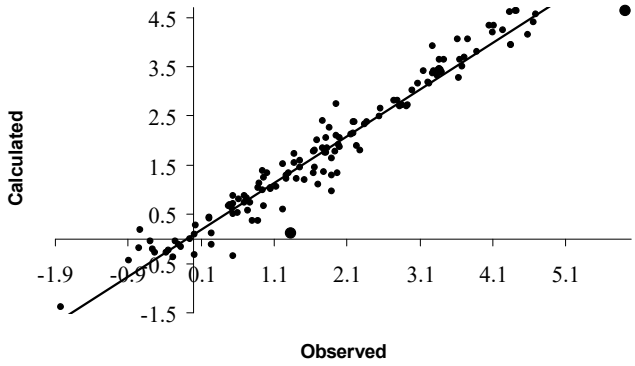

Fig. 1 Scatter-plot of the Observed/Calculated values of AP

One can compare the obtained QSAR with the QSAR of the Abraham Model [24]. This model uses five predictors ($p = 5$), the calibration set used includes inorganic anesthetics, such as noble gases and nitrogen oxides (N = 148), and the prediction of AP values is very good ($r^2 = 0.9850$ Q = 0.9517 F = 1878.1 SEE = 0.192). However, the predictors of the Abraham Model are "experimental descriptors", i.e. descriptors for which the value was not calculated, but measured experimentally, by chromatographic methods [31-33]. Thus, the Abraham Model requires synthesis and chromatographic analysis of the molecules for which the QSAR equation calculates the AP values.

Study #2

In this QSAR study the calibration set includes 34 phenol derivatives, see the molecules **P01 – P34** in Table 2. The dependent property is the toxicity T against protozoan *Tetrahymena pyriformis*, widely used in laboratory research due to its sensitivity to water pollution. The values of the toxicity T, weighted here within the range [0, 2.638], are from literature [34, 35].

In this QSAR study the prediction set is present (molecules **P35 – P50** in Table 2). To avoid subjectivity in choosing the prediction set we ordered the molecules in Table 2 according to the observed values of toxicity. The molecules having rank 3, 6, 9, 12, etc. have been included in the prediction set.

Table 2 The structure and observed value of Toxicity of phenol's derivatives

Mol ID	Substituent(s)	T_{obs}	Mol ID	Substituent(s)	T_{obs}
P01	none	0.000	**P26**	2,4,6-trimethyl	2.126
P02	2,6-difluoro	0.827	**P27**	3,5-dimethyl, 4-chloro	1.634
P03	4-fluoro	0.448	**P28**	2,6-dichloro, 4-bromo	2.210
P04	3-fluoro	0.904	**P29**	2-methyl, 4-bromo, 6-chloro	1.708
P05	4-methyl	0.239	**P30**	2,4,6-tribromo	2.481
P06	2-chloro	0.708	**P31**	2-*tert*-butyl, 4-methyl	1.728
P07	4-chloro	0.976	**P32**	2-*iso*-propyl, 4-chloro, 5-methyl	2.293
P08	3-ethyl	0.660	**P33**	2,6-diphenyl	2.544
P09	2-ethyl	0.607	**P34**	2,4-dibromo, 6-phenyl	2.638
P10	4-bromo	1.112	**P35**	2-fluoro	0.679
P11	2,3-dimethyl	0.553	**P36**	3-methyl	0.369
P12	2,5-dimethyl	0.440	**P37**	2-bromo	0.935
P13	3,4-dimethyl	0.553	**P38**	2,4-dimethyl	0.559
P14	4-iodo	1.285	**P39**	3,5-dimethyl	0.544
P15	2-*iso*-propyl	1.234	**P40**	3-chloro, 4-fluoro	1.273
P16	3-*iso*-propyl	1.040	**P41**	2-chloro, 5-methyl	1.071
P17	4-*iso*-propyl	0.904	**P42**	3-iodo	1.549
P18	2,5-dichloro	1.559	**P43**	3-*tert*-butyl	1.161
P19	2,3-dichloro	1.702	**P44**	3,5-dichloro	1.993
P20	2-methyl, 4-chloro	1.131	**P45**	2,3,6-trimethyl	0.849
P21	3-methyl, 4-chloro	1.226	**P46**	3,4,5-trimethyl	1.361
P22	2,4-dichloro	1.467	**P47**	2,4,5-trichloro	2.531
P23	4-*tert*-butyl	1.344	**P48**	2,6-dimethyl, 4-bromo	1.709
P24	2-phenyl	1.525	**P49**	2,4-dimethyl, 6-*tert*-butyl	1.676
P25	2,4-dibromo	1.834	**P50**	2,6-di- *tert*-butyl, 4-methyl	2.219

This QSAR study did not analyze the huge number of sums and products of WM and 3D descriptors. However, the number of descriptors analyzed is more than two thousand.

