Journal of Computer-Aided Molecular Design, **16:** 357–369, 2002.
KLUWER/ESCOM

Predictive QSAR modeling based on diversity sampling of experimental datasets for the training and test set selection

Alexander Golbraikh and Alexander Tropsha*
The Laboratory for Molecular Modeling, School of Pharmacy, University of North Carolina, Chapel Hill, NC 27599-7360.

Summary

One of the most important characteristics of Quantitative Structure Activity Relashionships (QSAR) models is their predictive power. The latter can be defined as the ability of a model to predict accurately the target property (e.g., biological activity) of compounds that were not used for model development. We suggest that this goal can be achieved by rational division of an experimental SAR dataset into the training and test set, which are used for model development and validation, respectively. Given that all compounds are represented by points in multidimensional descriptor space, we argue that training and test sets must satisfy the following criteria: (i) Representative points of the test set must be close to those of the training set; (ii) Representative points of the training set must be close to representative points of the test set; (iii) Training set must be diverse. For quantitative description of these criteria, we use molecular dataset diversity indices introduced recently (Golbraikh, A., J. Chem. Inf. Comput. Sci., 40 (2000) 414–425). For rational division of a dataset into the training and test sets, we use three closely related sphere-exclusion algorithms. Using several experimental datasets, we demonstrate that QSAR models built and validated with our approach have statistically better predictive power than models generated with either random or activity ranking based selection of the training and test sets. We suggest that rational approaches to the selection of training and test sets based on diversity principles should be used routinely in all QSAR modeling research.

Introduction

A brief overview of QSAR modeling

Quantitative Structure-Activity Relationships (QSAR) analysis can be defined as application of mathematical and statistical methods to the problem of finding empirical equations (QSAR models) of the form $Y_i = F_i(X_1,X_2,\ldots,X_n)$, where Y_i are biological activities of molecules, and X_1, X_2,..., X_n are experimental or calculated structural properties (molecular descriptors) of compounds. Each compound can be represented by a point in a multidimensional space, in which descriptors X_1, X_2,...,X_n are independent coordinates of the compound. The goal of QSAR modeling is to establish a trend in descriptor values which correlate with a trend in biological activity. All QSAR approaches implement, directly or indirectly, a simple similarity principle, which for a long time has provided a foundation for the experimental medicinal chemistry: compounds with similar structures are expected to have similar biological activities. This implies that points representing compounds with similar activities in multidimentional descriptor space should be close to each other.

Based on the origin of molecular descriptors used in calculations, QSAR methods can be divided into three groups. One group is based on a relatively small number (usually many times smaller than the number of compounds in a dataset) of physicochemical properties and parameters describing hydrophobic, steric, electrostatic, etc. effects. Usually, these descriptors are used as independent variables in multiple regression approaches [1]. In the literature, these methods are referred to as Hansch analysis [2]. Other methods are based on quantitative characteristics of molecular graphs (molecular topological descriptors). They include molecular connectivity indices [3–5], molec-

*To whom correspondence should be addressed.

ular shape indices [6,7], topological [8], and electrotopological state indices [9–12], atom-pair descriptors [13, 14], etc. Sometimes topological descriptors are also combined with physical-chemical properties of molecules. Since molecular graphs or structural formulas are 'two-dimensional', these methods are referred to as two-dimensional (2D) QSAR. Different correlation methods are used to develop 2D QSAR models. They include linear (e.g., multiple linear regression (MLR) with variable selection [15], partial least squares (PLS) [16], etc.) as well as non-linear (e.g., *k*-Nearest Neighbors (*k*NN) [17, 18], artificial neural networks [19], etc.) methods. The third group of methods is based on using descriptors derived from spatial (three-dimensional) representation of molecular structures. Correspondingly, these methods are referred to as three-dimensional or 3D-QSAR. Most of 3D-QSAR methods, require 3D alignment of all molecules according to a pharmacophore model or based on docking to a ligand binding site of a receptor. Descriptors in this case (as in Comparative Molecular field Analysis, CoMFA [20, 21], and CoMFA-like methods [22–25]) could be electrostatic, steric, hydrophobic etc. field values in grid points surrounding the molecules.

Model validation: current approaches to estimate the predictive power of QSAR models

It has been shown that the more independent variables is involved in MLR QSAR analysis, the higher is the probability of a chance correlation between predicted and observed activities, even if only a small portion of variables is included in the final QSAR equation [26]. This conclusion is true not only for MLR QSAR, but also for any QSAR approach when the number of variables (descriptors) is comparable to or higher than the number of compounds in a dataset. Thus, model validation is one of the most important aspects of QSAR analysis.

To validate a QSAR model, most of researchers apply the leave-one-out (LOO) or leave-some-out (LSO) cross-validation procedures. The outcome from this procedure is a cross-validated correlation coefficient R^2 (q^2), which is calculated according to the formula

$$q^2 = 1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y})^2}, \quad (1)$$

where y_i, $\hat{y}_i$, and $\bar{y}$ are the actual, estimated, and averaged (over the entire dataset) activities, respectively. The summations in (1) are performed over all compounds, which are used to build a model for the training set. Frequently, q^2 is used as a criterion of both robustness and predictive ability of the model. Many authors consider high q^2 (for instance, $q^2 > 0.5$) as an indicator or even as the ultimate proof of the high predictive power of the QSAR model. They do not test the models for their ability to predict the activity of compounds of an external test set (i.e. compounds, which have not been used in the QSAR model development). Refs [27–31] provide several examples of recent publications, where the authors claim that their models have high predictive ability without validating them by using an external test set. Other authors validate their models using only one or two compounds that were not used in QSAR model development [32, 33], and still claim that their models are highly predictive. In contrast with such expectations, it has been shown that if a test set with known values of biological activities is available for prediction, there exists no correlation between LOO cross-validated q^2 and correlation coefficient R^2 between the predicted and observed activities for the test set [25, 34]. Another widely used approach to establish the model robustness is so called y-randomization (randomization of response, i.e. in our case, activities) [35]. It consists of repeating the calculation procedure with randomized activities and subsequent probability assessment of the resultant statistics. Frequently, it is used along with cross-validation. It is expected that models obtained for the dataset with randomized activity should have low values of q^2. However, sometimes models based on the randomized data have high q^2 values due to chance correlation or structural redundancy [36].

