Journal of Computer-Aided Molecular Design **17:** 241–253, 2003.

Rational selection of training and test sets for the development of validated QSAR models

Alexander Golbraikh, Min Shen, Zhiyan Xiao, Yun-De Xiao, Kuo-Hsiung Lee & Alexander Tropsha*

Division of Medicinal Chemistry and Natural Products, School of Pharmacy, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599-7360, USA

Received 11 November 2002; Accepted 14 January 2003

Summary

Quantitative Structure–Activity Relationship (QSAR) models are used increasingly to screen chemical databases and/or virtual chemical libraries for potentially bioactive molecules. These developments emphasize the importance of rigorous model validation to ensure that the models have acceptable predictive power. Using *k* nearest neighbors (*k*NN) variable selection QSAR method for the analysis of several datasets, we have demonstrated recently that the widely accepted leave-one-out (LOO) cross-validated R^2 (q^2) is an inadequate characteristic to assess the predictive ability of the models [Golbraikh, A., Tropsha, A. Beware of q2! *J. Mol. Graphics Mod.* 20, 269-276, (2002)]. Herein, we provide additional evidence that there exists no correlation between the values of q^2 for the training set and accuracy of prediction (R^2) for the test set and argue that this observation is a general property of any QSAR model developed with LOO cross-validation. We suggest that external validation using rationally selected training and test sets provides a means to establish a reliable QSAR model. We propose several approaches to the division of experimental datasets into training and test sets and apply them in QSAR studies of 48 functionalized amino acid anticonvulsants and a series of 157 epipodophyllotoxin derivatives with antitumor activity. We formulate a set of general criteria for the evaluation of predictive power of QSAR models.

Introduction

Quantitative Structure-Activity Relationship (QSAR) models are used increasingly in chemical data mining and combinatorial library design [1–3]. Thus, three-dimensional (3D) stereoelectronic pharmacophore based on QSAR modeling was used to search the National Cancer Institute Repository of Small Molecules to find new leads for inhibiting human immunodeficiency virus type 1 reverse transcriptase at the nonnucleoside binding site [4]. A QSAR model developed for cytochrome P450 1A2 (CYP1A2) inhibitors was successfully used to search Maybridge and Metabolite (MDL) databases [5]. A descriptor pharmacophore concept was introduced recently [6] on the basis of variable selection QSAR; the descriptor pharmacophore is defined as a subset of molecular descriptors that afford the most statistically significant QSAR model. It has been demonstrated that chemical similarity searches using descriptor pharmacophores as opposed to using all descriptors afford more efficient mining of chemical databases or virtual libraries to discover compounds with the desired biological activity [1, 6]. These recent developments emphasize the growing role of QSAR models for drug discovery by the means of database mining.

The process of QSAR model development can be generally divided into three stages: data preparation, data analysis, and model validation. These steps represent a standard practice of any QSAR modeling, and their implementations are generally determined by the researchers' interests, experience, and software availability.

The first stage includes the selection of a molecular dataset, calculation of molecular descriptors, and the choice of the QSAR approach in terms of the statistical

*To whom correspondence should be addressed. E-mail: alex_tropsha@unc.edu

methods of data analysis and correlation. The second part of QSAR modeling procedure involves building models that correlate descriptor values with those of biological activity. Many different algorithms and computer software are available for this purpose. Most are based on linear (multiple linear regression (MLR) with variable selection [7], partial least squares (PLS) [8], etc) as well as non-linear (e.g., *k*-nearest neighbors [9, 10], artificial neural networks [11]) methods. In all approaches, descriptors represent independent variables, and biological activities serve as dependent variables.

The final part of QSAR model development is the model validation [12], when the predictive power of the model and hence its ability to reproduce biological activities of untested compounds is established. Most of the QSAR modeling methods implement the leave-one-out (LOO) (or leave-some-out) cross-validation procedure. The outcome of this procedure is cross-validated $R^2(q^2)$, which is commonly regarded as an ultimate criterion of both robustness and predictive ability of the model. A widely used approach to establish the model robustness is so called y-randomization (randomization of response, i.e. in our case of activities) [13]. It consists of repeating the calculation procedure with randomized activities and subsequent probability assessment of the resultant statistics. Often, it is used along with the cross-validation. However, it is still uncommon to test QSAR models (characterized by a reasonably high LOO q^2) for their ability to predict accurately biological activities of compounds not included in the training set. Still, many authors claim that their models have high predictive ability in the absence of external validation [14–18]. Some researchers validate their models using only one or two compounds that were not used in QSAR model development [19, 20]. 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 the LOO cross-validated q^2 and the correlation coefficient R^2 between the predicted and observed activities for the test set, [21–23]. In a recent review, we have emphasized the importance of validation in developing reliable QSAR models [24].

In order to obtain a validated and, therefore, predictive QSAR model, an available dataset should be divided into the training and test sets. For the prediction statistics to be reliable, the test set must include at least five compounds [25]. Ideally, the division into the training and test set must satisfy the following three conditions [25]: (i) All representative compound-points of the test set in the multidimensional descriptor space must be close to those of the training set. (ii) All representative points of the training set must be close to those of the test set. (iii) The representative points of the training set must be distributed within the whole area occupied by the entire dataset. In this paper we consider three closely related methods for the rational division of a dataset into training and test sets based on sphere exclusion algorithms [26], which are widely used for comparison of chemical databases or chemical libraries [2]. These methods satisfy conditions (i) and (iii) automatically [25]. To estimate predictive power of a QSAR model, we employ a set of statistical characteristics introduced recently [12].

In this report, we describe the development and validation of the QSAR models for 48 functionalized amino acid anticonvulsants [27] and a series of 157 epipodophyllotoxin derivatives with antitumor activity [28]. Molconn-Z descriptors were used [29] and *k*-nearest-neighbors (*k*NN) QSAR method [9] was applied to build QSAR models. We argue that rational selection of training and test sets using sphere-exclusion algorithms leads to QSAR models with generally higher predictive ability than models based on alternative approaches to training and test set selection.

Methods

Descriptors

Descriptors for the development of QSAR models were obtained using Molconn-Z [29] software. Prior to the development of QSAR models, descriptors were normalized using the following formula,

$$X_{\mathrm{ij}}^{\mathrm{n}} = \frac{X_{ij} - X_{j,\min}}{X_{j,\max} - X_{j,\min}} \qquad (1)$$

where X_{ij} and X_{ij}^{n} are the non-normalized and normalized j-th ($j = 1,\ldots,\mathrm{K}$) descriptor values for compound i ($i = 1,\ldots,\mathrm{N}$), correspondingly, and $X_{j,\min}$ and $X_{j,\max}$ are the minimum and maximum values for the *j-th* descriptor. Thus, for all descriptors, $\min(X_{\mathrm{ij}}^{\mathrm{n}}) = 0$ and $\max(X_{\mathrm{ij}}^{\mathrm{n}}) = 1$. 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. After normalization, identical descriptors were deleted.