Number of calculated descriptors is 2,189

Number of selected descriptors after Phase #1 is 1,214

Number of selected descriptors after Phase #2 is 978

Computation time in Phase #2 of selection was 1 hour 11 minutes.

Number of selected descriptors after Phase #3 is 889

The list of significant molecular fragments and correlation with T:

HO	r = - 0.5787
Br	r = 0.4897
C_6H_2OH	r = 0.4343
C_6H_4	r = - 0.4067
C_6H_2	r = 0.3319

According to the sign and the absolute values of *r,* a high percentage in weight of Br (bromine), C_6H_2OH (tri-substituted cycle conjugated with OH) and C_6H_2 (tri-substituted cycle, weak conjugated with OH) fragments increases the toxicity.

A high percentage in weight of OH (hydroxyl, weak conjugated with cycle) and C_6H_4 (mono-substituted cycle, weak conjugated with OH) fragments decreases the toxicity. All molecules include just one OH group. Consequently, "a high percent in weight of OH decreases the toxicity" actually means "a high molecular mass increases the toxicity".

The best type (2) QSAR:

$C_0 = - 0.7363$
$C_1 = 0.2819$
D_1 is solvation connectivity index χ_0 [1, 37]
$C_2 = - 0.0072$
D_2 is the sum d_1+d_2 where
d_1 is the percent in weight of C_6H_4 molecular fragment
d_2 is the percent in weight of C_6H_3 molecular fragment

The minimum square correlation predictor / T is $r^2 = 0.3237$ for D_2 predictor.
The intercorrelation of predictors is $r^2 = 0.1112$.

The physical/chemical meaning of descriptors and the algebraic sign of coefficients emphasizes, in our opinion, the favorable influence on toxicity of a high molecular volume. Consequently, the presence of high-volume substituents should increase the toxicity, as revealed earlier Duchowicz *et. al.* [36] using different statistical methods, although I (iodine) and C_6H_5 were not identified here as "significant" molecular fragments.

Computation time in Phase #4 of selection was 54 minutes.

The quality of (prediction by) the best QSAR:

p = 2 N = 34 r^2 = 0.8615 Q = 0.8108 F = 99.5 SEE = 0.261

According to PRECLAV criterion there are no outlier molecules in the calibration set, from the point of view of the QSAR obtained.

The columns "by QSAR #2" in Table 3 present the calculated values of T in the calibration and prediction set.

Figure 2 presents the scatter-plot of the Observed/Calculated values of T in the calibration set.

As a rule, the molecules in the prediction set are new, not yet synthesized and the observed (experimental) values of the dependent property are unknown. However, here the experimental values of T for the prediction set molecules are known. Therefore, the prediction for molecules **P35 – P50** in the prediction set**,** not used for QSAR building, is, actually, a validation test for the obtained QSAR. The quality of prediction for the molecules in the prediction set is acceptable (r^2 = 0.6978 SEE = 0.379 PSL = 0.501 and figure 3).

There are no molecules in the prediction set outside of the Applicability Domain of QSAR #2.

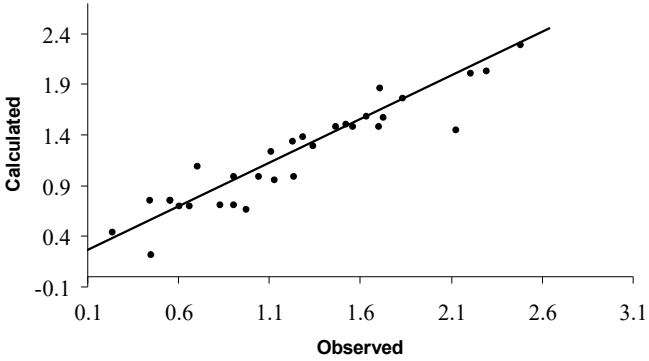

Fig. 2 Scatter-plot of the Observed/Calculated values of T in calibration set (QSAR #2)