Several authors, including our group, have suggested that the only way to estimate the true predictive power of a QSAR model is to compare the predicted and observed activities of an (sufficiently large) external test set of compounds that were not used in the model development [25, 34, 37–39]. To estimate predictive power of a QSAR model, we recommended using the following statistical characteristics of the test set [34]: (i) correlation coefficient R between the predicted and observed activities; (ii) coefficients of determination [40] (predicted versus observed activities R_0^2, and observed versus predicted activities $R_0'^2$); (iii) slopes k and k' of the regression lines through the origin. We consider a QSAR model predictive, if the following conditions are satisfied [34]:

$$q^2 > 0.5; \quad (2)$$

$$R^2 > 0.6; \quad (3)$$

$$\frac{(R^2 - R_0^2)}{R^2} < 0.1 \text{ or } \frac{(R^2 - R_0'^2)}{R^2} < 0.1; \quad (4)$$

$$0.85 \leq k \leq 1.15 \text{ or } 0.85 \leq k' \leq 1.15. \quad (5)$$

The lack of the correlation between q^2 and R^2 has been noticed in Refs [25,37,38] and in our recent publication [34] where we demonstrated that all of the above criteria are necessary to adequately assess the predictive ability of a QSAR model. We suggested [34] that the external test set must contain at least five compounds, representing the whole range of both descriptor and activities of compounds included into the training set.

Rational division of the SAR dataset into a training and test sets

As discussed above, in order to obtain a reliable (validated) QSAR model, an available dataset should be divided into the training and test sets. Ideally, this division must be performed such that points representing both training and training set are distributed within the whole descriptor space occupied by the entire dataset, and each point of the test set is close to at least one point of the training set. This approach ensures that the similarity principle can be employed for the activity prediction of the test set. Unfortunately, as we shall see below, this condition can not always be satisfied.

Many authors use external test sets for validation of QSAR models, but do not provide any rationale as to how and why certain compounds were chosen for the test set [41, 42]. One of the most widely used method for dividing a dataset into training and test sets is a mere random selection [43, 44]. Some authors assign whole structural subgroups of molecules to the training or the test set [45, 46]. Another frequently used approach is based on the activity sampling. The whole range of activities is divided into bins, and compounds belonging to each bin are randomly (or in some regular way) assigned to the training or test sets [47, 48]. Obviously, these methods [43, 47, 48] can not guarantee that the training set compounds represent the entire descriptor space of the original dataset, and that each compound-point of the test set is close to at least one point of the training set.

In several publications, the division of a dataset into training and test sets is performed using the Kohonen's Self-Organizing Map (SOM) [49]. Representative points falling into the same areas of the SOM are randomly selected for the training and test sets [50, 51]. SOM preserves the closeness between points (points which are close to each other in the multidimensional descriptor space, are close to each other on the map). Therefore, it is anticipated that the training and test sets must be scattered within the whole area occupied by representative points in the original descriptor space, and that each point of the test set is close to at least one point of the training set. The drawback of this method is that the quantitative methods of prediction use exact values of distances between representative points: since SOM is a nonlinear projection method, the distances between points in the map are distorted.

The division of a dataset into the training and test sets can be performed using various clustering techniques. In Refs [52] and [53] K-means clustering algorithm [54] was used, and from each cluster one compound for the training set was randomly selected. In Refs [55], to select a representative subset from a dataset, hierarchical clustering and maximum dissimilarity method [56–58] were used. The authors showed that both methods choose representative subsets of compounds much better than the random selection. Compounds selected using the maximum dissimilarity method were used as training sets in 3D-QSAR studies, with all remaining compounds comprising the test set. In Ref. [43] Kennard-Stone [59–61] method, which is similar to the maximum dissimilarity method, was applied to the classification of NIR spectra and QSAR analysis. The drawbacks of clustering methods are that different clusters contain different number of points, and have different densities of representative points. Therefore, the closeness of each point of the test set to at least one point of the training set is not guaranteed. Maximum dissimilarity and Kennard-Stone methods guarantee that the points of the training set are distributed more or less evenly within the whole area occupied by representative points, and the condition of closeness of the test set points to the training set points is satisfied. The maximum distance between training and test set points in these methods does not exceed the radius of the probe sphere.

To select a representative subset of samples from the whole dataset, factorial designs [62, 63], and D-optimal designs [64] were used [43, 50, 65]. Factorial designs presume that different sample properties (such as substituent groups at certain positions) are divided into groups. Training set includes one representative for each combination of properties. For diverse dataset this approach is impractical, and fractional factorial designs are used, in which only a part of all combinations is included into the training set. Generally,

this approach does not guarantee the closeness of the test set points to the training set points in the descriptor space. D-optimal design algorithms select samples that maximize the |X′X| determinant, where X is the information (variance-covariance) matrix of independent variables (descriptors) [66–68]. The points maximizing the |X′X| determinant are spanned across the whole area occupied by representative points. They can be used as a training set, and the points not selected then used as the test set [43, 50].

In Ref. [43] four methods of sample selection (random, SOM, Kennard-Stone design and D-optimal design) were compared. The best models were built when Kennard-Stone and D-optimal designs were used. SOM was better than random selection, and D-optimal design was slightly better than the random selection.

In this paper we consider three closely related methods for rational division of a dataset into training and test sets. Our approaches are based on sphere exclusion algorithms [58], which are widely used for comparison of chemical databases or chemical libraries [69]. We have applied these methods to build and validate models for a dataset of 29 dopamine D_1 receptor ligands [17] using Molconn-Z descriptors [70] and *k*-nearest-neighbors (*k*NN) QSAR method [18]. In addition, we have developed validated models for a set of 66 Histamine H_1 receptor ligands [71] using a combination of Molconn-Z descriptors [70] and our topological chirality descriptors introduced recently [72, 73]. We show that rational selection of test and training sets using sphere exclusion algorithms leads to QSAR models with higher predictive ability than models based on alternative approaches to training and test set selection.

Methods

Descriptors

Descriptors were obtained using Molconn-Z [70] software and included simple and valence path, cluster, path/cluster and chain molecular connectivity indices [3–5], kappa molecular shape indices [6, 7], topological [8] and electrotopological state indices [9–12], differential connectivity indices [74], graph's radius and diameter [75], Wiener [76] and Platt [77] indices, Shannon [78] and Bonchev-Trinajstić [79] information indices, counts of different vertices [70], counts of paths and edges between different kinds of vertices [70].

Chirality descriptors [72, 73] used for the development of QSAR models for Histamine H1 receptor ligands [71] included [72, 73]: modified overall Zagreb indices [80], molecular connectivity indices [3–5], extended connectivity indices [81], and overall connectivity indices [82, 83]. They were calculated using the same formulas used for the conventional descriptors, but with modified vertex degrees for asymmetric atoms: a quantity named chirality correction was added to or subtracted from the conventional vertex degrees of atoms in R or S configurations, respectively. The detailed description of these novel topological chirality indices is given elsewhere [72, 73].

Prior to the development of QSAR models, descriptors were normalized according to the following formula

$$X_{ij}^{n} = \frac{X_{ij} - X_{j,\min}}{X_{j,\max} - X_{j,\min}}, \qquad (6)$$

where X_{ij} and X_{ij}^{n} are the non-normalized and normalized j-th ($j = 1, \ldots, K$) descriptor values for compound i ($i = 1, \ldots, N$), correspondingly, and $X_{j,\min}$ and $X_{j,\max}$ are the minimum and maximum values for j-th descriptor. Thus, for all descriptors, $\min(X_{ij}^{n}) = 0$ and, $\max(X_{ij}^{n}) = 1$. The total volume V occupied by the representative points in the normalized descriptor space was equal to one. The volume corresponding to one point was equal to 1/N.