Division of a dataset into training and test sets

The following three sphere-exclusion algorithms were used to divide a set of N compounds into training and test sets [25].

1. Select a compound with the highest activity.
2. Include this compound into the training set.
3. Construct a sphere with radius $R = c(V/N)^{1/K}$ with the center in the representative point of this compound. Here, K is the number of descriptors (dimensionality of the descriptor space), and c is the dissimilarity level [25, 30]. Dissimilarity level was varied to construct different training and test sets.
4. Include compounds, corresponding to representative points within this sphere, except for the sphere's center, 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. If n > 0, in the case of *algorithm 1* select the next compound randomly and go to step 2. Otherwise 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 compounds left to the sphere centers. Select a compound with the smallest d_{ij} (*algorithm 2*) or largest d_{ij} (*algorithm 3*) that corresponds to the $\max_i \min_j d_{ij}$ and go to step 2.
7. Stop.

Frequently, data points are distributed unevenly in the descriptor space, and the point density in some areas may be much higher than in the others. Probe spheres in high density areas can include many points, and according to algorithms 1–3, only one of them will be assigned to the training set, and all other to the test set. This may make prediction of activities of compounds including into the test set unreliable, and the model may have low predictive power. To avoid this situation, we have developed modified versions of algorithms 1-3, in which more than one point inside the probe sphere is included into the training set. The modified procedure allows specifying the maximum number of compounds included in the test set successively, after which the next point is included into the training set. In this study, this number was equal to one. Further, the modified algorithms are referred to as algorithms 1M, 2M, and 3M.

Algorithm 4 is based on ranking of compound activities. The algorithm is as follows [25].

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. 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. Include this number of more active compounds of each group in the training set, and the remaining compounds in the test set.

Finally, *Algorithm 5* is merely the random division of a dataset into training and test sets.

To compare divisions of datasets into training and test sets, the following parameters of the training and test sets resulting from the division of the original dataset have been employed: 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 $M_{train,test}$ [25, 30]. $M_{test,train}$ is defined as follows. The volume corresponding to one representative point in multidimensional descriptor space is equal to V/N = 1/N, where N is the number of compounds. We construct spheres with radius $R = c(V/N)^{1/K}$ with centers in the representative points of the test set. Here, K denotes the number of descriptors (dimensionality of the descriptor space), and c is the dissimilarity level. Let N_a be the number of points of the test set, for which the spheres contain no points of the training set. Then

$$M_{test,train}(c) = N_a/N_{test}, \quad (2)$$

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

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

$$M_{train,test}(c) = N_b/N_{train}, \quad (3)$$

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 the test set points to the training set points. The lower $M_{test,train}$, the better the condition of closeness of the test set points to the training set points is satisfied. Training and test sets obtained with our algorithms 1–3 and 1M–3M with dissimilarity level c satisfy condition $M_{test,train}(c) = 0$ automatically. $M_{train,test}$ characterizes the closeness of points of the training set to those of the test

set. Lower $M_{train,test}$ correspond to the better division of compounds into the training and test sets. If training and test sets are obtained with a sphere-exclusion algorithm with dissimilarity level c and $N_{train} > N_{test}$ (which is true in the majority of QSAR studies), then the optimal value of $M_{train,test}(c) = (N_{train}\text{-}N_{test})/N_{train} > 0$. In this case each of the $(N_{train}–N_{test})$ points of the training set is close to exactly one point of the test set. If $N_{train} \leq N_{test}$ then the optimal value of $M_{train,test}$ (c) is zero. $M_{train,test}$ (c) for training and test sets obtained with algorithms 1M–3M are usually lower than those obtained with algorithms 1–3. Theoretically, if any of the modified algorithms 1M–3M is used, $M_{train,test}$ (c) can reach zero, even if $N_{train} > N_{test}$.

In order to characterize the diversity of the training set, we have employed the diversity index I_{train},[25, 30] which is defined as follows. We construct spheres with radius $R = c(V/N)^{1/K}$ with their centers at the points corresponding to the training set compounds. Let N_c be the number of points of the training set, for which the spheres contain no other points of the training set. Then

$$I_{train}(c) = N_c/N_{train}. \tag{4}$$

Algorithms 1–3 automatically provide the highest value of I_{train}(c), which is equal to one. Algorithms 1M–3M give lower values of I_{train} (c) because the local density of the training sets generated by these methods correlates with the local density of all representative points. Nevertheless, points of the training set are distributed within the whole area occupied by the representative points.

Diversity indices for training and test sets obtained with the algorithms 4 and 5 can not be estimated, since dissimilarity levels c for these methods are not defined. However, if the number of compounds in the training and test set are equal or close to those obtained by a sphere-exclusion algorithm with a certain dissimilarity level, this dissimilarity level can be used also for training and test sets obtained with algorithms 4 and 5. Theoretically, for algorithms 4 and 5 the optimal value for $M_{train,test}$ (c) may be lower than $(N_{train}\text{-}N_{test})/N_{train}$, nevertheless we will compare our $M_{train,test}$ (c) with this critical value.

It is possible to obtain minimum c, which corresponds to the optimal values of $M_{test,train} = 0$ and $M_{train,test} = 0$ and maximum c, corresponding to the optimum of $I_{train} = 1$.

$$c_{\min}(M_{test,train} = 0) = \frac{\max\limits_{i}\min\limits_{j} d_{ij}}{(1/N)^{1/K}} \tag{5}$$

$$c_{\min}(M_{train,test} = 0) = \frac{\max\limits_{j}\min\limits_{i} d_{ij}}{(1/N)^{1/K}} \tag{6}$$

$$c_{\max}(I_{train} = 1) = \frac{\min d_{ij}}{(1/N)^{1/K}} \tag{7}$$

In equations (5) and (6) $i = 1,2,\ldots, N_{test}$ and $j = 1,2,\ldots, N_{train}$ are the numbers of points of the test and training set, respectively, and in Equation (7) both i and j are the numbers of points of the training set.