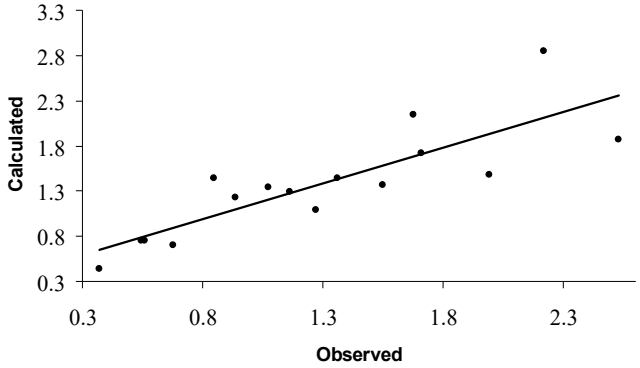

Fig. 3 Scatter-plot of the Observed/Calculated values of T in validation set (QSAR #2)

Table 3 The calculated value of Toxicity of phenol's derivatives

MolID	by QSAR #2	by QSAR #3	No.	by QSAR #2	by QSAR #3
P01	0.705	-0.031	**P26**	1.441	1.625
P02	0.705	0.641	**P27**	1.582	1.647
P03	0.214	0.6	**P28**	2.005	2.083
P04	0.705	0.864	**P29**	1.864	1.550
P05	0.442	0.233	**P30**	2.287	2.402
P06	1.091	0.822	**P31**	1.570	1.771
P07	0.663	0.935	**P32**	2.026	2.207
P08	0.699	0.8	**P33**	2.951	2.624
P09	0.699	0.622	**P34**	2.882	2.769
P10	1.232	1.102	**P35**	0.705	0.507
P11	0.751	0.602	**P36**	0.442	0.414
P12	0.751	0.463	**P37**	1.232	0.982
P13	0.751	0.735	**P38**	0.751	0.627
P14	1.373	1.265	**P39**	0.751	0.701
P15	0.991	1.003	**P40**	1.091	1.543
P16	0.991	1.11	**P41**	1.336	1.248
P17	0.991	0.921	**P42**	1.373	1.290
P18	1.477	1.56	**P43**	1.289	1.597
P19	1.477	1.727	**P44**	1.477	1.853
P20	0.956	1.083	**P45**	1.441	1.613
P21	1.336	1.383	**P46**	1.441	0.987
P22	1.477	1.64	**P47**	1.864	2.292
P23	1.289	1.349	**P48**	1.723	2.623
P24	1.505	1.442	**P49**	2.146	11.524
P25	1.759	2.092	**P50**	2.850	13.096

If *all* molecules in Table 2 are included in the calibration set (and so, the validation is avoided) the results are very different. For instance, the list of the significant molecular fragments and correlation with toxicity is different:

HO	r = - 0.5363
C_6H_2OH	r = 0.4549
C_6H_4	r = - 0.3858
Br	r = 0.3747
C_6H_3	r = - 0.3286
Cl	r = 0.3189

Study #3

This QSAR study used the same database (calibration set + prediction set) as QSAR study #2, see Table 2. However, this QSAR study analyzed all descriptors used in QSAR study #2 plus the sums and products of WM and 3D descriptors. Consequently, the initial number of the analyzed descriptors is almost 94,000.

Number of calculated descriptors is 93,888

Number of selected descriptors after Phase #1 is 92,707

Number of selected descriptors after Phase #2 is 979

Computation time in Phase #2 of selection was 149 hours 43 minutes.

Number of selected descriptors after Phase #3 is 975

The best type (2) QSAR:

C_0 = - 3.3665
C_1 = - 1.3391
D_1 is Verhaar model of Fish base-line toxicity from MlogP,
i.e. D_1 = -0.85 • MLogP - 1.39 [1, 38]
C_2 = - 1.1535
D_2 is the sum d_1+d_2 where
d_1 is Maximum parallax for probe atom #80
d_2 is Rejection force sum on probe atom #88
C_3 = 1233.3871
D_3 is the product $d_3 \bullet d_4$ where
d_3 is Rejection force sum on probe atom #104
d_4 is Resultant electrostatic force on probe atom #5

The minimum square correlation predictor / T is $r^2 = 0.2049$ for D_3 predictor.
The maximum intercorrelation of predictors is $r^2 = 0.4350$ for D_1/D_2 pair.

