MATCH
Communications in Mathematical and in Computer Chemistry

MATCH Commun. Math. Comput. Chem. **75** (2016) 511-532

ISSN 0340 - 6253

Influence of Calibration and Validation Sets' Similarity on the Result of External Validation Test

Laszlo Tarko

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

E-mail: tarko_laszlo@yahoo.com

(Received September 9, 2015)

Abstract

Formulas and procedures are proposed for the computation of complexity, similarity and diversity of molecules in the analyzed group and of the similarity of two groups (as a whole), based on the Shannon Entropy formula. The article reports certain QSPR studies regarding twenty pairs of calibration/validation sets with high/low complexity and diversity of the included molecules and high/low similarity of sets. A high complexity of structures in the calibration/validation set decreases the quality of the prediction for the validation set. A high similarity of the calibration and validation sets (as a whole) increases the quality of the prediction for the validation set. The diversity of molecules in the calibration/validation set should be directly proportional with the similarity of the calibration and validation sets. If the similarity of the calibration and validation sets is high, a high quality of prediction for the calibration set (*cause*) increases the quality of the prediction for the validation set (*effect*) and the validation test is useful because of this *cause-effect* relation. On the contrary, if the similarity of the calibration and validation sets is low, the influence of the prediction's quality for calibration set on the prediction's quality for the validation set is low and the validation test is useless. The text proposes a formula/criterion for identification of "the best" QSPR/QSAR in the presence of validation/prediction set.

1. Introduction

In *practical* QSPR (*Quantitative Structure-Property Relationship*) studies one uses *calibration set* molecules (having known values of the dependent property) and *prediction set* molecules (having unknown values of the dependent property and not used for QSPR building). The structure of molecules in the prediction set is effect of certain drug design exertion. The goal of this effort is to obtain 'more valuable' new structures. Consequently, the similarity of the structures in calibration set and the structures in prediction set is frequently low enough.

The calibration set is used to identify the QSPR mathematical formula which gives the minimum sum of square differences between the observed and computed values of the dependent property. The best QSPR equation includes a number of computable features of the molecules, named *predictors*, and it is used to compute the value of the dependent property for the molecules in prediction set.

On the contrary, in *academic* publishable QSPR studies, the prediction set, which is not used for model building, includes molecules having known value of activity. Consequently, the computed values can be compared with the observed values. In this case the prediction set is named *validation set* and the comparison is an *external cross-validation test*. The agreement between the observed and the computed values of the dependent property for the molecules in validation set is considered a measure of the QSPR's quality. There are many papers and documents which emphasize the importance of the external validation test [1-20] which is made *after* computation of QSPR equation. The result of the validation test depends on the descriptors used, features of molecules in calibration/validation sets and features of the calibration/validation sets as a whole. Therefore, the reliability of external validation tests is a debatable subject in literature [12b, 21-29].

There are few proposed procedures for extraction of calibration and validation sets from the initial database [28, 30-32]. As a rule, these procedures ensure a high similarity of calibration and validation sets, whilst the similarity is not quantitatively measured. Generally speaking, the similarity of sets seems to emphasize the similarity of structure-property relationship in the calibration set and validation set, respectively. Some authors studied the effect of the composition and similarity ('feature range') of the sets on the quality of prediction in QSPR studies [33, 34].

Besides similarity, another difficult statistical problem is the diversity of the molecules. If the diversity is too low, the statistical methodology cannot identify as

'significant' some descriptors and the best QSPR cannot include, as predictors, these descriptors. On the contrary, if the diversity is too high, the calibration / prediction / validation sets can be non-homogeneous, i.e. they can include few classes, different from the viewpoint of structure-property relationship. The databases used in QSPR studies include frequently molecules similar enough in their chemical structure but not similar from other points of view or molecules different from all points of view.

Some authors emphasized the inability of some equations to correctly estimate the bio-activity for 'completely new chemicals', despite a good result of cross-validation [34]. According to Guidance Documents of the OECD Principles, the quality of certain QSPR must be verified by cross-validation, but any QSPR can be applied only for chemicals included in the Applicability Domain AD [4, 35].

If the QSPR axiom (*similar structures present similar values of the properties*) is true, there are two possibilities:

a) the new molecules, not yet synthesized, imagined by drug design, are inside the AD, their properties can be more or less correctly estimated, but they are more or less similar with the properties of molecules in calibration set and thus they are not interesting

b) the new molecules are outside the AD, their properties cannot be correctly estimated, but are dissimilar with the properties of the calibration set molecules and so they are interesting

Thus, a) + b) seem to highlight an internal contradiction of the QSPR methodology. One can theorize the more interesting molecules are inside AD and close to the borders of AD.

The goal of our paper is to study the result of external validation test as effect of the

- complexity of chemical structure of the analyzed molecules
- homogeneity of molecules in the calibration set from the point of view of structure–property relationship
- diversity of molecules in calibration and validation sets from the point of view of the values of the dependent property
- diversity of molecules in calibration and validation sets from the point of view of molecular features
- similarity of the two sets from the point of view of values of the dependent property and molecular features
- quality of prediction for the molecules in calibration set

2. Methods and formulas

Similarity of two objects

A group including N objects and a criterion K used to compare objects are considered. The values of K for objects are k_1, k_2, … , k_N. There are a maximum value k_{max} and a minimum value k_{min} of K. The similarity S_{ij} of any two *i* and *j* objects in group, from the point of view of K, is computed by the proposed formula (1). The value of S_{ij} is in the range {0, 1}.

$$S_{ij} = 1 - \frac{|k_i - k_j|}{k_{max} - k_{min}} \quad (1)$$

The similarity calculated by formula (1) is 'relative', not 'absolute'. For example, the similarity of the numbers 4 and 7 in group (4, 7, 22), S_{ij} = 0.8333, is much higher than the similarity of the same numbers 4 and 7 in group (4, 7, 9), S_{ij} = 0.4000 and the similarity of the same numbers 4 and 7 in group (4, 5, 7), S_{ij} = 0.0000.