k-nearest neighbors QSAR

k-nearest neighbors (kNN) QSAR method [9] uses LOO cross-validation procedure and a simulated-annealing algorithm for the descriptor selection. The procedure starts with the random selection of a predefined number of descriptors from all descriptors. Estimated activities $\hat{y}_i$ of compounds excluded by LOO procedure are calculated using the following formula

$$\hat{y}_i = \frac{\sum\limits_{j=1}^{k} a_j w_{ij}}{\sum\limits_{j=1}^{k} w_{ij}}, \tag{8}$$

where weights w_{ij} are defined as

$$w_{ij} = \left(1 - \frac{d_{ij}}{\sum\limits_{j'=1}^{k} d_{ij'}}\right), \tag{9}$$

and d_{ij} are the distances between compound i and its k nearest neighbors. After each run, cross-validated R^2 (q^2) is calculated

$$q^2 = 1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y})^2}, \tag{10}$$

where y_i, and $\bar{y}$ are the actual and average values of activity. The summation in (10) is performed over all compounds. After each run, a predefined small number of descriptors are randomly replaced by other descriptors from the original pool, and the new value of q^2 is obtained. If q^2 (new) $> q^2$ (old), the new set of descriptors is accepted. If q^2 (new) $\leq q^2$ (old), the new set of descriptors is accepted with probability $p = \exp(q^2 \text{ (new)} - q^2 \text{ (old)})/T$, or rejected with probability $(1 - p)$, where T is a simulated 'temperature' annealing parameter. During this process, T

is decreased until a predefined threshold. Thus, the optimal (highest) q^2 is achieved (see Ref. [9] for additional details). For the prediction, the final set of selected descriptors is used, and expressions (8) and (9) are applied to predict activities of compounds of the test set. Prediction is impossible, if a representative point of a compound is 'too far' from the k nearest neighbor points representing training set compounds with known activities. To limit the domain of model applicability, we use a distance cutoff value [9] D = <d>+Zσ, where <d> is the average of Euclidean distances between k nearest neighbors of all points of the training set used in model derivation, σ the standard deviation of these distances, and Z is the empirical parameter, which in this paper was varied from 0.5 to 3.0 with step 0.5.

To estimate the predictive power of a QSAR model, the following criteria were used [12]. (i) correlation coefficient R between the predicted and observed activities; (ii) coefficients of determination [31] (predicted versus observed activities R_0^2, and observed versus predicted activities R'^2_0); (iii) slopes k and k' of regression lines (predicted versus observed activities, and observed versus predicted activities) through the origin. We conclude that a QSAR model has an acceptable predictive power if the following conditions are satisfied [12]:

$$q^2 > 0.5; \tag{11}$$

$$R^2 > 0.6; \tag{12}$$

$$\frac{(R^2 - R_0^2)}{R^2} < 0.1 \text{ and } 0.85 \le K \le 1.15 \tag{13}$$

or

$$\frac{(R^2 - R_0^{12})}{R^2} < 0.1 \text{ and } 0.85 \le K' \le 1.15 \tag{14}$$

R_0^2 is a quantity characterizing linear regression with the Y-intercept set to zero (i.e., described by Y = kX, where Y and X are actual and predictive activity, respectively) which is different from conventional R^2 for the best fit linear regression (i.e., Y = aX + b). The reason why we introduce R_0^2 and require k (or k′) be close to one is that when one compares actual vs. predicted activity, an exact fit is required, not (just) a linear correlation. Thus, a model with k = 0.9 (i.e., Y = 0.9X) and high R_0^2 (e.g., 0.7) does have a high accuracy whereas a model described by Y = 0.8X + 2 and higher R^2 (e.g., 0.9), but low R_0^2 (e.g., 0.5) actually implies poor accuracy. Most authors ignore this caveat and present resulting statistics in the form of the best-fit linear regression between actual and predicted activities. In our opinion, this practice is insufficient, since in fact we are looking for models capable of *reproducing* the experimental data as opposed to producing predicted activity values that *correlate* with the experimental data. Earlier, we suggested [12], that preferably both values of R_0^2 and R'^2_0 must satisfy condition (13). This condition appears to be too strong; thus, herein we employed a less stringent condition: thus, we have required that the difference between R_0^2 and R'^2_0 should not be too high as follows:

$$|R_0^2 - R'^2_0| < 0.3. \tag{15}$$

Datasets

The QSAR analysis of a dataset of 48 functionalized amino acid anticonvulsant agents was published recently [27]. ED_{50} values for these compounds are given in Table 1 of Ref. [27]. In our study, the procedures based on the sphere-exclusion algorithms for dividing datasets into training and test sets (Section 2.2) were repeated with the dissimilarity levels between 0.8 and 2.2 and a step of 0.2. The number of descriptors selected by the *k*NN-QSAR procedure varied from 10 to 30 with step 2. All models with $q^2 > 0.5$ were validated using the corresponding test sets. For algorithm 4 (Section 2.2), calculations were performed for eight different training and test sets differing in the number of groups and the number of compounds belonging to the training and test sets in each group. The number of compounds in training and test sets was varied from 24 to 42 and from 24 to 6, respectively. Using algorithm 5 (random division of the dataset into the training and test sets), 15 different training and test sets were generated. The number of compounds in the test sets varied from 6 to 20 with step 2 (i.e., eight divisions into the training and test sets). Seven additional training and test sets were created with the number of compounds equal to those obtained with the algorithms 1 to 4. For each number of descriptors selected and each training and test set 10 QSAR models were obtained. Thus, for each training and test set the total number of QSAR models was 11 × 10 = 110. All calculations were repeated with randomized activities of the training set compounds as well to evaluate model robustness (y-randomization test [24]).

As a second example, we have considered a dataset of 157 epipodophyllotoxin derivatives with antitumor

Table 1. Predictive QSAR models built for a series of 48 functionalized amino acid anticonvulsant agents using algorithms 1–5.

Model No.	Algorithm for training/test set selection	Dissimilarity level	Training set size, cmpds	Test set size, cmpds	q^2	R^2	k or k'	R_0^2 or R'^2_0
1	1	0.8	42	6	0.62	0.81	0.97	0.80
2	1	1.0	42	6	0.71	0.67	0.98′	0.67′
3	2	0.8	42	6	0.63	0.73	0.97′	0.73′
4	2	1.0	41	7	0.68	0.80	0.91′	0.79′
5	2	1.6	34	14	0.72	0.63	0.94′	0.63′
6	3	0.8	42	6	0.61	0.81	0.95′	0.80′
7	3	1.0	42	6	0.59	0.94	0.94′	0.94′
8	3	1.2	40	8	0.54	0.79	0.99′	0.79′
9	4	–	42	6	0.73	0.87	0.99	0.86
10	4	–	40	8	0.74	0.81	1.07	0.79
11	4	–	39	9	0.70	0.72	1.03	0.70
12	4	–	35	13	0.67	0.69	0.95	0.67
13	5	–	40	8	0.58	0.93	1.06′	0.88′
14	5	–	39	9	0.72	0.72	1.00	0.72
15	5	–	38	10	0.80	0.68	0.96	0.63
16	5	–	34	14	0.56	0.75	0.98	0.75
17	5	–	33	15	0.77	0.62	0.97′	0.59′
18	5	–	28	20	0.64	0.62	0.97′	0.61′

Table 2. Predictive QSAR models built for a series of 156 epipodophyllotoxin derivatives using algorithms 1–5.