Training and test set compounds selection

The following sphere-exclusion algorithms were used to divide a set of N compounds into training and test sets.

Algorithms 1 and 2 (sphere exclusion)

1. Select a compound with the highest activity.
2. Include this compound into the training set.
3. Construct a sphere with the center in the representative point of this compound with radius $R = c(V/N)^{1/K}$ (cf. Ref. [84]). Here, K is the number of descriptors (dimensionality of descriptor space), and c is the dissimilarity level [84]. Dissimilarity level was varied to construct different training and test sets.
4. Include compounds, corresponding to representative points within this sphere, except for the center of it, in the test set.
5. Exclude all points within this sphere from the initial set of compounds.
6. Let n be the number of remaining compounds. If n = 0, go to the last step, otherwise go to step 7.

7. Let m be the number of spheres already constructed. Calculate the distances d_{ij}, i=1,. . . ,n, j=1,. . . ,m of the representative points of the compounds left to the sphere centers.
8. Select a compound with the smallest (*algorithm 1*) or largest (*algorithm 2*) d_{ij}.
9. Go to step 2.
10. Stop.

Algorithm 3 (sphere exclusion)
Steps 1 to 5 of this algorithm are the same as for the first two algorithms.
Select randomly one of the remaining compounds, if any, and go to step 2.

Algorithm 4 (activity ranking)
To divide a dataset into training and test sets, this method uses only activities of compounds

1. Sort compounds by activity.
2. Specify the size of a group of compounds. Include the specified number of the most active compounds into the first group, the same number of the next most active compounds into the second group, etc. Of course, the last group of compounds may be smaller than the specified size.
3. Specify the number of compounds in each group, which will be included into the training set. Select this number of most active compounds of each group for the training set, and the remaining compounds for the test set.

Algorithm 5 (random selection)
The compounds for the test set are selected randomly from the whole dataset. The total number of compounds for the test set may be specified.

The first three algorithms allow constructing training sets that are distributed within the entire area of the descriptor space occupied by all representative points of the original dataset. The higher the dissimilarity level c, the smaller is the training set and the larger is the test set. Since activities of similar compounds (represented by points located closely to each other in the descriptors space) are supposed to be similar, the range of compound activities in the test set must be practically the same as the activity range for compounds in the training set. It is expected that in general, the predictive power of QSAR models decreases when dissimilarity level increases.

If there are no large gaps in activity values in the whole dataset, algorithm 4 in principle allows constructing test sets, which represent the whole range of activities. However, compounds with similar activities do not necessarily have similar structures. Therefore, algorithm 4 does not guarantee the closeness of the training and test set points in the descriptors space.

Quantitative estimates of the closeness of training and test set points in the descriptor space and the diversity of the training set

The following criteria for the division of a dataset into training and test sets have been implemented: the diversity index of the test set with respect to the training set $M_{test,train}$ and the diversity index of the training set with respect to the test set $M_{train,test}$ [84], which are defined as follows. The volume corresponding to one representative point is equal to V/N=1/N, where N is the number of compounds. We construct spheres with centers in the test set points with radius $R = c(V/N)^{1/K}$. Here, K denotes the number of descriptors (dimensionality of the descriptor space), and c is the dissimilarity level [84]. Let N_a be the number of test set points, for which the spheres contain no points of the training set. Then

$$M_{test,\ train}(c) = N_a/N_{test}, \tag{7}$$

where N_{test} is the number of compounds in the test set.

$M_{train,test}$ is defined as follows. We construct spheres with centers in the training set points with radius $R = c(V/N)^{1/K}$. Let N_b be the number of training set points, for which the spheres contain no points of the test set. Then

$$M_{train,test}(c) = N_b/N_{train}, \tag{8}$$

where N_{train} is the number of compounds in the training set. Obviously, $N_{train}+N_{test}=N$. $M_{test,train}$ characterizes the closeness of test set points to training set points. The lower is $M_{test,train}$, the better the condition of closeness of test set points to training set points is satisfied. Training and test sets obtained with algorithms 1 to 3 with dissimilarity level c satisfy the condition $M_{test,train}(c)= 0$ automatically. $M_{train,test}$ characterizes the quality of validation procedure of a QSAR model. The lower is $M_{train,test}$, the better is the validation procedure. If, for example $M_{train,test}=0$, then all areas of the descriptor space, in which the points of the training set are distributed, are tested. Of course, $M_{train,test}$ depends not only on the algorithm of the division of a dataset into the training and test

sets, but also on the distribution of points of the whole dataset. We calculate $M_{test,train}$ and $M_{train,test}$ for all our training and test sets.

To estimate the diversity of the training set the index I_{train} [84] is used, which is defined as follows. We construct spheres with centers in the training set points with radius $R = c(V/N)^{1/K}$. Let N_c be the number of points in the training set, for which the spheres contain no other points of the training set. Then

$$I_{train(c)} = N_c/N_{train}. \quad (9)$$

Algorithms 1 to 3 provide the highest value of $I_{train}(c)$ which is equal to 1 automatically.

k-Nearest neighbors QSAR

K-nearest neighbors (*k*NN) QSAR method [18] was applied for QSAR studies of all examples considered in this paper. The method uses LOO cross-validation procedure and an evolutionary simulated-annealing algorithm for descriptor selection. The procedure starts with the random selection of a predefined number of descriptors out of all descriptors. Activities of compounds excluded in LOO procedure are estimated using the following formula

$$\hat{y} = \frac{\sum_{\text{nearest neighbors}} y_i \exp(-d_i)}{\sum_{\text{nearest neighbors}} \exp(-d_i)}, \quad (10)$$

where d_i are the distances between nearest neighbors and this compound. After each run, q^2 is calculated:

$$q^2 = 1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(\bar{y} - \hat{y}_i)^2}, \quad (11)$$

where y_i, $\hat{y}_i$ and $\bar{y}$ are the actual, predicted and mean values of activity. The summation in (11) is performed over all compounds. After each run, a fraction of descriptors are randomly replaced by other descriptors from the original pool, and the new value of q^2 is obtained. If $q^2(\text{new}) \leq q^2(\text{old})$, the new set of descriptors is accepted with probability $p = \exp(q^2(\text{new}) - q^2(\text{old}))/T$, and rejected with probability (1-p), where T is a simulated 'temperature' annealing parameter. During this process, T is decreasing until the predefined value. Thus, q^2 is optimized (see Ref. [18] for additional details). In the prediction process, the final set of descriptors selected is used, and expression (10) is applied to predict activities of the test set compounds.