If the values k_1, k_2, … , k_N would be equally distributed in the range {k_{min}, k_{max}} then S_{ij} = (N – 2) / (N – 1). Therefore, the ratio (N – 2) / (N – 1) can be viewed as limit value S_{limit} of S_{ij}. In practice, this value proves suitable, despite its proximity to 1.

Classification of objects in the group

The N objects in the group are included in classes (categories), according to similarity S_{ij}. Each pair of objects in a certain class fulfils the condition (2).

$$S_{ij} \geq S_{limit} \quad (2)$$

Here, we used the next proposed classification procedure including five steps.

Step #1 identification of the first 'seed', i.e. the object having minimum sum of similarities ΣS_{ij} with the other N-1 objects; the first seed is included in the first class

Step #2 identification of the next 'seeds', i.e. objects having similarity (with each seed) *smaller* than S_{limit} and minimum sum of similarities ΣS_{ij} (with the other 'seeds')

Each 'seed' is included in a new class. Step #2 calculations run in a loop until the number of 'seeds' becomes, as a rule, zero. After *n* times running of Step #2 there are *n* + 1 classes, each class includes 1 object and the number of non-classified objects is N – *n* – 1.

Step #3 identification of the object having maximum sum of similarities ΣS_{ij} with the objects included in classes

Step #4 identification of the class having features a) and b)

a) all similarities of included objects with the object identified in Step #3 fulfil the condition (2)
b) greatest mean value of similarities of included objects with the object identified in Step #3

The object identified in Step #3 is the most suitable to be classified. The class identified in Step #4 is the most suitable to include the object identified in Step #3. Step #3 + Step #4 calculations run in a loop. After $N - n - 1$ times running of Step #3 + Step #4 there are $n + 1$ classes also, each class includes few objects and number of the non-classified objects becomes, as a rule, zero. However, sometimes, the last analyzed object remains non-classified, because all classes include one or more object(s) that have a poor similarity (i.e. *smaller* than S_{limit}) similarity to the last object.

Step #5 the non-classified object, if it exists, becomes the last 'seed' of a new (last) class

Diversity of objects in a group

After classification, one calculates the entropy of objects in group using the Shannon equation [36], see formula (3). Here, we used the natural logarithm. The value of entropy SE is in the range {0, log (N)}.

$$SE = -\sum_{i=1}^{k} p_i \cdot \log p_i \qquad (3)$$

where

k is number of classes including minimum one object
$p_i = n_i / N$
n_i is the number of objects in class i ($n_i > 0$)

If all objects are very similar with the all other objects then $k = 1$, $n_i = N$, $p_i = 1$ and SE = 0. On the contrary, if all objects are very non-similar with the all other objects then $k = N$, $n_i = 1$, $p_i = 1/N$ and SE = log (N). Here, SE of objects in the group is considered the measure of

group's complexity. The diversity of objects in group, from the point of view of criterion K, is the ratio SE/ log (N) its value being in the range {0, 1}.

Similarity of two groups of objects

To compute the similarity of two groups G_1 and G_2 of objects, each group considered as a whole, we applied the proposed classification procedure on aggregate $G_1 + G_2$, which includes $N_1 + N_2$ objects. After classification, there are some classes including objects in group G_1 (not G_2), some classes including objects in group G_2 (not G_1) and some classes including objects in group G_1 *and* objects in group G_2. Using the formula (3) one computes the entropy SE_1 of objects of G_1 in aggregate, the entropy SE_2 of objects of G_2 in aggregate and the entropy SE_{12} of aggregate. The diversity of objects in aggregate is SE_{12}/ log ($N_1 + N_2$). The similarity of the two groups is calculated by formula (4).

$$SIM_E = r_1 \cdot r_2 \qquad (4)$$

If $SE_1 < SE_{12}$ then $r_1 = SE_1 / SE_{12}$ else $r_1 = SE_{12} / SE_1$. If $SE_2 < SE_{12}$ then $r_2 = SE_2 / SE_{12}$ else $r_2 = SE_{12} / SE_2$. Consequently, the value of SIM_E is in the range {0, 1}.

Any object can be considered, at a suitable level, a group of objects. For instance, a molecule can be considered a group of atoms, chemical bonds or molecular fragments, a living creature can be considered a group of organs, tissues or cells etc. Therefore, the proposed formulas/algorithm can be used for complexity/diversity/similarity analysis of any (groups of) objects (stars, states, factories, cars, persons, microbes, molecules etc.), compared by a suitable criterion K.

Building of molecules, geometry optimization and calculation of descriptors

Here, the analyzed (groups of) objects are molecules.

The virtual building of the molecules and the geometry's optimization were done using the molecular mechanics program PCModel [37]. The more rigorous geometry's optimization and calculation of some descriptors were performed by MOPAC [38].

Based on the output files created by MOPAC, the PRECLAV software [27, 39] calculated, for each molecule, more than 450 *whole molecule* descriptors, specific to this program. In addition, we used the descriptors calculated by the DRAGON software [40]. PRECLAV was used for the identification of molecular fragments and similarity / diversity / statistical computations.

Identification of molecular fragments

The molecular fragments are identified according to a previously described algorithm [41-43]. Two linked heavy atoms (different from hydrogen) are included within the same fragment (together with the attached hydrogen atoms) if $B > 1.051$, where B is the bond order. For example, the molecule $CH_3–CH_2–O–CH_2–OH$ includes always five fragments, CH_3 + CH_2 + O + CH_2 + OH. The molecule $PhCOOCH_3$ includes, depending on conjugation and B value, two fragments $PhCOO + CH_3$, three fragments $PhCO + O + CH_3$ / $Ph + COO + CH_3$ or four fragments $Ph + CO + O + CH_3$. Similarly, the molecule $CH_3NHCOCH_3$ includes three fragments $CH_3 + NHCO + CH_3$ or four fragments $CH_3 + NH + CO + CH_3$.