No.	Algorithm	Dissimilarity level	Training set	Test set	q^2	R^2	$k$or k'	R_0^2 or R'^2_0
1	1	0.4	144	12	0.58	0.65	0.98	0.65
2	1	0.6	137	19	0.56	0.78	0.97′	0.75′
3	2	0.4	144	12	0.53	0.68	0.96′	0.68′
4	2	0.6	135	21	0.58	0.67	0.97′	0.67′
5	3	0.6	134	22	0.58	0.73	0.98′	0.72′
6	3	0.8	116	40	0.61	0.67	0.99′	0.65′
7	4	–	144	12	0.52	0.85	0.98′	0.85′
8	4	–	139	17(16)	0.53	0.83	1.01′	0.82′
9	5	–	142	14(13)	0.56	0.65	0.95′	0.65′
10	5	–	139	17(16)	0.61	0.61	0.94′	0.61′
11	5	–	135	21(20)	0.58	0.68	0.95′	0.68′
12	5	–	134	22	0.52	0.67	0.97′	0.67′
13	5	–	130	26(25)	0.59	0.61	0.98′	0.59′
14	5	–	114	42(41)	0.65	0.62	0.97′	0.60′

activity. These compounds are topoisomerase II inhibitors. The QSAR analysis of this dataset was reported recently [28]. Activity data for these compounds are given in Tables 1 and 2 in Ref. [28]. A compound with the lowest activity (its activity was significantly lower than those of all other compounds) was excluded from our analysis. Thus, the total number of compounds in our investigation was 156. All procedures based on the sphere-exclusion algorithms for dividing datasets into training and test sets (Section 2.2) were repeated with the dissimilarity levels ranging from 0.4 to 1.0 with step 0.2. The number of descriptors selected by the *k*NN-QSAR procedure varied from 10 to 50 with step 4. All models with $q^2 > 0.5$ were validated using corresponding test sets. For algorithm 4 (activity based selection; cf. Section 2.2), calculations were performed for eight different training and test sets differing in the number of groups and the number of compounds in the training and test sets in each group. Thus, the number of compounds in training and test sets varied from 104 to 144 and from 52 to 12, respectively. For algorithm 5, 16 different training and test sets were generated. The number of compounds in the test sets varied from 12 to 52 with step 10 (five divisions into training and test sets). Eleven training and test sets were created additionally with the number of compounds equal to those obtained with algorithms 1 to 4. For each predefined number of descriptors 10 QSAR models were obtained. Thus, for each training and test set the total number of QSAR models built was $11 \times 10 = 110$. All calculations were repeated for the test sets with randomized activities as well.

Results

48 anticonvulsant agents

As expected, with the increase in the dissimilarity level from 0.8 to 2.2 (see algorithms 1–3 in Section 2.2), the number of compounds in the training sets was gradually decreasing from 42 (for all three algorithms) to 24 (algorithm 1), 25 (algorithm 2) and 27 (algorithm 3), and the number of compounds in the corresponding test sets was gradually increasing from 6 to 24, 23, and 21, respectively. For calculation of $M_{test,train}$, $M_{train,test}$ and I_{train} for training and test sets obtained with the algorithms 4 and 5, we used dissimilarity levels, with which training and test sets of the same size were generated by algorithms 1–3. For training and test sets generated with algorithms 4 and 5, $M_{test,train}$ values were significantly higher than zero and all I_{train} values were significantly lower than one. The lowest $M_{test,train}$ equal to 0.38 was obtained with the algorithm 4 when training and test sets included 35 and 13 compounds, respectively. The majority of other $M_{test,train}$ values varied between 0.6 and 0.8. The lowest I_{train} equal to 0.39 was obtained with algorithm 5 when the training set contained 28 compounds. The majority of other I_{train} values varied between 0.6 and 0.8. $M_{train,test}$ values for training and test sets generated with algorithms 1–3 and with algorithms 4 and 5 were similar to each other, except for the training and test set of 28 and 20 compounds, respectively, generated with the algorithm 5, for which $M_{train,test}(2.0) = 0.43$. This value is still significantly higher than the critical value, which in this case is equal to 0.28. However, as expected, for higher dissimilarity levels $M_{train,test}$ had lower values. $c_{min}(M_{test,train} = 0)$ (see Section 2.2) for training and test sets obtained with algorithms 1–3 must be lower than the dissimilarity levels used to obtain these training and test sets. c_{min} $(M_{test,train} = 0)$ for training and test sets obtained with algorithms 4 and 5 were significantly higher than these dissimilarity levels. $c_{max}(I_{train} = 1)$ for the training and test sets obtained with algorithms 1–3 must be higher than dissimilarity levels used to obtain these training and test sets. All $c_{max}(I_{train} = 1)$ for training and test sets obtained with algorithms 4 and 5 were around 0.5. This value corresponds to the distance between the closest points of the training set. On average, $c_{min}(M_{train,test} = 0)$ for training and test sets obtained with algorithms 1-3 were slightly higher than for those obtained with algorithms 4 and 5.

QSAR models with the highest predictive power are presented in Table 1. All compounds of the test sets were within the cutoff distance with $Z = 3.0$ (cf. Section 2.3) from their nearest neighbors in the corresponding training sets. However, for all models in Table 1 based on the training sets generated with algorithms 1–3 (except for Model 4), all compounds of the test sets were within much smaller cutoff distance with $Z = 0.5$ from their nearest neighbors in the corresponding training sets. Prediction for all compounds of the test set of Model 4 was possible with $Z = 1.0$. Using algorithm 2, for dissimilarity level of 0.5 another model with slightly worse statistics than that of Model 4 was found to satisfy all conditions (11)-(15): $q^2 = 0.63$, $R^2 = 0.80$, $k' = 0.95$, $R'^2_0 = 0.77$, and the corresponding Z value was 0.5. For all models based on the training sets generated with the algorithm 4, except for Model 10, all compounds in the test sets

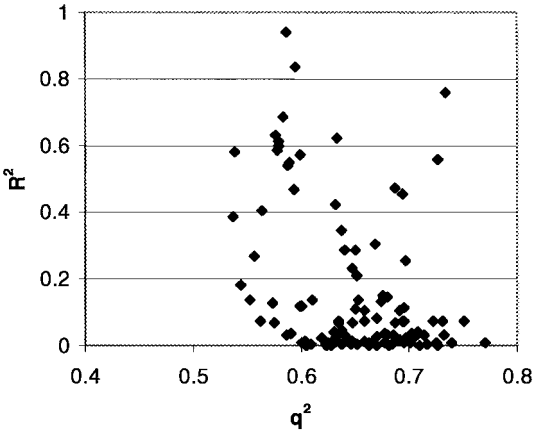

Figure 1. R^2 versus q^2 for anticonvulsant dataset. Models were built and validated using the training and test set obtained with algorithm 3 (Section 2.2). Dissimilarity level was equal to 1.0.