Experimental datasets

29 Dopamine D_1 Receptor Ligands

This data set includes a variety of tetrahydroisoquinolines, tetrahydrobenzazepines, and phenylaminotetralins, as well as representatives of several other series [17]. The competitive binding affinities of these compounds to the dopamine D_1 receptor were measured and expressed as $-\log(K_{0.5})$ [17].

Procedures based on the first three algorithms for dividing a dataset into training and test sets (Section 2.2) were repeated using different values of the dissimilarity level. The starting dissimilarity level was set to 0.4. Then it was increased with step 0.2. The minimum dissimilarity level for which QSAR models were built was defined to allow at least five compounds in the test set. The maximum dissimilarity level was equal to 3.6. The number of descriptors for the *k*NN QSAR procedure was varied from 10 to 40 with step 2. For each number of descriptors 10 QSAR models were obtained. Thus, for each algorithm the total number of QSAR models built for each value of dissimilarity level was $16 \times 10 = 160$. All models with $q^2 > 0.5$ were validated using test sets.

For algorithm 4 (Section 2.2), calculations were performed for ten different training and test sets differing in the number of groups and the number of compounds in each group (cf., Section 2.2). For algorithm 5 (Section 2.2), the number of compounds in the training set varied from 6 to 22 with step 2.

66 Histamine H_1 receptor antagonists

This dataset includes 35 analogs of 1-phenyl-3-amino-1,2,3,4-tetrahydronaphtalenes [1-phenyl-3-aminotetralins (PATs)] and 31 non-PATs [71]. Binding affinities $K_{0.5}$ to the histamine H_1 receptor of these compounds are available [71]. Standard CoMFA and CoMFA/q^2-GRS [23] models were built previously using the training set containing 50 of these compounds [71]. The models were validated using the predictions of $K_{0.5}$ values for the external test set consisted of the remaining 16 compounds.

QSAR models were developed using a combination of various chirality descriptors [72, 73] with conventional Molconn-Z [70] descriptors as well as non-chiral overall Zagreb indices [80], molecular connectivity indices [3–5], extended connectivity indices [81], and overall connectivity indices [82, 83].

Algorithm 1 for dividing a dataset into training and test sets (Section 2.2) was applied for different dissimilarity levels. The starting dissimilarity level was equal

to 0.5, and then it was increased with step 0.5. The minimum dissimilarity level was defined to allow at least five compounds in the test set. The maximum dissimilarity level was defined as the one, for which either the test set contained at least half of all compounds or no good (predictive) QSAR model was obtained.

The number of descriptors for the *k*NN QSAR procedure was varied from 10 to 40 with step 2. For each fixed number of descriptors 10 QSAR models were developed. Thus, for each value of dissimilarity level as many as $16 \times 10 = 160$ QSAR models were built. All QSAR models with $q^2 > 0.5$ were validated using corresponding external test sets. Models, which satisfied criteria (2)-(5) (cf., Introduction), were considered as good models. To estimate predictive ability of our models, we also considered the F-test and the α- or *p*-value (which is the boundary significance level between null hypothesis H_0 that assumes that the model does not predict better than the average activity value and the opposite H_1 hypothesis). α-value is one minus *p*-value [40].

Results

29 Dopamine D1 receptor ligands

Training and test sets obtained with Algorithms 1–3

As a general trend, with the increase in the dissimilarity level, the number of compounds in training sets gradually decreases, and the number of compounds in the corresponding test sets gradually increases (see Tables 1–3). The deviations from this trend can be partially explained by a non-uniform distribution of the representative points of compounds in the descriptor space. It is expected that for smaller training sets (and for higher dissimilarity levels) the predictive ability of QSAR models must generally decrease.

As mentioned previously, for all the training and test sets $M_{test,train}(c)=0$ and $I_{train}(c)=1$, which are the optimal values of these indices. The results presented in Tables 1–3, indicate that for all three algorithms $M_{train,test}$ values follow the same trend. First, with the increase in the dissimilarity level, $M_{train,test}$ decreases, and after achieving some optimal value it increases. This behavior can be explained as follows. If the dissimilarity level is low, the test set is small, and the representative points of the test set cannot be distributed within the whole area occupied by representative points in the descriptor space. If the number of points in the test set increases, they are distributed over increasingly bigger part of the area occupied by the whole dataset. Thus, the $M_{train,test}$ values are decreasing. At the same time, the number of points of the training set is decreasing. If a dataset includes isolated points, they will belong to the training set, even if the dissimilarity level is high. With the number of points of the training set decreases, these points will constitute increasingly bigger part of the training set. Thus, in this case, $M_{train,test}$ will increase. Since lower $M_{train,test}$ values are preferred, we can conclude that the Algorithm 2 divides our dataset into training and test sets slightly better than Algorithm 1, while Algorithm 3 is slightly worse than Algorithm 1.

QSAR models developed and validated using training and test sets obtained with Algorithms 1–3

Best QSAR models are presented in Tables 1–3, which demonstrates that the best results were obtained with the first two algorithms. Models obtained using training and test sets obtained with Algorithm 3 (Table 3) are slightly worse, but still comparable to those obtained with Algorithms 1 and 2 (Tables 1 and 2). Almost all models built and validated using training and test sets obtained with Algorithms 1 and 2 with the dissimilarity level values up to 3.2 satisfy our conditions (2)–(5) (see Introduction). Training and test sets obtained with the Algorithm 3 gave models satisfying these conditions up to the dissimilarity level 2.6. All models in Tables 1–3 have α-values close to one.

Training and test sets obtained using Algorithm 4

The numbers of compounds in training and test sets obtained with Algorithm 4 are shown in Table 4. In this case, we can find the minimum dissimilarity level, for which $M_{test,train}=0$ as follows. For each point of the test set we define the distances to training set points d_{ij}, where $i = 1,\ldots,N_{test}$ and $j = 1,\ldots, N_{train}$. Then

$$c_{\min}(M_{\text{test,train}} = 0) = \frac{\max\limits_{i}\min\limits_{j} d_{ij}}{(1/N)^{1/K}} \qquad (12)$$

$c_{min}(M_{test,train})$ values are presented in Table 4. These values are significantly higher than dissimilarity levels, which correspond to the same (or approximately the same) number of compounds in Tables 1–3. It means that the condition of closeness of the test set to the training set is not satisfied.

Similarly, we can define the maximum dissimilarity level, for which $I_{train}=1$. If d_{ij} are the distances between points of the training set, i and $j=1,\ldots,N_{train}$, then

Table 1. Statistics for best models obtained using algorithm 1 for each value of dissimilarity level (DL).

