ELSEVIER

Journal of Molecular Graphics and Modelling 20 (2002) 269–276

Journal of Molecular Graphics and Modelling

www.elsevier.com/locate/JMGM

Beware of q^2!

Alexander Golbraikh, Alexander Tropsha*

Laboratory for Molecular Modeling, Division of Medicinal Chemistry and Natural Products, School of Pharmacy, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599, USA

Received 27 March 2001; accepted 6 April 2001

Abstract

Validation is a crucial aspect of any quantitative structure–activity relationship (QSAR) modeling. This paper examines one of the most popular validation criteria, leave-one-out cross-validated R^2 (LOO q^2). Often, a high value of this statistical characteristic ($q^2 > 0.5$) is considered as a proof of the high predictive ability of the model. In this paper, we show that this assumption is generally incorrect. In the case of 3D QSAR, the lack of the correlation between the high LOO q^2 and the high predictive ability of a QSAR model has been established earlier [Pharm. Acta Helv. 70 (1995) 149; J. Chemomet. 10 (1996) 95; J. Med. Chem. 41 (1998) 2553]. In this paper, we use two-dimensional (2D) molecular descriptors and *k* nearest neighbors (*k*NN) QSAR method for the analysis of several datasets. No correlation between the values of q^2 for the training set and predictive ability for the test set was found for any of the datasets. Thus, the high value of LOO q^2 appears to be the necessary but not the sufficient condition for the model to have a high predictive power. We argue that this is the general property of QSAR models developed using LOO cross-validation. We emphasize that the external validation is the only way to establish a reliable QSAR model. We formulate a set of criteria for evaluation of predictive ability of QSAR models.

Keywords: QSAR modeling; LOO cross-validation; Training and test sets; *k*NN QSAR

1. Introduction

Rapid development of combinatorial chemistry and high throughput screening methods in recent years has significantly increased a bulk of experimental structure–activity relationship (SAR) datasets. These developments have emphasized a need for reliable analytical methods for biological SAR data examination such as quantitative SAR (QSAR). QSAR has been traditionally perceived as a means of establishing correlations between trends in chemical structure modifications and respective changes of biological activity [1]. However, in many cases of chemical library design, the number of compounds that could be practically synthesized and tested is much smaller than the total size of exhaustive virtual chemical libraries. There is a need for developing virtual library screening tools, and QSAR modeling can be adapted to the task of targeted library design [2–4]. Of course, any QSAR modeling should ultimately lead to statistically robust models capable of making accurate and reliable predictions of biological activities of compounds. However, the application of QSAR models for virtual screening places a special emphasis on statistical significance and predictive ability of these models as their most crucial characteristics. This paper examines the validity of one of the most popular criteria of QSAR model predictive ability, leave-one-out cross-validated R^2 (LOO q^2).

The process of QSAR model development can be generally divided into three stages: data preparation, data analysis, and model validation. The first stage includes selection of a molecular dataset for QSAR studies, calculation of molecular descriptors, and selection of a QSAR (statistical analysis and correlation) method. These steps represent a standard practice of any QSAR modeling, and their specific details are generally determined by the researchers' interests and software availability.

The second part of QSAR model development consists of an application of statistical approaches for QSAR model development. Many different algorithms and computer software are available for this purpose. Most are based on linear (multiple linear) regression with variable selection [5], partial least squares (PLS) [6], etc.) as well as non-linear (genetic algorithms [7], artificial neural networks [8], etc.) methods. In all approaches, descriptors serve as independent variables, and biological activities as dependent variables.

The last and as we emphasize in this paper, most important part of QSAR model development is the model *validation*. Most of the QSAR modeling methods implement the leave-one-out (or leave-some-out) cross-validation procedure. The outcome from the cross-validation procedure is

* Corresponding author. Tel.: +1-919-966-2955; fax: +1-919-966-0204.
E-mail address: tropsha@email.unc.edu (A. Tropsha).