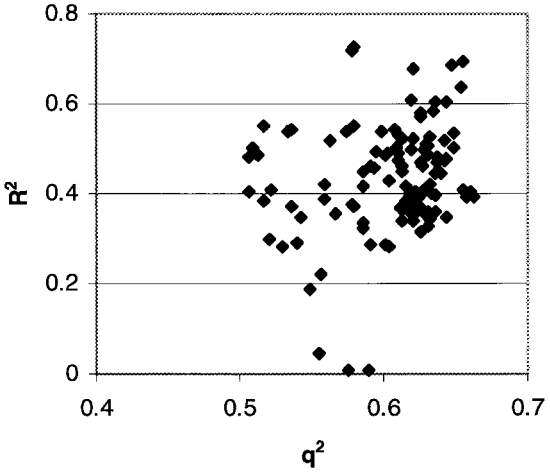

Figure 2. R^2 versus q^2 for anticancer agents. Models were built and validated using the training and test set obtained with algorithm 3 (Section 2.2). Dissimilarity level was equal to 0.6.

were within the cutoff distance with $Z = 0.5$ from their nearest neighbors in the corresponding training sets. For Model 10 all compounds of the test set were within the cutoff distance from their nearest neighbors of the corresponding training set only with $Z = 3.0$. The lowest Z value, when reliable predictions ($q^2 = 0.75$, $R^2 = 0.71$, $k' = 0.91$, $R'^2_0 = 0.70$) were obtained for all compounds in the test set used in Model 10, was 1.5. Only for one model (Model 14) based on the training sets generated by random division of compounds into training and test sets (algorithm 5) all compounds of the test set were within the cutoff distance with $Z = 0.5$ from compounds of the corresponding training set. For Models 13, 15, 16, 17 and 18, the lowest Z values were 2.5, 2.0, 1.5, 1.5, and 3.0, respectively. Minimal Z values, for which the models were able to predict all compounds of the corresponding test sets, were 1.0, 2.0, 1.0, 1.5 and 2.0, respectively. Thus, in general, rational selection of training and test sets (algorithms 1–3) afford more conservative estimates (i.e., lower Z value) and, therefore, higher accuracy of the predicted activity of the test set.

As in the studies of other datasets considered elsewhere [12, 21–23], there was no correlation between q^2 and R^2 values. Figure 1 shows a typical plot of R^2 versus q^2 for all models built with the training and test set obtained with the algorithm 3 and dissimilarity level of 1.0. All 110 of these models had $q^2 > 0.5$, seven of them had $R^2 > 0.6$, but only three of them satisfied all conditions (11)-(15). We note that for all training sets the number of models with $q^2 > 0.5$ was between 105 and 110. In total, 98 models satisfying conditions (11)-(15) were obtained. Calculations performed for training sets with randomized activities gave the following results. For all training sets the number of models with $q^2 > 0.5$ varied between 2 and 110. In total, only three models had $R^2 > 0.6$, and one of them satisfied all conditions (11)-(15): $q^2 = 0.52$, $R^2 = 0.75$, $k = 1.07$, $R^2_0 = 0.75$. Training and test set for this model were generated by algorithm 3; dissimilarity level was equal to 0.8, and the test set included 6 compounds. High q^2 values for datasets with randomized activities can be explained by a structural similarity (structural redundancy) of compounds in question, a chance correlation

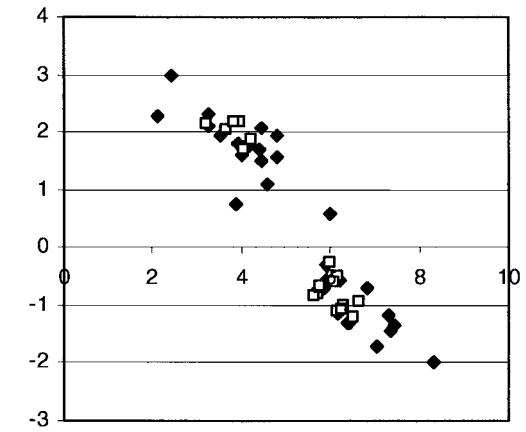

Figure 3. Projection of the anticonvulsant dataset in the descriptor space onto a plane defined by first two principal components obtained with the Singular Value Decomposition (SVD) method [35]. Training set (29 compounds) and test set (19 compounds) were obtained with the algorithm 3. Points of the training and test set are denoted by black diamonds and white squares, respectively.

between randomized and non-randomized activities, and a high total number of descriptors in comparison with the number of compounds. While there was no chance correlation between randomized and non-randomized activities, the majority of compounds had very similar structures (see Table 1 in Ref. [27]), and the total number of descriptors (before variable selection) was 157, which is much higher than the number of compounds. So there was a high probability of selecting descriptors which could afford high q^2 values. A small number of compounds in the test set increases the possibility of a chance correlation between the predicted and observed activities of these compounds. A chance correlation is probably responsible for the fact that one model had good statistics for the external prediction. The question arises, if under these conditions we are able to trust our predictive models. The answer to this question probably lies in the fact that the total number of predictive models found using real datasets was 98, i.e. the probability of a chance correlation for the external prediction was only about $1/98 = 1.02 \times 10^{-2}$. Since this probability is very low, our models with high predictive ability can be reliably used for external prediction. In our example, to be more confident in our results we must discard five models with good statistics based on the training set generated with the algorithm 3 with the dissimilarity level 0.8. It has been shown that most accurate predictions can be obtained by averaging predictions from multiple predictive QSAR models [27]. This approach was applied to predict activities of 13 compounds not included in the original dataset of 48 compounds (see Table 8 in Ref. [27]). All training and test sets in [27] were obtained with algorithm 2 with the random selection of the first compound of the training set.