No	DL[a]	Training set	Test set	$M_{train,test}$	q^2	R	R^2	R_0^{2b} or R'^2_0	k or k'^b	F
1	1.6	21	8	0.67	0.76	0.96	0.91	0.91	1.01	62.0
2	1.8	20	9	0.60	0.67	0.99	0.97	0.96	0.97	221.2
3	2.0	16	13	0.63	0.88	0.86	0.74	0.74	1.03	31.8
4	2.2	14	15	0.36	0.71	0.90	0.80	0.79	0.94	52.1
5	2.4	12	17	0.25	0.87	0.75	0.56	0.56	1.02	19.3
6	2.6	12	17	0.17	0.65	0.80	0.64	0.60	1.01	26.1
7	2.8	11	18	0.18	0.66	0.82	0.67	0.67	0.97	32.3
8	3.0	9	20	0.22	0.76	0.78	0.60	0.56	1.03	27.3
9	3.2	9	20	0.22	0.76	0.78	0.60	0.60	0.99	27.5
10	3.4	8	21	0.25	0.73	0.75	0.56	0.55	0.99	24.1

[a]dissimilarity level
[b]see conditions (2)–(5)

Table 2. Statistics for best models obtained using algorithm 2 for each value of dissimilarity level.

No	DL[a]	Training set	Test set	$M_{train,test}$	q^2	R	R^2	R_0^{2b} or R'^2_0	k or k'^b	F
1	1.6	21	8	0.67	0.72	0.98	0.97	0.96	0.97	181.2
2	1.8	20	9	0.60	0.74	0.98	0.97	0.96	1.01	202.4
3	2.0	18	11	0.56	0.70	0.96	0.92	0.91	1.00	107.1
4	2.2	15	14	0.33	0.78	0.77	0.59	0.59	0.96	17.6
5	2.4	13	16	0.23	0.52	0.84	0.71	0.70	0.99	34.6
6	2.6	12	17	0.17	0.77	0.83	0.69	0.68	1.03	33.0
7	2.8	12	17	0.17	0.71	0.82	0.67	0.67	0.96	30.2
8	3.0	11	18	0.18	0.83	0.89	0.79	0.75	0.97	61.3
9	3.2	9	20	0.22	0.82	0.83	0.69	0.66	0.94	39.6
10	3.4	8	21	0.25	0.79	0.72	0.52	0.51	1.03	20.9

[a]dissimilarity level
[b]see conditions (2)–(5)

Table 3. Statistics for best models obtained using algorithm 3 for each value of dissimilarity level.

No	DL[a]	Training set	Test set	$M_{train,test}$	q^2	R	R^2	R_0^{2b} or R'^2_0	k or k'^b	F
1	1.6	21	8	0.67	0.72	0.99	0.98	0.96	1.01	254.9
2	1.8	20	9	0.60	0.78	0.98	0.96	0.95	1.00	155.0
3	2.0	18	11	0.56	0.80	0.86	0.74	0.74	1.06	25.7
4	2.2	14	15	0.36	0.83	0.75	0.56	0.56	0.96	16.6
5	2.4	11	18	0.27	0.84	0.81	0.65	0.64	0.96	29.4
6	2.6	12	17	0.17	0.77	0.83	0.69	0.68	0.98	32.7
7	2.8	8	21	0.25	0.92	0.71	0.50	0.43	0.95	18.9
8	3.0	9	20	0.22	0.75	0.71	0.51	0.46	1.06	18.3
9	3.2	9	20	0.22	0.88	0.65	0.42	0.30	0.98	13.1
10	3.4	8	21	0.25	0.77	0.76	0.57	0.55	1.03	25.4
11	3.6	6	23	0.40	0.75	0.75	0.57	0.54	1.02	27.6

[a]dissimilarity level
[b]see conditions (2)–(5)

Table 4. The number of compounds in training and test sets using algorithm 4 (see Section 2.2)

No	Size of a group	Compounds in the training set from one group	Training set	Test set	$c_{min}(M_{test,train})=0$	$c_{max}(I_{train}=1)$
1	2	1	15	14	5.60	1.41
2	3	2	20	9	4.52	0.87
3	3	1	10	19	6.16	0.87
4	4	3	22	7	4.52	1.32
5	4	2	15	14	4.52	1.60
6	4	1	8	21	6.26	1.65
7	5	4	24	5	2.02	0.87
8	5	3	18	11	4.47	1.48
9	5	2	12	17	4.59	1.73
10	5	1	6	23	6.02	1.73

$$c_{\max}(I_{\text{train}} = 1) = \frac{\min d_{ij}}{(1/N)^{1/K}} \quad (13)$$

$c_{max}(I_{train}=1)$ are presented in Table 2. These values are significantly lower than dissimilarity levels (may be even not the maximum ones), which correspond to the same (or approximately the same) number of compounds in Tables 1–3. It means that training sets obtained using Algorithm 4 are less diverse than training sets obtained using Algorithms 1–3.

QSAR models developed and validated using training and test sets obtained with Algorithm 4

Best QSAR models are presented in Table 5, which shows that the models built using different training sets of the same size may be either highly predictive (Model 5) or not (Model 1). Most of the models presented in Table 5 have worse statistics than the models in Tables 1–3. α-values are also lower than for models built for training and test sets obtained with Algorithms 1–3, and for Model 2 $\alpha=0.9292$, i.e. H_1 hypothesis cannot be accepted with the 95% confidence coefficient. This can be explained by the lack of closeness of the test set to the training set and a lower diversity of the training set as compared to the training sets obtained with our sphere-exclusion algorithms.

Training and test sets obtained with Algorithm 5

$c_{min}(M_{test,train})$ values calculated using formula (12) are presented in Table 6. These values are significantly higher than dissimilarity levels for training and test sets of similar size in Tables 1–3. This means that the condition of closeness of the test set to the training set is not satisfied. $c_{max}(I_{train}=1)$ values calculated using formula (13) are also presented in Table 6. These values are significantly lower than dissimilarity levels for training and test sets of similar size in Tables 1–3. This implies that training sets obtained with random selection from the whole dataset are less diverse than training sets obtained using our Algorithms 1–3.

QSAR models developed and validated using training and test sets obtained with Algorithm 5

Best QSAR models are presented in Table 6. Most of the models presented in Table 6 have a worse statistics than for models in Tables 1–3 built for training sets of roughly similar sizes. α-values are also lower than for the models built for training and test sets obtained with Algorithms 1–3, and for Model 1 $\alpha=0.8914$, i.e. H_1 hypothesis cannot be accepted even with the 90% confidence coefficient. As in the previous case, it can be explained by the lack of closeness of the test set to the training set and a lower diversity of the training set with respect to the training sets obtained using our sphere-exclusion algorithms.

66 Histamine H_1 receptor ligands

QSAR models were developed initially using training and test sets obtained with Algorithm 1 and using the same training and test sets as in Ref. [71]. The best model built with the training and test sets from Ref. [71] was characterized by the following statistics $q^2=0.69$, $R^2=0.72$, $R_0^2=0.72$, $k=1.02$, $F=35.4$, and

Table 5. Statistics for best models obtained using algorithm 4 for each training and test set.