The physical/chemical meaning of descriptors and the algebraic sign of coefficients emphasizes the influence on toxicity of the 3D position of the atomic net charges ("electrical shape"). Computation time in Phase #4 of selection was 1 hour 56 minutes.

The quality of (prediction by) the best QSAR #3:

p = 3 N = 34 r^2 = 0.9585 Q = 0.8739 F = 238.7 SEE = 0.145

The columns "by QSAR #3" in Table 3 present the calculated values of T in the calibration and prediction set.

The figure 4 presents the scatter-plot of the Observed/Calculated values of T in the calibration set. According to the PRECLAV criterion, the molecule **P26** is an outlier from the point of view of the obtained QSAR. The big black point indicates the outlier value.

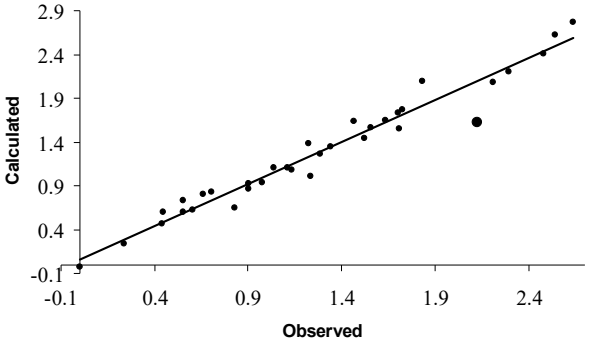

Fig. 4 Scatter-plot of the Observed/Calculated values of T in calibration set (QSAR #3)

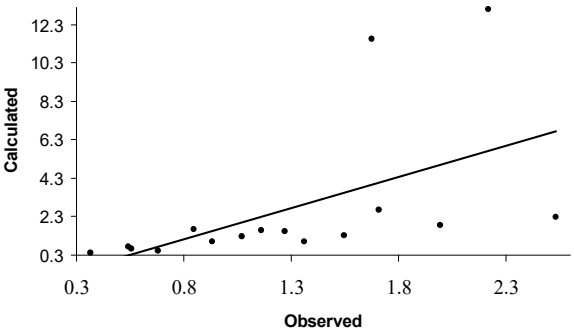

Fig. 5 Scatter-plot of the Observed/Calculated values of T in validation set (QSAR #3)

The leverages of the molecules **P49** and **P50** are much greater than the leverages of the other molecules in the prediction set. However, the molecules **P49** and **P50** are inside the Applicability Domain of QSAR #3. Therefore, there are no molecules in the prediction set outside the Applicability Domain of QSAR #3.

The quality of prediction by the best QSAR #3, for molecules in the *calibration* set is much higher than the quality of prediction by the best QSAR #2. However, the quality of prediction for molecules in the *prediction* set is much lower from the point of view of r^2 and SEE ($r^2 = 0.2889$ SEE = 4.445 and figure 5). The calculated values of T for prediction set molecules are ordered quite right, despite a low agreement with the observed values. In the author's opinion the value of the difference $r^2_{cal} - r^2_{val}$ in the last QSAR study emphasizes "overfitting", actually a very low similarity between calibration and validation sets from the point of view of QSAR #3. The presence or absence of "overfitting" depends on the number and type of descriptors in the initial set.

If the observed values of T for the prediction set molecules are unknown, we should choose for prediction QSAR #2 (PSL = 0.501) and not QSAR #3 (PSL = 0.777).

Study #4

This QSPR study allows a comparison with the ACO (Ant Colony Optimization) method [4, 47] combined with MLR (Multi Linear Regression), applied by Atabati *et al.* [44] in calculation of λ_{max} (nm) for 66 derivatives of 9,10-anthraquinone. We used the same calibration set (first 36 molecules in Table 4) and validation set (last 30 molecules in Table 4) as quoted authors.