Figure 1 presents, as example, the five fragments CH_3, NHCO, O, Cl and C_6H_4 identified in N-methyl-2-chloro-phenyl urethane. The fragment 'Amide' is present.

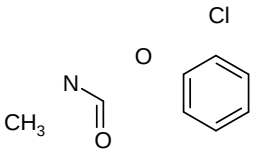

Figure. 1 Five identified molecular fragments in N-methyl-2-chloro-phenyl urethane

Diversity and similarity from the viewpoint of various molecular features

Diversity and similarity calculation from the viewpoint of the dependent property uses the values of the dependent property and formulas (1) – (4). In addition, to calculate the similarity of calibration and validation sets we used the much simpler formula (5), where P_{cal} and P_{val} are the mean values of the dependent property in the calibration and validation set, respectively.

$$SIM_p = 1 - \frac{|P_{cal} - P_{val}|}{\max(P_{cal}, P_{val})} \qquad (5)$$

There are thousands computable molecular features (descriptors). We selected here, as criteria K for comparison of molecules, descriptors for size, hydrophilicity, flexibility, chemical structure and shape.

Diversity and similarity calculation from the viewpoint of molecular size uses the values of COSMO volume [44], computed by MOPAC. In addition, we used the formula (5).

Diversity and similarity calculation from the viewpoint of molecular hydrophilicity uses the values of descriptor AHF (average hydrophilicity of fragments). In addition, we used the formula (5).

The hydrophilicity of a certain molecular fragment is calculated here as the difference between the maximum value S_{max} of hydrogen atoms' net charges and the minimum value S_{min} of heteroatoms' net charges, see formula (6).

$$\Delta = S_{max} - S_{min} \quad (6)$$

If the hydrogen atoms in fragment are missing $S_{max} = 0$. If the heteroatoms in fragment are missing $S_{min} = 0$. Therefore, $\Delta = 0$ for fragments which includes only carbon atoms (C in carbon tetrachloride, C_2 in tetrachloroethylene, C_6 in totally substituted benzene etc.). Using the values of Δ for all fragments one can compute the value of AHF. For example, dodecane and dodecanol present close values of AHF. On the contrary, methanol and dodecanol present quite different values of AHF.

Diversity and similarity calculation from the viewpoint of molecular flexibility uses the percentage values of rotatable bonds [45]. In addition, we used the formula (5).

Calculation of diversity and similarity from the viewpoint of chemical structure and molecular shape is much more difficult because the chemical structure and shape of a certain molecule cannot be defined by an only one real number. Here, the chemical similarity calculation uses the result of molecular fragments' identification [33].

All identified fragments are classified according to criteria #1 and #2. If the number of heavy atoms included in analyzed fragment is 1 the value of criterion #1 is 1. If the number of heavy atoms included is 2 or 3 the value of criterion #1 is 2. If the number of heavy atoms included is > 3 the value of criterion #1 is 3. Criterion #2 is the string of symbols of included elements, in alphabetical order. If the value of criterion #1 *and* criterion #2 is the same the analyzed fragments are considered *in the same class*.

Then, the Shannon Entropy of classes in the analyzed molecule is computed with formula (3), where p_i are the percentages in weight of classes. The similarity of two molecules, considered two groups of classes' fragments, is calculated by formula (4). The classification of molecules, from the point of view of chemical structure, uses, in formula (2), the value $S_{limit} = 0.9$.

The classification of molecules, from the point of view of molecular shape, uses, in formula (2), the similarity S_{ij} calculated by *Ultrafast Shape Recognition* method [46, 47] and the value $S_{limit} = 0.9$.

Databases used

The ten databases used in our computations are presented in Table 1. The number of molecules in each database is in the range {49, 112}.

To avoid arithmetical problems, the references' values of the dependent properties are placed in the positive range, see the column no. 4 in table. For instance, Celsius degrees were replaced by Kelvin degrees. Other times all values V were replaced by corrected values $V_{cor} = V-V_{min}$.

Table 1 Analyzed databases

Database	**Chemical class**	**Dependent property**	**Range**	**Ref.**
1	Anthranilic acids	Anti-inflammatory activity	{1.000, 4.125}	[48]
2	Cyclic ureas	log(1/K)	{5.30, 11.01}	[49]
3	Flavones	logK	{0.000, 5.301}	[50]
4	Fluoro-alkans	Boiling Point	{195.2, 383.2}	[51]
5	Guanidines	Sweetness power	{0.000, 2.768}	[52]
6	HEPTs	Anti-HIV activity	{3.66, 8.57}	[53]
7	PAHs	PADA*	{0.7, 50.0}	[54]
8	Phenols	Toxicity	{0.000, 2.638}	[55]
9	Quinolines	Anti-HIV activity	{3.46, 6.70}	[56]
10	Urethanes	Toxicity	{2.30, 7.48}	[57]

*Percent of Applied dose Dermally Absorbed over 24 hours

Selection of calibration and validation sets

The molecules in each database in Table 1 were ordered according to the values of the dependent property. Then, each database was used to make a) and b) type pairs of calibration/validation sets:

a) calibration set and validation set are presumed to be 'similar enough'

b) calibration set and validation set are presumed to be 'non-similar'

To obtain type a) pair we extracted, as validation set, the ordered molecules having ranks 2, 4, 6, etc. or 2, 5, 8, etc. or 2, 4, 7, 9, 12, 14, 17 etc., depending on desired percentage, in the range {30, 50}, of validation set within database.

To obtain type b) pair we extracted, as validation set, the molecules having greatest / smallest values of dependent property.

Each pair calibration set/validation set is used in a distinct diversity/similarity computation and distinct QSPR study. Table 2 presents the number of molecules in calibration/validation sets and the similarity value SIM_P calculated with formula (5).