No	Training set	Test set	q^2	R	R^2	R_0^{2a} or R'^2_0	k or k'^a	F
1	15	14	0.79	0.81	0.65	0.64	1.00	22.2
2	20	9	0.82	0.63	0.39	0.26	0.98	4.53
3	10	19	0.86	0.69	0.48	0.47	1.00	15.8
4	22	7	0.83	0.94	0.88	0.79	0.96	35.8
5	15	14	0.87	0.79	0.62	0.61	0.93	19.2
6	8	21	0.73	0.74	0.55	0.54	1.06	23.3
7	24	5	0.73	1.00	1.00	0.96	0.98	868
8	18	11	0.81	0.69	0.48	0.47	0.92	8.28
9	12	17	0.80	0.74	0.54	0.48	0.95	17.7
10	6	23	0.70	0.60	0.37	0.34	1.04	12.1

[a]see conditions (2)–(5)

Table 6. Statistics for best models obtained using algorithm 5 for each value of dissimilarity level.

No	Training set	Test set	$c_{min}(M_{test,train})=0$	$c_{max}(I_{train}=1)$	Q^2	R	R^2	R_0^{2a} or R'^2_0	k or k'^a	F
1	23	6	3.07	0.86	0.85	0.72	0.51	0.44	1.06	4.24
2	21	8	3.07	0.86	0.82	0.81	0.66	0.66	1.05	11.7
3	19	10	3.07	1.31	0.88	0.73	0.53	0.51	1.08	8.9
4	17	12	3.07	1.31	0.73	0.87	0.76	0.75	1.05	31.8
5	15	14	4.44	1.31	0.90	0.73	0.53	0.53	0.95	13.6
6	13	16	4.44	1.31	0.95	0.74	0.55	0.54	0.93	17.0
7	11	18	4.44	1.31	0.92	0.77	0.60	0.48	0.90	23.6
8	9	20	4.56	1,31	0.95	0.69	0.47	0.42	0.91	16.0
9	7	22	6.04	1.59	0.71	0.74	0.55	0.54	0.98	24.1

[a]see conditions (2)–(5)

α=1–3.5489$\times 10^{-5}$. Predictive power of this model is illustrated in Figure 1a. Several models built for the training set containing about 50 compounds selected with Algorithm 1, have much better predictive ability than the models built using the training set of a similar size in Ref. [71]. Predictive power of one of these models (q^2=0.55, R^2=0.83,R_0^2=0.83, k=0.97, F=69.1, and α=1–8.7402$\times 10^{-7}$) is illustrated in Figure 1b. Better predictive ability of these models can be explained by better (rational) division of all compounds into the training and test sets used in these calculations. The rational methodology for selecting training and test sets affords predictive QSAR models for relatively large test sets. Thus, one model built using the training set containing only 36 molecules and validated using the test set containing 30 compounds had the following statistics: q^2=0.67, R^2=0.61, R_0^2=0.60, k=0.97, F=43.2, and α=1–3.9767$\times 10^{-7}$ (see also Figure 2).

Final notes and conclusions

In this paper we have considered one of the most important aspects of predictive QSAR analysis: selection of training and test sets of compounds. We established the following criteria, which both the training and test sets must satisfy.

(i) Closeness of the representative points of the test set to representative points of the training set in the multidimensional descriptor space. The concept of closeness is based on the general assumption underlying all QSAR theories: similar compounds

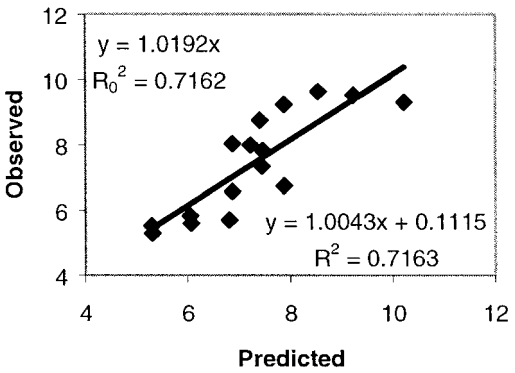

Figure 1a. Histamine H_1 receptor ligands. Observed vs predicted $K_{0.5}$ values for the external test set the same as in Ref. [71].

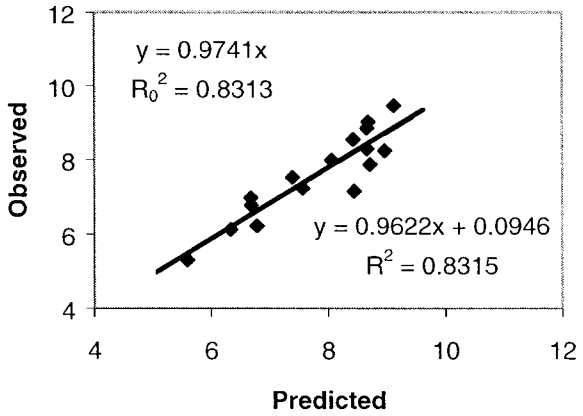

Figure 1b. Histamine H_1 receptor ligands. Observed vs predicted $K_{0.5}$ values for the external test set consisting of 16 compounds.

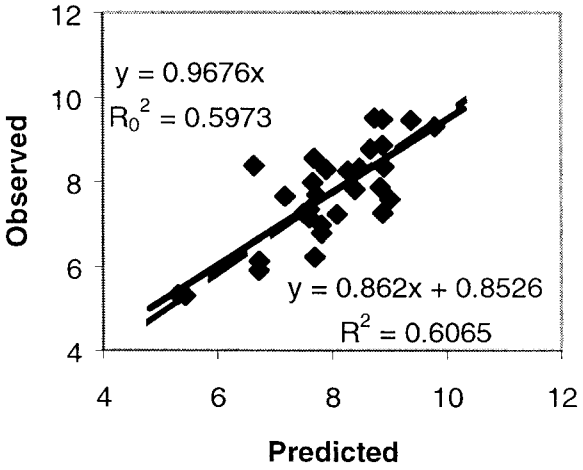

Figure 2. Histamine H_1 receptor ligands. The model was built using the training set consisted of 36 compounds. Observed $K_{0.5}$ values for the external test set consisting of 30 compounds.

have similar activities. If a compound with unknown activity is represented by a point in the descriptor space, which is close to at least one point representing a compound with known activity, the prediction of activity for this compound is possible based on similarity principle; otherwise the prediction is unreasonable. Thus, if one wishes to validate a QSAR model, the points of the test set must be close to the points of the training set in the descriptor space. This closeness can be measured by the diversity index of the test set with respect to the training set, $M_{test,train}$ [84]. This index can take values in the interval [0,1]. It depends on the radius of the probe spheres built around the test set points. The radius is proportional to a quantity named dissimilarity level and depends on the total number of compounds in the dataset and the dimensionality of the descriptor space. At any given dissimilarity level, the lower values of the diversity index of the test set with respect to the training set correspond to higher closeness of the test set to the training set.