156 epipodophyllotoxin derivatives

As in the previous example, with the increase of the dissimilarity level from 0.4 to 1.0 (see algorithms 1–3 in Section 2.2), the number of compounds in training sets was gradually decreasing from 144 (for all three algorithms) to 96 (algorithm 1), 95 (algorithm 2) and 99 (algorithm 3), and the number of compounds in the corresponding test sets was gradually increasing from 12 to 60, 61, and 57, respectively. For calculation of $M_{test,train}$, $M_{train,test}$ and I_{train} for the training and test sets obtained with algorithms 4 and 5, we used dissimilarity levels, with which training and test sets of the same size were generated by algorithms 1–3. As in the previous example, similar tendencies were

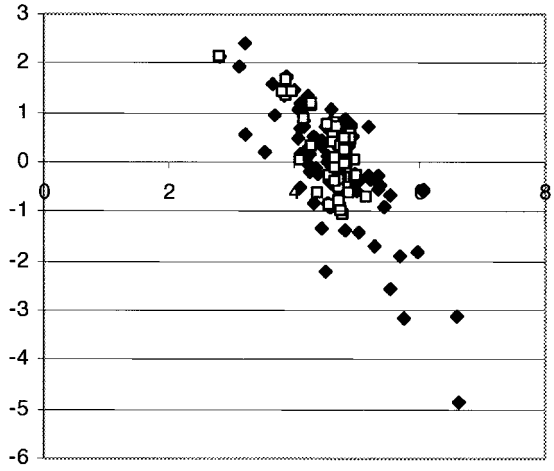

Figure 4. Projection of the anticancer dataset in the descriptor space onto a plane defined by first two principal components obtained with the SVD method [35]. Training set (95 compounds) and test set (61 compounds) were obtained with the algorithm 3. Points of the training and test set are denoted by black diamonds and white squares, respectively.

observed for parameters $M_{test,train}$, $M_{train,test}$ and I_{train}. For training and test sets generated with algorithms 4 and 5, $M_{test.train}$ values were significantly higher than zero, and all I_{train} values were significantly lower than 1. For higher c, $M_{train,test}$ (c) had lower values. For this dataset, $M_{train,test}$ had slightly lower values for the training and test sets generated with algorithms 1–3. However, $c_{min}(M_{train,test} = 0)$ were lower for the training and test sets generated with algorithms 4 and 5. As in the previous example, $c_{min}(M_{test,train} = 0) < c$ for training and test sets obtained with algorithms 1-3, and $c_{min}(M_{test,train} = 0) \gg c$ for training and test sets obtained with algorithms 4 and 5.

QSAR models with the highest predictive power are presented in Table 2. High predictive ability of models 1–6 (Table 2) was demonstrated by using these models to mine the National Cancer Institute (NCI) database [32]. The 15 most active compounds of the original dataset were used as probes. The Euclidean distances between the probes and the compounds in the databases were calculated in the multidimensional descriptor subspace defined by descriptors selected by the QSAR procedure. More than 10 hits were found that were different from the 157 compounds in the original dataset. All of them were independently reported as topoisomerase II inhibitors [see Refs. [28], [33] and [34]]. This result demonstrates the potential of the models to identify novel inhibitors in available

Table 3. Predictive QSAR models built for a series of 48 functionalized amino acid anticonvulsant agents using algorithms 1M–3M.

Model No.	Algorithm for training/test set selection	Dissimilarity level	Training set size, cmpds	Test set size, cmpds	q^2	R^2	k or k'	R_0^2 or R'^2_0
1	1M	0.9	43	5	0.61	0.81	0.97	0.75
2	1M	1.1	41	7	0.66	0.79	1.00	0.78
3	1M	1.5	40	8	0.74	0.88	0.90′	0.87′
4	1M	1.6	39	9	0.66	0.73	0.92′	0.72′
5	1M	1.7	37	11	0.66	0.86	1.03′	0.81′
6	1M	2.0	34	14	0.72	0.65	1.00′	0.63′
7	2M	0.9	43	5	0.76	0.78	1.01	0.77
8	2M	1.1	41	7	0.63	0.76	0.94′	0.75′
9	2M	1.5	40	8	0.68	0.72	1.01′	0.71′
10	2M	2.0	36	12	0.62	0.69	0.98′	0.67′
11	2M	2.9	29	19	0.59	0.63	1.02′	0.62′
12	3M	1.1	42	6	0.67	0.70	0.98′	0.69′
13	3M	1.4	41	7	0.75	0.76	0.96′	0.73′
14	3M	1.5	40	8	0.71	0.82	0.96′	0.82′
15	3M	1.7	38	10	0.61	0.83	0.99′	0.75′
16	3M	1.8	37	11	0.73	0.71	1.04′	0.69′
17	3M	2.0	36	12	0.63	0.72	1.04′	0.72′

databases and the high predictive power of our *k*NN QSAR models.

Z value for all the models included in Table 2 (Section 2.3) was equal to 3.0. If the prediction was not possible for all compounds (due to the model applicability limitation), the number of compounds of the test set for which prediction was possible is given in parentheses. For all models in Table 2 obtained for the training sets generated with algorithms 1–3, all compounds in the corresponding test sets were already within the cutoff distance corresponding to $Z = 0.5$. None of the models obtained for the training sets generated with the algorithm 4 was found for which all compounds from the corresponding test sets were within the cutoff distance corresponding to $Z = 0.5$. Model 7 afforded prediction of the test set with Z equal to 1.0. Model 8 enabled the activity prediction for 16 out of 17 compounds of the test set, and the lowest Z was 1.5. For none of the models for the training sets generated by random division of the dataset into training and test sets (algorithm 5) were all compounds of the test sets within the cutoff distance with $Z = 0.5$ from their nearest neighbors of the corresponding training sets. For Models 9–14, the prediction of activities for the number of compounds of the test sets listed in Table 2 was possible when the lowest Z values were equal to 2.0, 1.5, 2.0, 1.5, 1.0, and 2.5, respectively. Minimum Z value, for which a predictive model was found to be able to predict 20 out of 21 compounds of the test set used by Model 11, was 1.0 ($q^2 = 0.63$, $R^2 = 0.66$, $k' = 0.96$, $R'^2_0 = 0.66$). For the training and test sets utilized by Models 8–11, 13, and 14, we continued to increase Z values with step 0.5. We found that Model 8 predicts all 17 compounds of the test set, if $Z = 4.0$ ($q^2 = 0.53$, $R^2 = 0.82$, $k' = 1.01$, $R'^2_0 = 0.81$). Model 11 predicts all 21 compound of the test set, if $Z = 4.5$ ($q^2 = 0.58$, $R^2 = 0.65$, $k' = 0.94$, $R'^2_0 = 0.65$). For higher Z no one model was found to be predictive for all compounds of test sets corresponding to models 9, 10, 13 and 14. We conclude that sphere-exclusion algorithms provide closeness of the test set to the training set not only in the original descriptor space, but also in the subspace of descriptors selected by the *k*NN-QSAR procedure. Thus, the sphere-exclusion algorithms generate training and test sets for building predictive QSAR models with lower cutoff distances, i.e. higher confidence. This is one of the advantages of these algorithms.