Table 2 Number of molecules and similarity of calibration/validation sets

Database	Study	N_{cal}	N_{val}	SIM_P	Database	Study	N_{cal}	N_{val}	SIM_P
1	I	56	56	0.989	**6**	XI	40	40	0.989
	II	79	33	0.685		XII	56	24	0.691
2	III	47	32	0.991	**7**	XIII	36	24	0.972
	IV	40	39	0.780		XIV	30	30	0.447
3	V	39	39	0.970	**8**	XV	34	16	0.948
	VI	54	24	0.467		XVI	30	20	0.426
4	VII	52	25	0.980	**9**	XVII	35	15	0.992
	VIII	46	31	0.846		XVIII	30	20	0.742
5	IX	34	15	0.997	**10**	XIX	46	30	0.992
	X	29	20	0.570		XX	38	38	0.700

In QSPR studies I, III, V, VII etc., the calibration and validation sets are presumed to be 'similar enough', at least from the point of view of values of dependent property. In QSPR studies II, IV, VI, VIII etc., the calibration and validation sets are presumed to be 'non-similar'.

Used statistical methods

We used in all twenty QSPR studies:

- the same initial set of descriptors, i.e. DRAGON descriptors, PRECLAV *whole molecule* descriptors and PRECLAV descriptors of molecular fragments; aromaticity descriptors [42] and 3D descriptors were not used
- the same QSPR algorithm, i.e. the selection of significant descriptors and equations by formulas and *forward stepwise* procedure of PRECLAV
- the 'best' QSPR for prediction, i.e. the equation having maximum value of quality *Q*, see formula (7), obtained with calibration set *including outliers*

$$Q = r^2 \cdot \left(1 - \frac{p}{N}\right) \tag{7}$$

where

r^2 is square linear correlation of observed/computed values
p is number of predictors
N is number of molecules in calibration set

The value of the function *Q* is in the range {0, 1}. According to used *forward stepwise* procedure, the value of *p* increases in the range {2, $\log_2$ (N)} and the value of *Q* increases, reaches a maximum, then decreases. Therefore, the value of $\log_2$ (N), rounded off to integer, and the value of *Q* are criteria to stop the calculation.

Other criteria for quality of obtained QSPR are r^2 in formula (7), being in the range {0, 1} and the Fisher function, see formula (8), being in the range {0, ∞}.

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

The relative utility of a certain predictor is computed by formula (9).

$$U = \frac{R^2 - r^2}{1-r^2} \qquad (9)$$

where

R^2 is the square correlation between the observed/computed values of the dependent property (using the QSPR with all *p* predictors)

r^2 is the square correlation between the observed/computed values of the dependent property (using the QSPR with *p-1* predictors, i.e. the equation *without* the analyzed predictor)

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

PRECLAV calculates the square of cross-validated linear correlation r^2_{CV} using the *Leave Half Out* method. However, this usual method is applied after ordering of molecules in the calibration set using the observed values of the dependent property. Consequently, the function r^2_{CV} is viewed here as a measure of the homogeneity of the calibration set from the viewpoint of structure-activity relationship, not as a result of a quite drastic internal validation test. If $r^2_{CV} > 0.4$ the calibration set can be considered 'homogeneous enough'.

The outlier index O of a certain molecule is the usual ratio $|P_{obs} - P_{calc}| / SEE$, where P_{obs} and P_{calc} are the observed and calculated values of the dependent property and SEE is the standard error of estimation. If the quality of the prediction for the calibration set is high the value of the difference $|P_{obs} - P_{calc}|$ is small and the value of SEE is also small. Thus, the value of ratio O can be great or small, regardless of the quality of prediction. Here, other measures of non-homogeneity (diversity) of molecules in calibration set, from the viewpoint

of structure-activity relationship, are the four sums of outlier indices ΣO_i (O > 1.5, O > 2.0, O > 2.5 and O > 3.0).

The quality of the prediction for calibration set molecules is measured by r^2_{cal}, *Q* and *F*. The quality of prediction for molecules in validation set is measured by r^2_{val}. Finally, the diversities, the similarities, various combinations of diversities / similarities (sums, products and ratios), r^2_{cal}, *F*, *Q* and r^2_{CV} (as a group including over 100 descriptors) and r^2_{val} (as dependent property) are used to obtain the equation of r^2_{val}. In this last statistical computation the selection of descriptors and equations is made by the same PRECLAV formulas and *forward stepwise* procedure, identification of 'outlier databases' included.

3. Results and discussions

The calculated values D_{cal} of diversity in the calibration sets, D_{val} of diversity in the validation sets and the similarities SIM_E and SIM_P of calibration set/validation set, from the point of view of six criteria, for all twenty QSPR studies, are presented in Table A in Supplementary Data appendix [58].

The highest correlation between diversity and similarity SIM_E is for D_{val} (hydrophilicity, $r^2 = 0.3295$). The highest correlation between diversity and similarity SIM_P is for D_{cal} (size, $r^2 = 0.1544$). Consequently, there is a low correlation between diversity descriptors and similarity descriptors.

Correlation between SIM_E and SIM_P is low: $r^2 = 0.7044$ (dependent property), $r^2 = 0.2248$ (size), $r^2 = 0.2642$ (hydrophilicity), $r^2 = 0.2545$ (flexibility). Hence, the formulas (4) and (5) describe in different manner the similarity of calibration and validation sets.

Correlation of similarity of calibration and validation sets, from the point of view of molecular features, with similarity SIM_E of dependent property is low: $r^2 = 0.5783$ (size), $r^2 = 0.2518$ (hydrophilicity), $r^2 = 0.4129$ (flexibility), $r^2 = 0.4185$ (chemical structure), $r^2 = 0.1746$ (shape). Therefore, the QSPR axiom seems to be fulfilled to a small extent.