Table 4 Structure and λ_{max} of analyzed anthraquinone derivatives

MolID	Substituents	Obs. λ_{max}	Calc. λ_{max}	
			PRECLAV	ACO + MLR
lam01	none	327	306.9	323.3
lam02	2,3-diBr	330	317.0	329.2
lam03	1,8-diCl	344	341.5	348.7
lam04	1,5-diCl	347	345.4	350.6
lam05	1,4-diCl	350	367.3	358.8
lam06	2-OH	365	354.6	350.6
lam07	1,8-diOCH_3	385	390.9	457.3
lam08	2-NH_2, 3-Br	406	415.2	416.1
lam09	1-$NHCOCH_3$	410	409.0	410.3
lam10	1-$NHCOC_6H_5$	415	419.7	448.4
lam11	2-NH_2, 3-NO_2	420	464.1	421.1
lam12	1,5-diOH	428	420.1	420.9
lam13	1,8-diOH	430	437.3	422.4

lam14	1-SCH_3	438	458.5	441.6
lam15	2,3-diNH_2	442	461.5	471.5
lam16	1-NH_2, 4-NO_2	460	449.8	484.2
lam17	1-NH_2, 2-CH_3	465	476.7	463.8
lam18	1-NH_2	465	460.6	465.7
lam19	1-NH_2, 4-Cl	466	481.4	472.9
lam20	1-NH_2, 6-Cl	470	472.0	448.8
lam21	1,4-diOH	476	494.1	442.5
lam22	1-NH_2, 6,7-diCl	477	476.3	432.0
lam23	1,8-diNH_2	492	485.1	527.3
lam24	1-NH_2, 4-OCH_3	500	519.0	499.2
lam25	1-$N(CH_3)_2$	504	462.7	482.3
lam26	1-NHC_6H_5	508	487.7	509.4
lam27	1-NO_2, 4,5,8-triOH	510	504.4	490.8
lam28	1-$NHCH_3$, 4-Br	510	499.8	503.4
lam29	1-OH, 4-NH_2	520	521.4	503.0
lam30	1-NH_2, 4-$NHCOC_6H_5$	532	536.3	526.4
lam31	1-$NHCH_3$, 4-OCH_3	540	541.9	534.8
lam32	1,4-diNH_2	550	555.0	571.9
lam33	1-NH_2 4-$NHCH_3$	590	583.0	558.3
lam34	1-NH_2, 4-NHC_6H_5	590	563.9	611.1
lam35	1,4,5,8-tetraNH_2	610	605.9	604.1
lam36	1,4-di$NHCH_3$	620	606.1	588.5
lam37	1,4-diNH_2, 2-NO_2	645	546.6	563.7
lam38	1-$NHCH_3$, 4-NHC_6H_5	625	586.7	613.2
lam39	1,4-diNHC_6H_5	620	573.7	620.0
lam40	1,5-diNH_2, 4,8-diOH	590	581.2	585.4
lam41	1-OH, 4-NHC_6H_5	566	542.9	535.0
lam42	1,4-diNH_2, 2-OCH_3	550	539.8	566.1
lam43	1-OH, 2,4-diNH_2	530	507.3	529.9
lam44	1-$NHCH_3$	508	490.2	491.0
lam45	1,4-di$NHCOC_6H_5$	490	486.1	505.7
lam46	1,5-diNH_2	480	459.3	480.0
lam47	1,2-diNH_2	480	502.9	478.2
lam48	1-NH_2, 2-$NHCOC_6H_5$	475	471.1	464.6
lam49	1-NH_2, 2-CH_3, 4-Br	473	487.0	481.5
lam50	2-$NHCH_3$	470	441.3	476.7
lam51	1-NH_2, 5-OCH_3	460	452.1	499.5
lam52	1,2-diOH	416	410.9	383.1
lam53	2-NH_2, 3-Cl	414	417.1	405.2
lam54	2-NH_2	410	417.0	431.7
lam55	1-NO_2, 2-NH_2	410	435.0	415.5
lam56	1-Cl, 2-NH_2	405	413.4	423.2
lam57	1-OH	405	402.0	386.0
lam58	1-OCH_3	380	385.4	410.3
lam59	2-OCH_3	363	377.2	380.2
lam60	1-Cl	337	345.1	346.5
lam61	1-NO_2, 4-Cl	335	325.7	423.0
lam62	2,7-diCl	330	323.6	307.0
lam63	2,6-diCl	330	320.2	305.3

lam64	2,3-diCl	330	332.9	309.8
lam65	2-Cl	330	323.0	320.0
lam66	2-F	325	299.5	321.6

Number of calculated descriptors is 47,612

Number of selected descriptors after Phase #1 is 46,594

Number of selected descriptors after Phase #2 is 976

Computation time in Phase #2 of selection was 130 hours 23 minutes.