(ii) Closeness of representative points of the training set to representative points of the test set in the multidimensional descriptor space. The closeness of the training set to the test set can be measured by the diversity index of the training set with respect to the test set, $M_{train,test}$ [84]. The index can take values in the interval [0,1]. It depends on the radius of probe spheres built around the test set points, which is proportional to a quantity named dissimilarity level. It also depends on the total number of compounds in the dataset and the dimensionality of the descriptor space. For the given dissimilarity level, lower values of the diversity index of the training set with respect to the test set correspond to higher closeness of the training set to the test set. We showed that this criterion depends not only on the procedure of the division of a dataset into training and test sets, but also on the dataset itself. If a dataset contains unique compounds dissimilar to all other compounds, it is impossible to satisfy both criteria (i) and (ii) at the same time.

(iii) Diversity of the training set. A QSAR model must predict activities of compounds, represented by points found within or in the vicinity of the area occupied by the representative points of the whole dataset. Thus, points representing the training set must be distributed within the entire area occupied by representative points for the whole dataset. In this paper we measured the diversity of

the training set by the diversity index of the training set with respect to itself I_{train} [84], which takes values within the interval [0,1]. Higher values of this index correspond to higher diversity of the training set. This criterion is particularly useful, if training set must include a predefined number of compounds. It is also important in cross-validation procedure to avoid the structural redundancy and overfitting, leading to an overestimate of q^2 values.

In summary, we have proposed three closely related sphere-exclusion algorithms, which divide a dataset into training and test sets such that $M_{test,train}=0$ and $I_{train}=1$, i.e. these indices take their optimal values. We further proposed several statistical characteristics of the test and training set models that are used to evaluate both the robustness and predictive ability of the models. We used our algorithms to build training and test sets for a series of 29 Dopamine D_1 receptor ligands [17] and a series of 66 Histamine H_1 receptor ligands [71]. We showed that none of the alternative popular approaches to select training and test sets, i.e., activity ranking, and random selection, could satisfy conditions (i) and (iii) above. Furthermore, and most importantly, we also showed that these alternative approaches yield models with significantly lower predictive power than models obtained with the rational selection of the test and training sets. We conclude that rational division of experimental datasets into training and test sets for model building and validation, respectively, using diversity sampling algorithms such as described in this paper, should be adopted by the QSAR research community as guiding principles of predictive QSAR modeling.

References

1. Hansch, C., Fujita, T., J. Am. Chem. Soc., 86 (1964) 1616–1626.
2. Kubinyi, H., In: Mannhold, R. et al. (eds.) Methods and Principles in Medicinal Chemistry, VCH, Weinheim, 1993.
3. Randić, M., J. Am. Chem. Soc., 97 (1975) 6609–6615.
4. Kier, L.B. and Hall, L.H., Molecular Connectivity in Chemistry and Drug Research. Academic Press, New York, 1976.
5. Kier, L.B. and Hall, L.H., Molecular Connectivity in Structure-Activity Analysis. Wiley, New York, 1986.
6. Kier, L.B., Quant. Struct.-Act. Relat. 4 (1985) 109–116.
7. Kier, L.B., Quant. Struct-Act. Relat. 6 (1987) 8–12.
8. Hall, L.H. and Kier, L.B., Quant. Struct.-Act. Relat 9 (1990) 115–131.
9. Hall, L.H., Mohney, B.K. and Kier, L.B., Quant. Struct.-Act. Relat., 10 (1991) 43–51.
10. Hall, L.H., Mohney, B.K. and Kier, L.B., J. Chem. Inf. Comput. Sci., 31 (1991) 76–82.
11. Kier, L.B. and Hall, L.H., Molecular Structure Description: The Electrotopological State, Academic Press, 1999.
12. Kellogg, G.E., Kier, L.B., Gaillard, P. and Hall, L.H., J. Comput. Aid. Mol. Des. 10 (1996) 513–520.
13. Sheridan, R.P., Nachbar, R.B. and Bush, B.L., J. Comput.-Aid Mol. Des. 8 (1994) 323–340.
14. Matter, H., J. Medic. Chem. 40(8) (1997) 1219–1229.
15. Clementi, S. and Wold, S., In: Waterbeemd, H. van de (ed.), Chemometrics Methods in Molecular Design, VCH, (1995) 319–338.
16. Wold, S., In: Waterbeemd, H. van de (ed.), Chemometrics Methods in , VCH, (1995) 195–218.
17. Hoffman B., Cho S.J., Zheng W., Wyrick S., Nichols D.E. and Mailman R.B., J. Med. Chem. 42 (1999) 3217–3226.
18. Zheng, W. and Tropsha, A., J. Chem. Inf. Comput. Sci., 40 (2000) 185–194.
19. Ajay. J. Med. Chem. 36 (1993) 3565–3571.
20. Cramer III, R.D., Patterson, D.E. and Bunce, J.D., J. Am. Chem. Soc. 110 (1988) 5959–5967.
21. Marshall, G.R. and Cramer III, R.D., Trends Pharmacol. Sci. 9 (1988) 285–289.
22. Pérez, C., Pastor, M., Ortiz, AR. and Gago, F., J. Med. Chem. 41 (1998) 836–852.
23. Cho, S.J. and Tropsha, A., J. Med. Chem. 38 (1995) 1060–1066.
24. Klebe, G., In: Kubinyi, H., Folkers, G., Martin, Y.C., (eds.) 3D QSAR in Drug Design. Volume 3. Recent Advances, Kluwer/ESCOM: Dordrecht, (1998) pp. 87–104.
25. Kubinyi, H., Hamprecht, F.A. and Mietzner, T., J. Med. Chem., 41 (1998) 2553–2564.
26. Topliss, J.G. and Edwards, R.P., J. Med. Chem. 22 (1979) 1238–1244.
27. Gironés, X., Gallegos, A. and Ramon, C.-D., J. Chem. Inf. Comput. Sci. 46 (2000) 1400–1407.
28. Bordás, B., Kömíves, T., Szántó, Z. and Lopata, A., J. Agric. Food Chem. 48 (2000) 926–931.
29. Fan, Y., Shi, L.M., Kohn, K.W., Pommier, Y. and Weinstein, J.N., J. Med. Chem. 44 (2001) 3254–3263.
30. Randić, M. and Basak, S.C., J. Chem. Inf. Comput. Sci. 40 (2000) 899–905.
31. Suzuki, T., Ide, K., Ishida, M. and Shapiro, S., J. Chem. Inf. Comput. Sci. 41 (2001) 718–726.
32. Recanatini, M., Cavalli, A., Belluti, F., Piazzi, L., Rampa, A., Bisi, A., Gobbi, S., Valenti, P., Andrisano, V., Bartolini, M. and Cavrini, V., J. Med. Chem. 43 (2000) 2007–2018.
33. Morón, J.A., Campillo, M., Perez, V., Unzeta, M. and Pardo, L., J. Med. Chem. 43 (2000) 1684–1691.
34. Golbraikh, A. and Tropsha, A., J. Mol. Graphics Model. 20 (2002) 269–276.
35. Wold, S. and Eriksson, L., Statistical Validation of QSAR Results. In: Waterbeemd, H. van de (ed.), Chemometrics Methods in Molecular Design, VCH, (1995) 309–318.
36. Clark, R.D., Sprous, D.G. and Leonard, J.M., Validating Models Based on Large Dataset. In: Höltje, H.-D., Sippl, W., (eds.) Rational Approaches to Drug Design. Proceedings of the 13th European Symposium on Quantitative Structure-Activity Relationships. Aug 27 – Sept 1 (2000), Duesseldorf, Germany. Prous Science, (2001) 475–485.
37. Novellino, E., Fattorusso, C. and Greco, G., Pharm. Acta Helv. 70 (1995) 149–154.
38. Norinder, U., J. Chemomet. 10 (1996) 95–105.
39. Zefirov, N.S. and Palyulin, V.A., J. Chem. Inf. Comput. Sci. 41 (2001) 1022–1027.