If the diversity and similarity, from the point of view of the criterion K, are smaller than an empirically established value, the QSPR calculation in presence of prediction/validation set should be, as a rule, avoided, because of the inability of the calculated QSPR to make a reliable prediction for the prediction/validation set, regardless of good prediction for calibration set. However, this 'avoidance' should be applied only if, from the point of view of criterion K, the observance of QSPR axiom is well marked.

Correlation of the similarity of the calibration and validation sets, from the point of view of chemical structure, with similarity from the point of view of hydrophilicity (r^2 = 0.1455) and flexibility (r^2 = 0.2060) is low. Consequently, as expected, the type and percentages of molecular fragments are not enough to describe the molecular hydrophilicity and flexibility.

Supplementary Data includes the results of QSPR studies, i.e. the formulas of multilinear QSPRs I – XX and Tables I – XX of observed/computed values of dependent properties, rounded off to two decimals. In seven from twenty QSPRs the number of predictors is smaller than the maximum allowed value $\log_2$ (N).

Table 3 presents the correlations in each QSPR study.

Table 3 The correlations in QSPR studies

Study	$\mathbf{r^2_{cal}}$	**F**	**Q**	$\mathbf{r^2_{CV}}$	$\mathbf{r^2_{val}}$
I	0.8022	33.8	0.7162	0.5611	0.3191
II	0.7478	36.1	0.6910	0.5265	0.5322
III	0.9018	62.8	0.7867	0.6492	0.5437
IV	0.8569	41.9	0.7498	0.4407	0.0256
V	0.8827	51.2	0.7696	0.3647	0.1413
VI	0.8644	51.0	0.7683	0.1182	0.1127
VII	0.9591	281.3	0.8853	0.4607	0.9411
VIII	0.9269	133.2	0.8463	0.5330	0.4085
IX	0.9319	79.4	0.7949	0.5543	0.2410
X	0.8572	28.8	0.7094	0.2666	0.2462
XI	0.9396	140.1	0.8457	0.4865	0.3917
XII	0.9227	121.8	0.8403	0.6033	0.0050
XIII	0.8939	52.2	0.7697	0.2693	0.5131
XIV	0.8969	56.6	0.7773	0.5813	0.3288
XV	0.9037	97.0	0.8240	0.3804	0.9327
XVI	0.9479	118.3	0.8215	0.6141	0.3512
XVII	0.8256	28.4	0.7076	0.5281	0.3983
XVIII	0.8218	23.1	0.6848	0.2208	0.0029
XIX	0.9166	73.3	0.7971	0.4851	0.6112
XX	0.8605	40.7	0.7472	0.1969	0.1963

According to the values of r^2_{cal}, *F* and *Q* in Table 3, the prediction for calibration set molecules is good enough. The initial set of descriptors, as basis for selection of predictors and QSPRs, seems to be suitable.

The correlation of r^2_{val} with r^2_{cal} is positive, but very low (r = 0.2647). The correlation of r^2_{val} with *F* (r = 0.5198) and *Q* (r = 0.4162) is also low. There is no strong cause-effect relation between the quality of prediction for calibration set molecules and the quality of

prediction for validation set molecules. Actually, the value of r^2_{cal} is within narrow range {0.7478, 0.9591} and the value of r^2_{val} is within wide range {0.0029, 0.9411}.

The descriptor having greatest positive correlation (r = 0.5325) with r^2_{val} is the similarity of calibration/validation sets from the point of view of the dependent property, calculated using formula (4). The descriptor having the greatest negative correlation (r = -0.2377) with r^2_{val} is the diversity of molecules in validation set regarding the chemical structure.

Above comments regarding correlations with r^2_{val} are not very significant because 'the best' equation for description of r^2_{val} can include predictors having a low enough correlation with r^2_{val} *because of* their low intercorrelation. Moreover, the group of twenty pairs of calibration set/validation set in Table 2 can include, as a rule, some outlier pairs.

Using all twenty pairs calibration set/validation set we have not identified any outlier pair and we obtained the equation (10). This statistical study seems to be a model of *practical* QSPR studies, because the similarity of the calibration and prediction sets is not very high.

$$r^2_{val} = -1.0637 + 0.0834 \cdot D_1 - 0.2791 \cdot D_2 + 2.2581 \cdot D_3 \qquad (10)$$

where

D_1 is the sum of outlier indices (O > 2.5) (U = 1000)
D_2 is minimum complexity (Shannon Entropy of masses of molecular fragments) of the molecules in calibration set (U = 904)
D_3 is product Q • S, where Q is the quality (by r^2) of prediction for calibration set and S is similarity of *estimated* values of dependent property for calibration and validation sets (U = 871)

There is a good match between the values of r^2_{val} in Table 3 and the values calculated by formula (10): r^2 = 0.8091; F = 24.0; Q = 0.6877. We point out the direct dependency of r^2_{val} on the diversity of calibration set molecules, on the quality of the prediction for the calibration set molecules and on the similarity of estimated values of the dependent property for the calibration and validation sets. A high complexity of chemical structures in the calibration set decreases the value of r^2_{val}.

As a rule, the best QSPR/QSAR according to formula (7) and the best QSPR/QSAR according to formula (10) are different, because of frequently low similarity of the calibration and validation/prediction sets. Accordingly, the formula (7) can be used as a criterion for the identification of "the best" QSPR/QSAR in absence of the validation/prediction set and the formula (10) can be used as a criterion for identification of "the best" QSPR/QSAR in presence of validation/prediction set. This is a new viewpoint regarding the prediction method

in presence of validation/prediction set. In present, the published QSPR/QSAR studies use, in prediction for calibration and validation/prediction sets, the same formula.

See the scatter-plot of equation (10) in Figure 2.