1093-3263/02/$ – see front matter
PII: S1093-3263(01)00123-1

cross-validated R^2 (q^2), which 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 that the model is highly predictive. A widely used approach to establish the model robustness is so-called *y*-randomization (randomization of response, e.g. biological activities) [9]. It consists of repeating the calculation procedure with randomized activities and subsequent probability assessment of the resultant statistics. Often, it is used along with cross-validation. Sometimes, models are tested for their ability to predict accurately the activity of one or two compounds that were not used in model development (see, for instance [10,11]). However, it is still common not to test QSAR models (characterized by a reasonably high LOO q^2) for their ability to predict accurately biological activities of compounds from an external test dataset, i.e. those compounds, which were not used for the model development.

Although, the low value of q^2 for the training set can indeed serve as an indicator of a low predictive ability of a model, the opposite is not necessarily true. Indeed, the high q^2 does not imply automatically a high predictive ability of the model. In order to both develop the model and validate it, one needs to split the whole available dataset into the training and test set. Several methods can be used for this purpose. They include random selection, selection by groups of compounds where the whole group is included into a training or a test set, selection of training set compounds with major features varying in a systematic way, etc. [12]. Not all of these methods produce sufficiently reliable models. In fact, the lack of correlation between the high value of the training set q^2 and the high predictive ability of a QSAR model has been noticed earlier in the case of 3D QSAR [13–15]. These studies indicated that while high q^2 is the necessary condition for a model to have a high predictive power, it is not a sufficient condition. Apparently, the only way to estimate the true predictive power of a model is to test it on a sufficiently large collection of compounds from an external test set.

In this paper, using several published datasets for the analysis, we argue that the external validation is an absolute requirement for the development of a truly predictive QSAR model. As the first example, we consider a well-known group of ligands of corticosteroid binding globulin [16]. This dataset is frequently referred to as a benchmark [17] for the development and testing of novel QSAR methods. In [13], many 3D QSAR models have been built based on the divisions of this dataset into training and test sets and no relationship between high q^2 and predictive R^2 values was found. In this paper, we employ the *k nearest neighbors* (*k*NN) QSAR variable selection method that was recently developed in this laboratory [18]. *k*NN QSAR uses 2D descriptors of chemical structures such as connectivity indices and atom pairs. We show that the application of this approach to the steroid dataset [16] leads to the same observations as using 3D QSAR: high q^2 does not automatically imply a high predictive power of the model. We also consider 2D QSAR models built for two other examples: a set of 78 ecdysteroids [19] and 66 Histamine H_1 receptor ligands [20]. In all these cases, we consider training and test sets as they were defined in the original publications. We demonstrate the lack of any relationship between high q^2 and predictive R^2 in all cases. The lack of this relationship appears to be the common feature of the QSAR methods that must be always taken into account in QSAR studies.

On the basis of our analysis, we suggest that the test set must include no less than five compounds, whose activities and structures must cover the range of activities and structures of compounds from the training set. This requirement is necessary for obtaining reliable statistics for comparison between the observed and predicted activities for these compounds. We reason that in addition to a high q^2 a reliable model should be also characterized by a high correlation coefficient R (or R^2) between the predicted and observed activities of compounds from a test set. Finally, we introduce a notion of the "ideal" QSAR model and formulate a set of criteria for a reliable QSAR model based on its closeness to the "ideal" model.

2. Methods

2.1. Descriptors

The following Molconn-Z [21] descriptors were used to develop QSAR models: simple and valence path, cluster, path/cluster and chain molecular connectivity indices [22–24], kappa molecular shape indices [25,26], topological [27] and electrotopological state indices [28–31], differential connectivity indices [32], graph's radius and diameter [33], Wiener [34] and Platt [35] indices, Shannon [36] and Bonchev et al. [37] information indices, counts of different vertices [21], counts of paths and edges between different kinds of vertices [21]. Since datasets considered in this paper contained chiral molecules, chirality descriptors developed recently [38] were added to the Molconn-Z [21] descriptors. The chirality descriptors [38] included modified Zagreb group indices [39], molecular connectivity indices [22–24], extended connectivity indices [40] and overall connectivity indices [41,42].