Number of selected descriptors after Phase #3 is 974

The best PRECLAV equation:

C_0 = - 223.5456
C_1 = + 0.5659
D_1 is the empirical non-linear descriptor 7800 / ($E_{LUMO} - E_{HOMO}$) - 600
C_2 = + 249.6675
D_2 is the sum d_1+d_2 where
d_1 is Maximum bond order (C-C bonds)
d_2 is Maximum free valence (O atoms)
C_3 = + 41114.1133
D_3 is Resultant electrostatic force on probe atom #79 (3D descriptor)

The calculated values of λ_{max} are presented in fourth column of Table 4.

The quality of prediction for calibration set:

p = 3 N = 36 r^2 = 0.9609 Q = 0.8808 F = 270.3 SEE = 16.0

The quality of prediction for validation set:

p = 3 N = 30 r^2 = 0.9506 Q = 0.8555 F = 173.2 SEE = 26.4 PSL = 0.767

The best ACO + MLR equation [44] includes four DRAGON descriptors and the energy of HOMO molecular orbital:

$$\lambda_{max} = 1003.4 + 154.4 \bullet MATS4e + 654.5 \bullet RDF020v - 485.0 \bullet RDF020p - 80.3 \bullet RTe^{+} + 64.4 \bullet E_{HOMO}$$

The calculated values of λ_{max} are presented in fifth column of Table 4.

The quality of prediction for calibration set:

p = 5 N = 36 r^2 = 0.9195 Q = 0.7918 F = 70.8 SEE = 23.7

The quality of prediction for validation set:

p = 5 N = 30 r^2 = 0.9150 Q = 0.7625 F = 53.8 SEE = 30.8 PSL = 1.055

4. Conclusions

The proposed method for the selection of the descriptors and QSPRs is a version of the heuristic Forward Stepwise procedure, using specific criteria to define the "significant" descriptors, the quality of QSPRs, the maximum number of predictors in QSPR, the "near constant" descriptors, the minimum value of the correlation descriptor/Property, the maximum value of the intercorrelation of the descriptors and a specific criterion for stopping the calculation.

The computation time in the selection of the descriptors is directly proportional to the number of significant descriptors.

The method allows the selection of less than 1000 suitable descriptors from a group of tens of thousands significant descriptors. The method highlights the importance of the selection of the significant descriptors having a *low* correlation with the dependent property.

If the number of the selected significant descriptors is smaller than 1000, the heuristic Forward Stepwise procedure is a suitable method for the selection of the QSPR/QSAR equations.

Two different QSARs obtained with the same database should be compared through the proposed function of leverages, calculated for the prediction set molecules. The presence or absence of "overfitting" depends on the number and type of descriptors in the initial set.

According to the result of QSAR study #1 the paper suggests testing some linear hydroxy-ketones, containing 6-12 carbon atoms and various distances between the OH and CO groups, as new inhalation anesthetics.

In analysis of λ_{max} for 66 derivatives of 9,10-anthraquinone the quality of PRECLAV prediction for the same calibration and validation sets is better than the quality of prediction made by ACO + MLR method.

References

[1] DRAGON program is available from Talete srl., via V. Pisani, 13-20124, Milano, Italy; http://www.talete.mi.it

[2] in-house PRECLAV and DESCRIPT programs, documentation included, are available from Center of Organic Chemistry – Bucharest – Romanian Academy; ltarko@ccocdn.ro; tarko_laszlo@yahoo.com .

[3] CODESSA software is available from http://www.codessa-pro.com/

[4] M. Dorigo, C. Blum, Ant colony optimization theory: A survey, *Theor. Comp. Sci.* **344** (2005) 243–278.

[5] T. Hastie, R. Tibshirani, J. Friedman, *The Elements of Statistical Learning*, Springer, New York, 2009.

[6] A. Miller, *Subset Selection in Regression*, Chapman & Hall, New York, 2002.

[7] R. Leardi, *Nature-Inspired Methods in Chemometrics: Genetic Algorithms and Artificial Neural Networks, Data Handling in Science and Technology*, Elsevier, Amsterdam, 2003.