As always, there was no correlation between q^2 and R^2 values. In Figure 2, R^2 versus q^2 is presen-

Table 4. Predictive QSAR models built for a series of 156 epipodophyllotoxin derivatives using algorithms 1M–3M.

No,	Algorithm	Dissimilarity level	Training set	Test set	q^2	R^2	k or k'	R_0^2 or R'^2_0
1	1M	0.5	140	16	0.64	0.84	1.02′	0.84′
2	1M	0.6	131	25	0.55	0.82	1.01′	0.81′
3	1M	0.7	127	29	0.61	0.71	0.99′	0.69′
4	1M	0.8	120	36	0.56	0.77	1.00′	0.77′
5	2M	0.2	145	11	0.66	0.80	0.96′	0.74′
6	2M	0.4	139	17	0.59	0.82	0.99′	0.79′
7	2M	0.6	125	31	0.54	0.81	1.01′	0.78′
8	3M	0.6	138	18	0.63	0.77	0.98′	0.77′
9	3M	0.7	131	25	0.57	0.79	0.99′	0.78′
10	3M	0.8	126	30	0.57	0.74	0.98′	0.74′
11	3M	0.9	120	36	0.52	0.67	0.99′	0.67′
12	3M	1.2	105	51	0.60	0.63	0.98′	0.60′

ted for all models built with the training and test set obtained with algorithm 3 using the dissimilarity level 0.6. Although 109 out of 110 of these models had $q^2 > 0.5$, only nine of them had $R^2 > 0.6$ and satisfied all conditions (11)-(15). Calculations performed for training sets with randomized activities gave no models with $q^2 > 0.5$. For this dataset, the number of compounds was much higher than in the previous example, and it was comparable to the total number of descriptors, which in this case was equal to 131. Thus, the chance correlation between predicted and observed activities in LOO cross-validation procedure was much lower than in the previous example.

Discussion

In this report, we have presented rational approach to the selection of training and test sets for the development of validated, predictive QSAR models. We have demonstrated that for the majority of models built and validated using training and test sets generated with the sphere-exclusion algorithms, all compounds of the test sets were within cutoff distance with Z=0.5 from their nearest neighbors of the training sets, even when test sets included just over 20 (anticonvulsants) or about 60 compounds (anticancer agents). However, no predictive models were found for these test sets. We can explain this result, if we take into account the differences in the local density of the distribution of compounds-points in the descriptor space. When the sphere-exclusion algorithm is employed, the representative points of the test set are concentrated in the areas with the high point density. If the dissimilarity level were high, they would constitute the majority of all points in these parts of the descriptor space (Figures 3 and 4). In this case, the accurate prediction of activities of compounds from these areas would become impossible, particularly if the activities of compounds vary significantly. On the other hand, for the algorithm based on the activity range or the random division representative points of the test set may cover a bigger area: as we mentioned above, training and test sets generated with algorithms 4 and 5 were characterized by lower $c_{\min}(M_{\text{train,test}} = 0)$ than those generated by algorithms 1–3. This disadvantage of sphere-exclusion algorithms can be easily overcome. Thus, we have modified the original algorithms 1–3 to construct algorithms 1M–3M (cf. Methods section). Calculations were performed with dissimilarity levels starting from 0.2 with step 0.1. Statistical characteristics of the models with the highest predictive power are presented in Tables 3 and 4 for the anticonvulsant and epipodophyllotoxin datasets, respectively. Comparison of the results presented in Tables 1 and 2 with those in Tables 3 and 4 shows that the modified algorithms 1M–3M have higher predictive power than the original algorithms 1–3. Model 12 in Table 4 showed satisfactory predictive ability for the test set of as many as 51 compounds. In Tables 3 and 4, all predictions for external test sets were made with the cutoff value $Z = 0.5$. Predictions with higher cutoff values are generally less reliable than those with lower

cutoff values. Thus, predictions for test sets of similar size obtained with algorithms 1M–3M were more reliable than those obtained with algorithms 4 and 5. The ability to validate a model using all compounds of the test set is the obvious advantage of the rational division of a dataset into training and test sets. Additional algorithmic developments are required to divide datasets into training and test sets taking into account local densities of representative point distribution in the descriptor space. For instance, probe spheres with smaller radii can be used for the areas of higher density, and spheres with higher radii can be used for areas of lower density.

Conclusions

In this paper we have considered a problem of QSAR model validation. Based on the study of several datasets in this and one of our previous reports [12], as well as on the results of other authors [21–23] we confirm that LOO cross-validated $R^2(q^2)$ alone can not be used as a reliable characteristic of the predictive power of a QSAR model. The only way to evaluate the predictive ability of a QSAR model is its validation using an external test set of compounds (i.e. those, which were not included in the training set) with known activities. We argue that training and test sets must satisfy the following three conditions. (i) All representative points of the test set in the multidimensional descriptor space must be close to representative points of the training set. (ii) All (or the absolute majority) of representative points of the training set must be close to representative points of the test set. (iii) The representative points of the training set must be distributed within the whole area of the descriptor space covered by the dataset.

We have used three sphere-exclusion algorithms, and two alternative algorithms based on the range of activities and on the random division of datasets into training and test sets, and we employed the variable selection *k*NN-QSAR procedure [9] to build QSAR models. We have applied our approach to a series of 48 functionalized amino acid anticonvulsant agents [27] and a dataset of 157 epipodophyllotoxin derivatives with antitumor activity [28]. Our major observations are as follows.

- Results obtained for both examples demonstrate that all three sphere-exclusion algorithms are equivalent in terms of the prediction power of QSAR models built and validated using the training and test sets;
- representative points of compounds of the test sets generated with sphere-exclusion algorithms are close to the points of the corresponding training sets not only in the original descriptor space, but also in the subspace of descriptors selected by the *k*NN-QSAR procedure. This allows prediction of activities of test set compounds with lower cutoff distances (i.e., higher confidence) between them and their nearest neighbors of the training set than for the test sets obtained with the algorithm based on the range of activities or with the random division of a dataset into training and test sets.
- training and test sets obtained by any procedure considered in this paper do not satisfy condition (ii) above;
- if a training and test set are generated using a sphere-exclusion algorithm with low dissimilarity level (test sets are small in comparison with the corresponding training sets), the training set models predict activities of the test set compounds better than those based on the training sets of the same size obtained with alternative approaches based on the range of activity or on the random division of a dataset into training and test sets;
- if test sets are comparable in size with the corresponding training sets, prediction of activities of compounds of the test sets is better, if training and test sets are generated by the activity-range algorithm or by random division of a dataset into training and test sets.