In each case, descriptors were scaled according to the following formula:

$$X_{ij}^{n} = \frac{X_{ij} - X_{j,\min}}{X_{j,\max} - X_{j,\min}},$$

where X_{ij} and X_{ij}^{n} are the non-scaled and scaled *j*th descriptor values for compound *i*, respectively, and $X_{j,\min}$ and $X_{j,\max}$ are the minimum and maximum values for *j*th descriptor, respectively. Thus, for all descriptors, $\min(X_{ij}^{n}) = 0$ and $\max(X_{ij}^{n}) = 1$.

2.2. kNN QSAR

*k*NN variable selection QSAR method developed in this laboratory earlier [18] was used to build QSAR models for all datasets. *k*NN QSAR optimizes the descriptor selection to achieve a model with the highest LOO q^2 as follows. To initiate the procedure, the following input should be provided:

- the number of variables (descriptors) D to be selected from the whole set of descriptors for the final best model;
- the maximum number k of nearest neighbors;
- the number of descriptors M to be changed at each step of the stochastic descriptor sampling procedure that utilizes simulation annealing;
- the starting $T_{\max}$ and ending $T_{\min}$ of the simulation annealing parameter, "temperature", T, and the factor $d < 1$ to decrease T ($T_{\text{next}} = dT_{\text{previous}}$) at each step;
- the number of times N the calculations must be performed before lowering T, if q^2 is not improved.

In all calculations reported in this work, $k = 5$, $T_{\max} = 100$, $T_{\min} = 10^{-9}$, $d = 0.9$, and $M = 3$. D was varied from 10 to 40 with step 2.

In the LOO cross-validation procedure, every compound was eliminated from the dataset once and its activity was then predicted as a weighted average of the activities of its nearest neighbors using the following formula:

$$\hat{y} = \frac{\sum_{\text{nearest neighbors}} y_i \exp(-d_i)}{\sum_{\text{nearest neighbors}} \exp(-d_i)}, \tag{1}$$

where d_i are the distances between this compound and its *k*NN. LOO q^2 was calculated according to the following expression:

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

where y_i are the actual activities, $\hat{y}_i$ are defined by Eq. (1), and $\bar{y}$ the average activity. The summation in Eq. (2) is performed over all compounds. The following algorithm was used to derive QSAR models for the examples considered in this paper.

1. Set $T = T_{\max}$.
2. Select randomly a subset of D descriptors.
3. For each compound, predict its activity using Eq. (1).
4. Select the number of nearest neighbors, which gives the highest q^2 (Eq. (2)).
5. Exchange $M \ll D$ descriptors for the same number of descriptors selected randomly out of all descriptors.
6. Perform steps 3 and 4 for the new descriptor set defined in step 5.
7. If the new q^2 (q^2_{new}) is higher than the previous one (q^2_{old}), accept the new set of descriptors and go to step 5. Otherwise, accept it with the probability $P = \exp[-(q^2_{\text{old}} - q^2_{\text{new}})/T]$ and go to step 5, or reject it with the probability $(1 - P)$, and go to step 8.
8. If q^2 did not change after step 5 has been performed N times for current T, and if $T > T_{\min}$, decrease T and go to step 5, and if $T \leq T_{\min}$ stop. If step 5 has been performed less than N times for the current q^2, go to step 5.

Thus, the output from the procedure is a QSAR model, which is characterized by the set of D descriptors selected, the number k of nearest neighbors, and the value of q^2.

The activities for the test set compounds are predicted based on selected training set model as follows.

1. For each compound of the test set, find its closest *k*NN from the training set using the value of k and descriptors selected by the underlying *k*NN QSAR model.
2. Predict activities of compounds from the test set using Eq. (1).