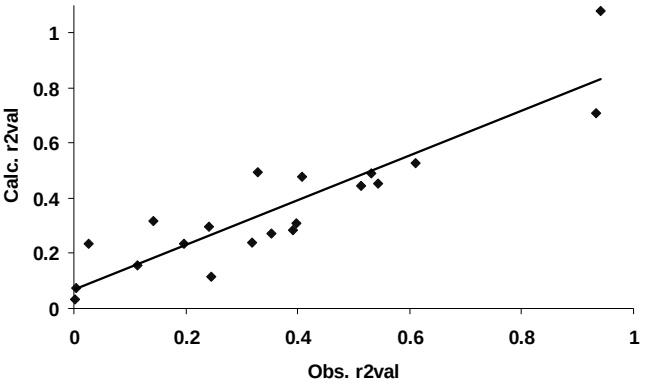

Figure 2 Scatter-plot of equation (10)

Using only the pairs calibration set/validation set I, III, V, VII etc., presumed to be 'similar enough', we have not identified any outlier pair and we obtained the equation (11). This statistical study seems to be a model of *academic* QSPR studies, because the similarity of calibration and prediction sets is high, as in the validation tests.

$$r^2_{val} = 2.8255 - 5.9615 \cdot D_1 + 3.0653 \cdot D_2 + 0.8052 \cdot D_3 \qquad (11)$$

where

D_1 is the ratio D_{cal}/D_{val} of the diversities of the molecules in calibration/validation sets from the point of view of shape (U = 1000)

D_2 is the diversity of molecules in calibration set from the point of view of hydrophilicity (U = 911)

D_3 is sum Q + S, where Q is the quality of prediction for the calibration set, by formula (7), and S is similarity of *estimated* values of the dependent property for the calibration and validation sets (U = 612)

The match between the values of r^2_{val} in Table 3 and the values calculated by formula (11) is very good, at least from the view point of r^2 and F: $r^2 = 0.9682$; F = 71.1; Q = 0.6777. The predictor D_3 has the smallest 'relative utility' in description of r^2_{val}.

Using only the pairs II, IV, VI, VIII etc., presumed to be 'non-similar', we have not identified any outlier pair and obtained the equation (12). Maybe, this statistical study is a

model of QSPR calculation for outside AD molecules, because of the low similarity of the calibration and prediction sets.

$$r^2_{val} = -1.2720 - 0.3356 \cdot D_1 + 2.9379 \cdot D_2 \qquad (12)$$

where

D_1 is maximum complexity (Shannon Entropy of topologic distances) of molecules in validation set (U = 1000)

D_2 is the similarity of molecules in the calibration/validation sets from the point of view of hydrophilicity (U = 931)

The match between the values of r^2_{val} in Table 3 and the values calculated by formula (12) is good: $r^2 = 0.8948$; F = 34.0; Q = 0.7158.

Table 4 includes the values of r^2_{val} in Table 3 and the values of r^2_{val} calculated by equations (10), (11) and (12). One observes very low quality of prediction of equation (11) for pairs II, IV, VI, VIII etc., and a very low quality of prediction of equation (12) for pairs I, III, V, VII etc. The equation which describes well enough the 'similar' pairs calibration/validation set cannot offer a good description for 'non-similar' pairs and vice versa.

Table 4 The values of r^2_{val}

Pair	**r^2_{val} in Table 3**	**Calc**		
		Eq. (10)	**Eq. (11)**	**Eq. (12)**
I	0.3191	0.240	0.344	0.501
II	0.5322	0.491	0.583	0.489
III	0.5437	0.455	0.480	0.117
IV	0.0256	0.234	0.455	0.091
V	0.1413	0.316	0.180	0.438
VI	0.1127	0.155	0.186	0.142
VII	0.9411	1.080	1.019	0.497
VIII	0.4085	0.479	1.816	0.494
IX	0.2410	0.297	0.297	0.048
X	0.2462	0.114	0.491	0.193
XI	0.3917	0.286	0.354	0.365
XII	0.0050	0.074	0.951	- 0.017
XIII	0.5131	0.443	0.478	0.584
XIV	0.3288	0.494	- 0.093	0.348
XV	0.9327	0.709	0.887	0.368
XVI	0.3512	0.271	0.433	0.253
XVII	0.3983	0.308	0.403	- 0.004
XVIII	0.0029	0.034	0.302	0.058
XIX	0.6112	0.529	0.591	0.330
XX	0.1963	0.234	0.437	0.160

The algebraic sign of predictors in equations (10), (11) and (12) highlights the influence (direct or inverse) of the complexity, diversity, similarity and predictive power for calibration set on the quality r^2_{val} of prediction for validation set, see Table 5.

Table 5 Influence of the complexity, diversity, similarity and predictive power

Influence of	**Similarity of calibration and validation sets (by dependent property)**		
	low (eq. 12)	**medium (eq. 10)**	**high (eq. 11)**
complexity of molecules in calibration set	not selected as predictor	inverse	not selected as predictor
complexity of molecules in validation set	inverse	not selected as predictor	not selected as predictor
diversity of molecules in calibration set	not selected as predictor	direct	not selected as predictor
diversity of molecules in validation set	not selected as predictor	not selected as predictor	direct
similarity of the calibration and validation sets (by molecular features)	direct	direct	direct
quality of the prediction for the calibration set	not selected as predictor	direct	direct

The calculation of diversity and similarity, from the point of view of molecular features, can use molecules having unknown value of the dependent property. In addition, unlike the calculation of the Applicability Domain, the calculation of diversity and similarity is made *before* QSPR calculation. The calculation of diversity and similarity requires less time than the computation of QSPR(s).

4. Conclusions

The influence of the size of the database and of the validation set on the quality of prediction for the validation set is low (i.e. the number and percentage of molecules are not selected as predictors for r^2_{val}).

A high complexity of structures in calibration/validation set decreases the quality of prediction for validation set.

A high similarity of calibration and validation sets (as a whole) increases the quality of prediction for validation set.

In selection of the validation set the diversity of molecules in calibration/validation set should be directly proportional with the similarity of the calibration and validation sets.

If the similarity of the calibration and validation sets is high, a high quality of prediction for the calibration set (*cause*) increases the quality of prediction for the validation set (*effect*) and the validation test is useful because of this cause-effect relation.