[8] R. Tibshirani, Regression shrinkage and selection via the Lasso, *J. Royal Stat. Soc. B* **58** (1996) 267–288.

[9] L. Bliss, *Particle Swarm Optimization*, Pace Univ. DPS program, 2002.

[10] S. Wold, *PLS for Multivariate Linear Modeling – Chemometric Methods in Molecular Design*, Wiley–VCH, Weinheim 1995.

[11] B. Ratner, Variable selection methods in regression: Ignorable problem, outing notable solution, *J. Targ., Meas. Anal. Mark.* **18** (2010) 65–75.

[12] I. Guyon, A. Elisseeff, An introduction to variable and feature selection, *J. Mach. Lear. Res.* **3** (2003) 1157–1182.

[13] PCModel program is available from Gajewski, J. J.; Gilbert, K. E., Serena Software, Box 3076, Bloomington, IN, USA.

[14] MOPAC program is available from Stewart, J.J.P., 15210 Paddington Circle, Colorado Springs, CO 80921; MrMOPAC@OpenMOPAC.net; http://www.openmopac.net/ .

[15] L. Tarko, C. T. Supuran, QSAR studies of sulfamate and sulfamide ionhibitors targeting human carbonic anhydrase isozymes I, II IX and XII, *Bioorg. Med. Chem.* **21** (2013) 1404–1409.

[16] L. Tarko**,** A statistical method for calculation of intramolecular synergy, *MATCH Commun. Math. Comput. Chem.* **75** (2016) 533–558.

[17] EPISuite program is available from http://www.epa.gov/oppt/exposure/pubs/episuitedl.htm .

[18] L. Tarko**,** Virtual fragmentation of molecules and similarity evaluation, *Rev. Chim.* **55** (2004) 539–546.

[19] C. E. Shannon, A mathematical theory of communication, *The Bell Sys. Tech. J.* **27** (1948) 379–423 and 623–656.

[20] M. A. Efroymson, *Multiple Regression Analysis*, Wiley, New York, 1960.

[21] R. R. Hocking, The analysis and selection of variables in linear regression, *Biometrics* **32** (1976) 1–49.

[22] N. Draper, H. Smith, *Applied Regression Analysis*, Wiley, New York, 1981.

[23] E.I. Eger, L. J. Saidman, B. Brandstater, Minimum alveolar anesthetic concentration: a standard of anesthetic potency, *Anesthesiology* **26** (1965) 756–763.

[24] M. H. Abraham, W. E. Acree, C. Mintz, S. Payne, Effect of anesthetic structure on inhalation anesthesia: implications for the mechanism, *J. Pharm. Sci.* **97** (2008) 2373–2384.

[25] E. I. Eger, M. J. Laster, The effect of rigidity, shape, unsaturation, and length on the anesthetic potency of hydrocarbons, *Anesth. Analg.* **92** (2001) 1477–1482.

[26] A. Won, I. Oh, M. Liao, J. M. Sonner, R. A. Harris, M. J. Laster, R. Brosnan, J. R. Trudell, E. I. Eger, The minimum alveolar anesthetic concentration of 2-,3-, and 4-alcohols and ketones in rats: Relevance to anesthetic mechanisms, *Anesth. Analg.* **102 (**2006) 1419–1426.

[27] E. I. Eger, Y. Xing, M. Laster, J. Sonner, J. F. Antognini, E. Carstens, Halothane and isoflurane have additive minimum alveolar concentration (MAC) effects in rats, *Anesth. Analg*. **96** (2003) 1350–1353.

[28] A. Won, I. Oh, M. J. Laster, J. Popovich, E. I. Eger, J. M. Sonner, Chirality in anesthesia I: Minimum alveolar concentration of secondary alcohol enantiomers, *Anesth. Analg.* **103** (2006) 81–84.

[29] E. I. Eger, M. J. Halsey, D. D. Koblin, M. J. Laster, P. Ionescu, K. Konigsberger, R. Fan, B. V. Nguyen, T. Hudlicky, The convulsant and anesthetic properties of cis-trans isomers of 1,2-dichlorohexafluorocyclo-butane and 1,2-dichloroethylene, *Anesth. Analg*. **93 (**2001) 922–927.