2.3. Estimation of the predictive ability of a QSAR model

Let us first define quantitative criteria of a predictive QSAR model. Let $\tilde{y}_i$ and y_i be the predicted and actual activities, respectively. If we plot y versus $\tilde{y}$ for the *ideal QSAR model*, the regression line will bisect the angle formed by positive directions of the orthogonal axes $\tilde{y}$ and y. The regression line can be described by expression $y^r = a\tilde{y} + b$, where [43]

$$a = \frac{\sum (y_i - \bar{y})(\tilde{y}_i - \bar{\tilde{y}})}{\sum (\tilde{y}_i - \bar{\tilde{y}})^2} \tag{3a}$$

and

$$b = \bar{y} - a\bar{\tilde{y}}. \tag{3b}$$

In Eqs. (3a) and (3b), $\bar{\tilde{y}}$ and $\bar{y}$ are the average values of the predicted and observed activities, respectively, and the summations in this and all the following equations are over all n compounds of the test set.

For the ideal model, the slope a is equal to 1, intercept b is equal to 0, and correlation coefficient:

$$R = \frac{\sum (y_i - \bar{y})(\tilde{y}_i - \bar{\tilde{y}})}{\sqrt{\sum (y_i - \bar{y})^2 \sum (\tilde{y}_i - \bar{\tilde{y}})^2}}, \tag{4}$$

for the regression of $\tilde{y}$ versus y is equal to 1. A *real QSAR model* may have a high predictive ability, if it is close to the ideal one. This may imply that the correlation coefficient R between the actual y and predicted $\tilde{y}$ activities must be close to 1 and regressions of y against $\tilde{y}$ or $\tilde{y}$ against y through the origin, i.e. $y^{r0} = k\tilde{y}$ and $\tilde{y}^{r0} = k'y$, respectively, should be characterized by at least either k or k' close to 1. Slopes k and k' are calculated as follows:

$$k = \frac{\sum y_i \tilde{y}_i}{\sum \tilde{y}_i^2}, \tag{5a}$$

$$k' = \frac{\sum y_i \tilde{y}_i}{\sum y_i^2}. \tag{5b}$$

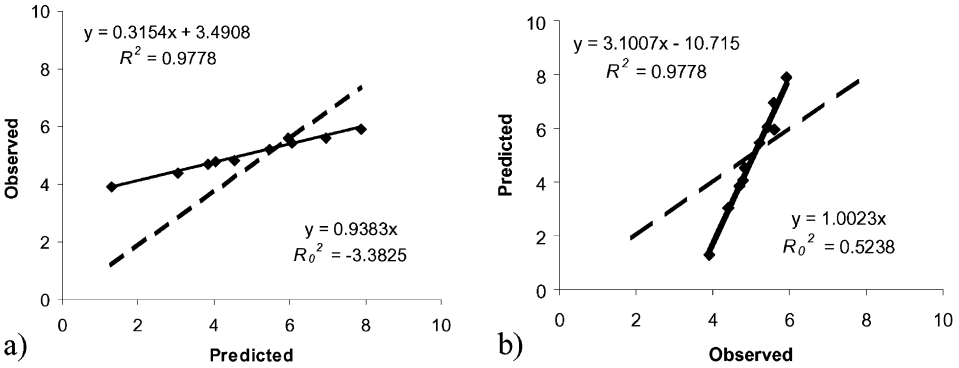

Fig. 1. An example of a regression between observed vs. predicted (a) and predicted vs. observed (b) activities for compounds from an external test set. Despite high R^2 value and both k and k' (cf. text) close to 1, the model is not highly predictive, because the regressions through the origin of the coordinate system are not close to the optimal regressions $y^r = a\tilde{y} + b$ and $\tilde{y}^r = a'y + b'$. Note that R_0^2 and $R_0'^2$ are substantially different from each other.