If the similarity of the calibration and validation sets is low, the influence of the prediction's quality for calibration set on the prediction's quality for validation set is low and the validation test is useless.

The newly proposed formula can be used as criterion for identification of "the best" QSPR/QSAR in presence of the validation/prediction set.

References

[1] D. W. Osten, Selection of optimal regression models via cross–validation, *J. Chemom.* **2** (1988) 39–48.

[2] L. Eriksson, J. Jaworska, A.P. Worth, M. T. D. Cronin, R. M. McDowell, P. Gramatica, Methods for reliability and uncertainty assessment and for applicability evaluations of classification– and regression–based QSARs, *Environ. Health Perspect* **111** (2003) 1361–1375.

[3] A. Tropsha, P. Gramatica, V. K. Gombar, The importance of being earnest: validation is the absolute essential for successful application and interpretation of QSPR models, *QSAR Comb. Sci.* **22** (2003) 69–77.

[4] OECD, Document ENV/JM/ MONO (2007). 2

[5] S. Weaver, M. P. Gleeson, The importance of the domain of applicability in QSAR modeling, *J. Mol. Graph. Model.* **26** (2008) 1315–1326.

[6] P. K. Ojha, I. Mitra, R. N. Das, K. Roy, Further exploring r_m^2 metrics for validation of QSPR models, *Chemom. Intell. Lab. Syst.* **107** (2011) 194–205.

[7] K. Roy, I. Mitra, S. Kar, P. K. Ojha, R. N. Das, H. Kabir, Comparative studies on some metrics for external validation of QSPR models, *J. Chem. Inf. Model.* **52** (2012) 396–408.

[8] K. Roy, P. Chakraborty, I. Mitra, P. K. Ojha, Kar S, R. N. Das, Some case studies on application of “rm2” metrics for judging quality of quantitative structure–activity relationship predictions: Emphasis on scaling of response data, *J. Comp. Chem.* **34** (2013) 1071–1082.

[9] H. Kubinyi, A. H. Fred, T. Mietzner, Three–dimensional quantitative similarity−activity relationships (3D QSiAR) from SEAL similarity matrices, *J. Med. Chem.* **41** (1998) 2553–2564.

[10] H. Kubinyi, From narcosis to hyperspace: The history of QSAR, *Quant. Struct. Act. Relat.* **21** (2002) 348–356.

[11] D. M. Hawkins, The problem of overfitting, *J. Chem. Inf. Comput. Sci.* **44** (2004) 1–12.

[12] a) K. Baumann, N. Stiefl, Validation tools for variable subset regression, *J. Comput. Aid. Mol. Design* **18** (2004) 549–562.

b) D. Baumann**,** K. Baumann, Reliable estimation of prediction errors for QSAR models under model uncertainty using double cross-validation, *J. Cheminf.* **6** (2014) 47.

[13] A. Tropsha, A. Golbraikh, Predictive QSAR modeling workflow, model applicability domains, and virtual screening, *Curr. Pharm. Des.* **13** (2007) 3494–3504.

[14] A. Tropsha, Best practices for QSAR model development, validation, and exploitation, *Mol. Inf.* **29** (2010) 476–488.

[15] K. H. Esbensen, P. Geladi, Principles of proper validation: use and abuse of re-sampling for validation, *J. Chemom.* **24** (2010) 168–187.

[16] N. Chirico, P. Gramatica, Real external predictability of QSAR models: how to evaluate it? Comparison of different validation criteria and proposal of using the concordance correlation coefficient, *J. Chem. Inf. Model.* **51** (2011) 2320–2335.

[17] T. M. Martin, P. Harten, D. M. Young, E. N. Muratov, A. Golbraikh, H. Zhu, A. Tropsha, Does rational selection of training sets improve the outcome of QSAR modeling? *J. Chem. Inf. Model.* **52** (2012) 2570–2578.

[18] N. Chirico, P. Gramatica, Real external predictivity of QSAR models. Part 2. New intercomparable thresholds for different validation criteria and the need for scatter plot inspection, *J. Chem. Inf. Model.* **52** (2012) 2044–2058.

[19] P. Gramatica, S. Cassani, P. P. Roy, S. Kovarich, C. W. Yap, E. Papa, QSAR modeling is not “push a button and find a correlation”: A case study of toxicity of (benzo-)triazoles on algae, *Mol. Inf.* **31** (2012) 817–835.

[20] R. Kiralj, M. M. C. Ferreira, Basic validation procedures for regression models in QSAR and QSPR studies: Theory and application, *J. Braz. Chem. Soc.* **20** (2009) 770–787.

[21] P. Gramatica, P. Pilutti, E. Papa, Ranking of volatile organic compounds for tropospheric degradability by oxidants: a QSPR approach, *SAR QSAR Environ. Res.* **13** (2002) 743–753.

[22] P. Gramatica, P. Pilutti, E. Papa, QSAR prediction of ozone tropospheric degradation, *QSAR Comb. Sci.* **22** (2003) 364–373.

[23] P. Gramatica, Principles of QSAR models validation: internal and external, *QSAR Comb. Sci.* **26** (2007) 694–701.

[24] D. M. Hawkins, S. C. Basak, D. Mills, Assessing model fit by cross–validation, *J. Chem. Inf. Comp. Sci.* **43** (2003) 579–586.

[25] C. Helma, Data mining and knowledge discovery in predictive toxicology, *SAR QSAR Environ. Res.* **15** (2004) 367–383.

[26] P. P. Roy, S. Paul, I. Mitra, K. Roy, On two novel parameters for validation of predictive QSAR models, *Molecules* **14** (2009) 1660–1701.

[27] 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.