The latter disadvantage of the original sphere-exclusion algorithms was explained by the uneven distribution of representative points of the datasets in the descriptor space. To overcome this shortcoming, sphere-exclusion algorithms have been modified: if the number of points inside probe spheres was higher than 2, more than one point from inside of probe-spheres was included in the training set. New models developed and validated using training and test sets obtained with the modified algorithms appeared to have better statistics than those based on the training and test sets obtained with algorithms 4 and 5.

In summary, we consider that a QSAR model is validated, if

- External test set includes at least five compounds and both training and test sets satisfy conditions (i)-(iii) above.
- The model satisfies all conditions (11)-(15)
- The probability of chance correlation is very low: generally, none of the models built with random-

ized activities of the training set satisfies conditions (11)-(15).

These conditions are general and independent of the types of descriptors and optimization algorithms used in QSAR model development. We suggest them as general guidelines that should be followed by QSAR practitioners to increase the reliability and predictive power of their models.

Acknowledgements

This research was supported in part by the NIH research grant MH60328 awarded to AT.

References

1. Tropsha, A., Cho, S. J., and Zheng, W. In: *Rational Drug Design: Novel Methodology and Practical Applications* (Parrill, A.L. and Reddy, M.R., Eds), ACS Symposium Series No 719, 1999, pp. 198–211.
2. Cho, S.J., Zheng, W., Tropsha, A.J., Chem. Inf. Comput. Sci., 38 (1998) 259.
3. Reynolds, C.H., Druker, R., Pfahler, L.B. J. Chem. Inf. Comput. Sci. 38 (1998) 305.
4. Gussio, R., Pattabiraman, N., Kellogg, G.E., Zaharevitz, D.W., Methods 14 (1998) 255.
5. Belkina, N.V., Skvortsov, V.S., Ivanov, A.S., Archakov, A.I., Vopr. Med. Khim. 44 (1998) 464.
6. Tropsha, A., Zheng, W., Curr. Pharm. Des. 7 (2001) 599.
7. Clementi, S., Wold, S. In: van de Waterbeemd, H. (Ed.), *Chemometrics Methods in Molecular Design*, VCH, 1995, pp. 319–338.
8. Wold, S. In: van de Waterbeemd, H. (Ed.), *Chemometrics Methods in Molecular Design*, VCH, 1995, pp. 195–218.
9. Zheng, W., Tropsha, A., J. Chem. Inf. Comput. Sci. 40 (2000) 185.
10. Hoffman, B., Cho, S.J., Zheng, W., Wyrick, S., Nichols, D.E., Mailman, R.B., Tropsha, A., J. Med. Chem. 42 (1999) 3217.
11. Ajay, A., J. Med. Chem. 36 (1993) 3565.
12. Golbraikh, A., Tropsha, A., J. Mol. Graphics Mod. 20 (2002) 269.
13. Wold, S., Eriksson, L. In: *Chemometrics Methods in Molecular Design*, van de Waterbeemd, H. (Ed.), VCH, 1995, pp. 309–318.
14. Gironés, X., Gallegos, A., Ramon, C.-D., J. Chem Inf. Comput. Sci. 46 (2000) 1400.
15. Bordás, B., Kömíves, T., Szántó, Z., Lopata, A., J. Agricult. Food Chem. 48 (2000) 926.
16. Fan, Y., Shi, L.M., Kohn, K.W., Pommier, Y., Weinstein, J.N., J. Med. Chem. 44 (2001) 3254.
17. Randić, M., Basak, S.C., J. Chem. Inf. Comput. Sci. 40 (2000) 899.
18. Suzuki, T., Ide, K., Ishida, M., Shapiro, S., J. Chem. Inf. Comput. Sci. 41 (2001) 718.
19. Recanatini, M., Cavalli, A., Belluti, F., Piazzi, L., Rampa, A., Bisi, A., Gobbi, S., Valenti, P., Andrisano, V., Bartolini, M., Cavrini, V., J. Med. Chem. 43 (2000) 2007.
20. Morón, J.A., Campillo, M., Perez, V., Unzeta, M., Pardo, L., J. Med. Chem. 43 (2000) 1684.
21. Kubinyi, H., Hamprecht, F.A., Mietzner, T., J. Med. Chem. 41 (1998) 2553.
22. Novellino, E., Fattorusso, C., Greco, G., Pharm. Acta Helv. 70 (1995) 149.
23. Norinder, U., J. Chemomet. 10 (1996) 95.
24. Tropsha, A., Gramatica, P., Gombar, V., Quant. Struct. Act. Relat. (2002) (in press).
25. Golbraikh, A., Tropsha, A., J. Comput.-Aided Molec. Des., 16 (2002) 357.
26. Snarey, M., Terrett, N.K., Willett, P., Wilton, D.J., J. Mol. Graphics Mod. 15 (1997) 372.
27. Shen, M., LeTiran, A., Xiao, Y.-D., Golbraikh, A., Kohn, H., Tropsha, A., J. Med. Chem. 45 (2002) 2811.
28. Xiao, Z., Xiao, Y.-D., Feng, A., Golbraikh, A., Tropsha, A., Lee, K.-H., J. Med. Chem. 45 (2002) 2294.
29. Molconn-Z. http://www.eslc.vabiotech.com/
30. Golbraikh, A., J. Chem. Inf. Comput. Sci. 40 (2000) 414.
31. Sachs, L. Applied Statistics. A Handbook of Techniques. Springer-Verlag, 1984, p. 349.
32. Xiao, Z. Design and Synthesis of Etoposide-Related Topo II Inhibitors by Conventional and Computational Approaches. Ph.D. Dissertation. The University of North Carolina at Chapel Hill, 2003.
33. Zhang, Y., Lee, K.H., Chin. Pharm. J. 46 (1994) 319.
34. Cho, S.J., Tropsha, A., Suffness, M., Cheng, Y.C., Lee, K.H., J. Med. Chem. 39 (1996) 1383.
35. Xie, D., Tropsha, A., Schlick, T., J. Chem. Inf. Comput. Sci., 40 (2000) 167.