Plotting both y against $\tilde{y}$ and $\tilde{y}$ against y may appear redundant; however, we shall see that these plots could be characterized by different statistics.

We shall show now that the criteria formulated above may not be sufficient for a QSAR model to be truly predictive. Fig. 1 examines a case when the correlation between the actual activities and those predicted by a QSAR model for an external test set is characterized by $R^2 = 0.98$, and both k and k' are close to 1. Despite these good statistics, the predictions are inaccurate. Thus, regression lines through the origin defined by $y^{r0} = k\tilde{y}$ and $\tilde{y}^{r0} = k'y$ (with the intercept set to 0) are not close to the optimum regression lines $y^r = a\tilde{y} + b$ and $\tilde{y}^r = a'y + b'$. Furthermore, both correlation coefficients for these lines R_0^2 and $R_0'^2$ have different values, which are quite different from that of R^2. R_0^2 and $R_0'^2$ are calculated as follows:

$$R_0^2 = 1 - \frac{\sum(\tilde{y}_i - y_i^{r0})^2}{\sum(\tilde{y}_i - \bar{\tilde{y}})^2}, \tag{6a}$$

$$R_0'^2 = 1 - \frac{\sum(y_i - \tilde{y}_i^{r0})^2}{\sum(y_i - \bar{y})^2}, \tag{6b}$$

where $y_i^{r0} = k\tilde{y}_i$ and $\tilde{y}_i^{r0} = k'y_i$, and the summations are over all n compounds in the test set.

The values of all coefficients in both regressions $y^r = a\tilde{y} + b$ and $\tilde{y}^r = a'y + b'$ are far from the ideal ones: a and a' are not close to 1, and b and b' are not close to 0.

This example demonstrates that we need to impose an additional, more strict condition for the QSAR model to have a high predictive ability: both R^2 and either R_0^2 or $R_0'^2$ must have similar values. In fact, it can be shown that $R^2 \geq \max(R_0^2, R_0'^2)$. The values of b and b' can be actually significantly different from 0 even for good models (for instance, see Fig. 2). If the angle between regression lines $y^r = a\tilde{y} + b$ and $y^{r0} = k\tilde{y}$ is small, then these lines are close to each other in the area close to their intersection. The permissible values of these angles for good models depend on the range of activities in the test set. The larger the range, the smaller this angle must be, and the closer a or a' must be to k or k'. Closeness of R to 1, R^2 to either R_0^2 or $R_0'^2$ and the corresponding slope k or k' to 1, guarantee that a and a' are sufficiently close to 1, and no additional condition is necessary.

Residual variance (or residual mean square error) are calculated as follows:

$$s_{\text{res}}^2 = \frac{\sum(\tilde{y}_i - y_i^r)^2}{n - 2}, \tag{7}$$

where s_{res} is the standard deviation of the predicted activities for given actual activities. The square of the deviation of the mean of the regression line is obtained as

$$s_r^2 = \sum(y_i^r - \bar{y})^2 \tag{8}$$

and F-ratio is calculated according to the formula

$$F = \frac{s_r^2}{s_{\text{res}}^2}. \tag{9}$$

s_{res}^2 is also referred to as residual mean square error (RMSE). s_{res} is standard error of estimation [43]. In Eqs. (7) and (8), y_i^r correspond to the equation of regression $y^r = a\tilde{y} + b$.