[28] a) A. Golbraikh, A. Tropsha, Beware of q2!, *J. Mol. Graphics Modell.* **20** (2002) 269–276.

b) A. Golbraikh, M. Shen, Z. Xiao, Y.-D. Xiao, K.-H. Lee, A. Tropsha, Rational selection of training and test sets for the development of validated QSAR models, *J. Comp. Aid. Mol. Des.* **17** (2003) 241–253.

[29] L. Tarko, QSPR for LogP using molecules without oxygen and nitrogen and quantum mechanics descriptors, *J. Iran. Chem. Soc.* **12** (2015) 359–365.

[30] J. T. Leonard, K. Roy, On selection of training and test sets for the development of predictive QSAR models, *QSAR Comb. Sci.* **25** (2006) 235–251.

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

[32] G. Xu, J. M. Hughes–Oliver, J. D. Brooks, J. L. Yeatts, R. E. Baynes, Selection of appropriate training and validation set chemicals for modelling dermal permeability by U- optimal design, *SAR QSAR Environ. Res.* **24** (2013) 135–156.

[33] L. Tarko, The effect of the diversity of molecules in sets and similarity of sets on the quality of prediction in QSAR studies, *J. Math. Chem.* **52** (2014) 948–965.

[34] P. Gramatica, External evaluation of QSAR models, in addition to cross-validation: verification of predictive capability on totally new chemicals, *Mol. Inf.* **33** (2014) 311–314.

[35] I. V. Tetko, I. Sushko, A. K. Pandey, H. Zhu, A. Tropsha, E. Papa, T. Oberg, R. Todeschini, D. Fourches, A. Varnek, Critical assessment of QSAR models of environmental toxicity against Tetrahymena pyriformis: focusing on applicability domain and overfitting by variable selection, *J. Chem. Inf. Model.* **48** (2008) 1733–1746.

[36] C. E. Shannon, A mathematical theory of communication, *Bell Syst. Techn. J.* **27** (1948) 379–423.

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

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

[39] in-house PRECLAV program, documentation included, is available from Center of Organic Chemistry – Bucharest – Romanian Academy; ltarko@cco.ro; tarko_laszlo@yahoo.com

[40] DRAGON program is available from Talete srl., via V. Pisani, 13-20124, Milano, Italy; http://www.talete.mi.it, Nov 2014.

[41] L. Tarko, Virtual fragmentation of molecules and similarity evaluation, *Rev. Chim. (Bucuresti)* **55** (2004) 539–544.

[42] L. Tarko, A procedure for virtual fragmentation of molecules into functional groups, *ARKIVOC* **xiv** (2004) 74–82.

[43] L. Tarko, Aromatic molecular zones and fragments, *ARKIVOC* **xi** (2008) 24–45.

[44] A. Klamt, G. Schümann, COSMO: a new approach to dielectric screening in solvents with explicit expressions for the screening energy and its gradient, *J. Chem. Soc. Perkin Trans.* **2** (1993) 799–805.

[45] L. Tarko, Using the bond order calculated by quantum mechanics to identify the rotatable bonds, *Rev. Chim. (Bucureşti)* **62** (2011) 135–138.

[46] P. J. Ballester, W. G. Richards, Ultrafast shape recognition for similarity search in molecular databases, *Proc. R. Soc. A* **463** (2007) 1307–1321.

[47] E. O. Cannon**,** F. Nigsch**,** J. B. O. Mitchell, A novel hybrid ultrafast shape descriptor method for use in virtual screening, *Chem. Cent. J.* **2** (2008) 3.

[48] C. Menye, C. M. Ngabireng, S. F. Kouam, Prediction of anti-inflammatory activity of anthranylic acids using structural molecular fragment and topochemical models, *Der Pharm. Lett.* **5** (2013) 88–98.

[49] S. Avram, C. Bologa, M. L. Flonta, Quantitative structure–activity relationship by CoMFA for cyclic urea and nonpeptide-cyclic cyanoguanidine derivatives on wild type and mutant HIV-1 proptease, *J. Mol. Model.* **11** (2005) 105–115.

[50] P. R. Duchowicz, M. G. Vitale, E. A. Castro, J. C. Autino, G. P. Romanelli, D. O. Bennardi, QSAR Modeling of the interactions of flavonoids with GABA(A) receptor, *Eur. J. Med. Chem.* **43** (2008) 1593–1602.

[51] C. Rucker, M. Meringer, A. Kerber, QSPR using MOLGEN-QSPR: the challenge of fluoroalkane boiling points, *J. Chem. Inf. Model.* **45** (2005) 74–80.

[52] J. S. Barker, C. K. Hattotuwagama, M. G. B. Drew, Computational studies of sweet-tasting molecules, *Pure Appl.Chem.* **74** (2002) 1207–1217.

[53] M. J. Heravi, F. Parastar, Use of artificial neural networks in a QSAR study of anti-HIV activity for a large group of HEPT derivatives, *J. Chem. Inf. Comput. Sci.* **40** (2000) 147–154.

[54] T. A. Roy, A. J. Krueger, C. R. Makerer, W. Neil, A. M. Arroyo, J. J. Yang, SAR models for estimating the percutaneous absorption of polynuclear aromatic hydrocarbons, *SAR QSAR Environ. Res.* **9** (1998) 171–185.

[55] L. H. Hall, T. A. Vaughn, QSAR of Phenol Toxicity using electrotopological state and kappa shape indices, *Med. Chem. Res.* **7** (1997) 407–416.

[56] X. H. Sun, J. Q. Guan, J. J. Tan, 3D-QSAR studies of quinoline ring derivatives as HIV-1 integrase inhibitors, *SAR QSAR Environ. Res.* **23** (2012) 683–703.

[57] R. L. Metcalf, T. R. Fukuto, Effects of molecular structure upon anticholinesterase and insecticidal activity of substituted phenyl N-methylcarbamates, *J. Agr. Food. Chem.* **15** (1967) 1022–1029.

[58] *Supplementary Data* are available from author by e-mail or from http://www.cco.ro/articles/tarko/TarkoSupplementaryDataMATCH.pdf