[30] S. M. Dancoff, H. Quastler, *Essays on the Use of Information Theory in Biology*, Univ. Illinois, Urbana, 1953 (see DRAGON software documentation).

[31] M. H. Abraham, A. Ibrahim, A. M. Zissimos, Determination of sets of solute descriptors from chromatographic measurements, *J. Chromat. A* **1037** (2004) 29–47.

[32] A. M. Zissimos, M. H. Abraham, C. M. Du, K. Valko, C. Bevan, D. Reynolds, J. Wood, K. Y. Tam, Calculation of Abraham descriptors from experimental data from seven HPLC systems: Evaluatio of five different methods of calculation, *J. Chem. Soc. Perkin Trans.* **2** (2002) 2001–2010.

[33] A. M. Zissimos, M. H. Abraham, M. C. Barker, K. J. Box, K. Y. Tam, Calculation of the Abraham descriptors for solvent-water partition coefficients in four different

systems: Evaluation of different methods of calculation, *J. Chem. Soc. Perkin. Trans.* **2** (2002) 470–477.

[34] L. H. Hall, T. A. Vaughn, QSAR of Phenol Toxicity using Electrotopological State and Kappa Shape Indices, *Med. Chem. Res.* **7** (1997) 407–416.

[35] K. Roy, G. Ghosh G., Introduction of extended topochemical atom (ETA) indices in the valence electron mobile (VEM) environment as tools for QSAR/QSPR studies, *Int. Elec. J. Mol. Des.* **2** (2003) 599-620.

[36] P. R. Duchowicz, A. G. Mercader, F. M. Fernández, E. A. Castro, Prediction of aqueous toxicity for heterogeneous phenolderivatives by QSAR, *Chemom. Int. Lab. Syst.* **90** (2008) 97–107.

[37] L. B. Kier, L. H. Hall, *Molecular Connectivity in Structure–Activity Analysis*, Wiley, Chichester, 1986 (see DRAGON software documentation).

[38] I. Moriguchi, S. Hirono, I. Nakagome, H. Hirano, Comparison of reliability of logP values for drugs calculated by several methods, *Chem. Pharm. Bull.* **42** (1994) 976–978.

[39] L. Tarko**,** Influence of calibration and validation sets' similarity on the result of external validation test, *MATCH*, *Commun. Math. Comput. Chem.* **75** (2016) 511–532.

[40] R. D. Cramer, R. J. Jilek, S. Guessregen,S. J. Clark, B. Wendt, R. D. Clark, "Lead hopping". Validation of topomer similarity as a superior predictor of similar biological activities, *J. Med. Chem.* **47** (2004) 6777-6791.

[41] J. C. Saeh, P. D. Lyne, B. K. Takasaki, D. A. Cosgrove, Lead hopping using SVM and 3D pharmacophore fingerprints *J. Chem. Inf. Comput. Sci.* **45** (2005) 1122–1133.

[42] D. S. Murrell, I. Cortes–Ciriano, G. J. P. van Westen, I. P. Stott, A. Bender, T. E. Malliavin, R. C. Glen, Chemically aware model builder (camb): an R package for property and bioactivity modelling of small molecules, *J. Cheminform.* **7** (2015) 45–54.

[43] M. Eklund, U. Norinder, S. Boyer, L. Carlsson, Choosing feature selection and learning algorithms in QSAR, *J. Chem. Inf. Model.* **54** (2014) 837-843.

[44] M. Atabati, K. Zarei, A. Borhani, Ant colony optimization as a descriptor selection in QSPR modeling: Estimation of the λ_{max} of anthraquinones-based dyes, *J. Saudi Chem. Soc.* **20** (2016) 547–551.

[45] J. C. Dearden, The history and development of quantitative structure–activity relationships (QSARs), *Int. J. Quant. Struct. Prop. Rel.* **1** (2016) 1–44.

[46] K. Roy, S. Kar, R. N. Das, *Understanding the Basics of QSAR for Applications in Pharmaceutical Sciences and Risk Assessment*, Elsevier, London, 2015.

[47] M. Goodarzi, M. P. Freitas, R. Jensen, Ant colony optimization as a feature selection method in the QSAR modeling of anti-HIV-1 activities of 3-(3,5-dimethylbenzyl)uracil derivatives using MLR, PLS and SVM regressions, *Chemom. Int. Lab. Syst.* **98** (2009) 123–129.