40. Sachs, L., Applied Statistics. A Handbook of Techniques. Springer-Verlag, (1984).
41. Huuskonen, J., J. Chem. Inf. Comput. Sci. 41 (2001) 425–429.
42. Tetko, I.V., Kovalishyn, V.V. and Livingstone D.J., J. Med. Chem. 44 (2001) 2411–2420.
43. Wu, W., Walczak, B., Massart, D.L., Heuerding, S., Erni, F., Last, I.R. and Prebble, K.A., Chemometr. Intell. Lab. Syst. 33 (1996) 35–46.
44. Yasri, A. and Hartsough, D., J. Chem. Inf. Comput. Sci. 41 (2001) 1218–1227
45. Bernard P., Kireev D.B., Chretien J.R., Fortier P.L. and Coppet L., J. Comput. Aided Mol. Des. 13 (1999) 355–371.
46. Takeuchi, Y., Shands, E.F.B., Beusen, D.D. and Marshall, G.R., J. Med. Chem. 41 (1998)3609–3623.
47. Kauffman, G.V. and Jurs, P.C., J. Chem. Inf. Comput. Sci. 41 (2001) 1553–1560.
48. Mattioni, B.E. and Jurs, P.C., J. Chem. Inf. Comput. Sci., in press.
49. Gasteiger, J. and Zupan, J., Angewandte chemie. 32(4) (1993) 503.
50. Loukas, Y.L., J. Med. Chem. 44 (2001) 2772–2783.
51. Bernard, P, Pintore, M, Berthon, J.Y. and Chretien, J.R., Eur. J. Med. Chem. 36 (2001) 1–19.
52. Burden, F.R. and Winkler, D.A., J. Med. Chem. 42 (1999) 3183–3187.
53. Burden, F.R., Ford, M.G., Whitley, D.C. and Winkler, D.A., J. Chem. Inf. Comput. Sci. 40 (2000) 1423–1430.
54. Adams, M.J., Chemometrics in Analytical Spectroscopy. The Royal Society of Chemistry, UK, 1995.
55. Potter, T. and Matter, H., J. Med. Chem. 41 (1998) 478–488.
56. Lajiness, M., Johnson, M.A. and Maggiora, G.M., In: Fauchere, J.L., (ed.), QSAR: Quantitative Structure-Activity Relationships in Drug Design Alan R. Liss Inc.: New York, (1989) pp. 173–176.
57. Taylor, R., J. Chem. Inf. Comput. Sci. 35 (1995) 59–67.
58. Snarey, M., Terrett, N.K., Willett, P. and Wilton, D.J., J. Mol. Graphics Mod. 15 (1997) 372–385.
59. Kennard, R.W. and Stone, L.A., Technometrics 11 (1969) 137–148.
60. Bourguignon, B., Deaguiar, P.F., Thorre, K. and Massart, D.L., J. Chromatogr. Sci. 32 (1994) 144–152.
61. Bourguignon, B., Deaguiar, P.F., Khots, M.S. and Massart, D.L., Anal. Chem. 66 (1994) 893–904.
62. Hellberg, S., Eriksson, L., Jonsson, J., Lindgren, F., Sjostrom, M., Skagerberg, B., Wold, S. and Andrews, P., Int. J. Pept. Protein. Res. 37 (1991) 414–424.
63. Eriksson, L. and Johansson, E., Chemometr. Intell. Lab. Syst. 34 (1996) 1–19.
64. Carlson, R., Design and Optimization in Organic Synthesis. Elsevier, (1992).
65. Martin, E.J. and Critchlow, R.E., J. Comb. Chem. 1 (1999) 32–45.
66. Miller, A. and Nguyen, N.-K., Appl. Stat. 43 (1994) 669–678.
67. Mitchell, T.J., Technometrics 16 (1974) 203–210.
68. Mitchell, T.J., Technometrics 42 (2000) 48–54.
69. Reynolds, C.H., Druker, R. and Pfahler, L.B., J. Chem. Inf. Comput. Sci. 38 (1998) 305–312.
70. Molconn-Z. http://www.eslc.vabiotech.com/
71. Bucholz, E., Brown, R.L., Tropsha, A., Booth, R.G. and Wyrick, S.D., J. Med. Chem. 42 (1999) 3041–3054.
72. Golbraikh, A., Bonchev, D., Xiao, Y.-D. and Tropsha, A., In: Rational Approaches to Drug Design. Proceedings of the 13^{th} European Symposium on quantitative Structure-Activity relationships, Prous Science, (2001) pp. 219–223.
73. Golbraikh A., Bonchev, D. and Tropsha, A., J. Chem. Inf. Comput. Sci. 41 (2001) 147–158.
74. Kier, L.B. and Hall, L.H., Quant. Struct.-Act. Relat. 10 (1991) 134–140.
75. Petitjean, M., J. Chem. Inf. Comput. Sci. 32 (1992) 331–337.
76. Wiener, H., J. Am. Chem. Soc. 69 (1947) 17.
77. Platt, J.R., J. Phys. Chem. 56 (1952) 328.
78. Shannon, C. and Weaver, W., Mathematical theory of Communication, University of Illinois, Urbana, (1949).
79. Bonchev, D., Mekenyan, O. and Trinajstic, N., J. Comput. Chem., 2 (1981) 127–148.
80. Gutman I., Ruscić, B., Trinajstić, N. and Wilcox, C.F., Jr., J. Chem. Phys., 62 (1975) 3399.
81. Rücker, G. and Rücker, C., J. Chem. Inf. Comput. Sci., 33 (1993) 683–695.
82. Bonchev, D., In: Devillers J., Balaban, A.T. (eds.), Topological Indices and Related Descriptors, Gordon and Breach, Reading, U.K. (1999) pp. 361–401.
83. Bonchev, D., SAR/QSAR Env. Res., 7 (1997) 23–43.
84. Golbraikh, A., J. Chem. Inf. Comput. Sci. 40 (2000) 414–425.