In most publications where F-ratio is calculated, authors assume that the higher F-value, the better is the model. This assumption is correct, if F-values with the same numbers of degrees of freedom [43] are compared. For simple regression, the degrees of freedom have values of 1 and $n - 2$ [43]. Most authors do not mention that F-ratio is used in hypothesis testing. Briefly, null hypothesis H_0 assumes that the model does not predict better than the average activity value. H_1 hypothesis is based on the opposite assumption. Usually H_1 is accepted or rejected with a certain significance level α, that means that the probability of H_1 being true or false is at least α. Typically, the significance level of 0.95 or 0.99 is used. To find, whether H_1 can be accepted with a certain significance level, F-distribution function is used. If F-ratio appears to be higher than the corresponding value of F-distribution function for given degrees of freedom, H_1 is

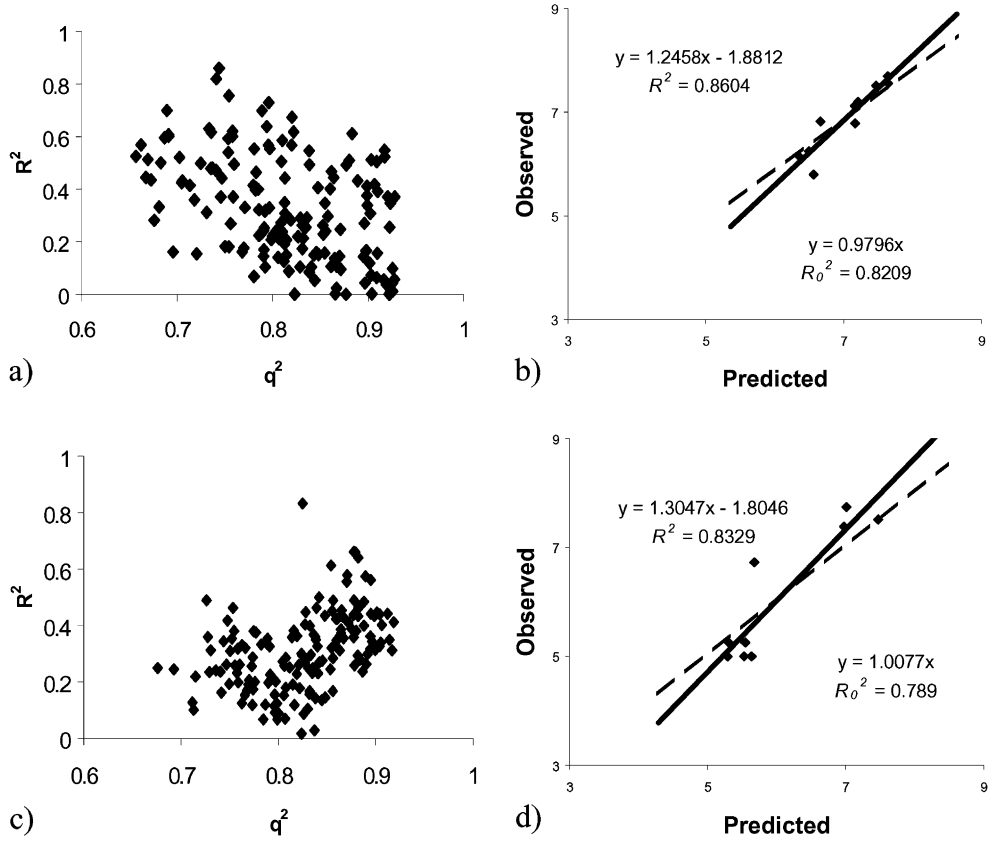

Fig. 2. QSAR modeling of Cramer's steroids [16]. (a) Predictive R^2 vs. q^2 for the models with $q^2 > 0.5$ for the same training and test sets. (b) Observed vs. predicted affinity to corticosteroid binding globulin by the best model obtained using compounds 1–21 as the training set and validated using compounds 22–31 as the test set. (c) Predictive R^2 vs. q^2 for the models with $q^2 > 0.5$ for the same training and test sets. (d) Observed vs. predicted affinity to corticosteroid binding globulin by the best model built using compounds 1–12 and 23–31 as the training set and validated using compounds 13–22 as the test sets.

accepted otherwise it is rejected [43]. Alternative approach is to find the boundary significance level α between H_1 and H_0. This boundary significance level can serve as an additional parameter of predictive power of a QSAR model. To find the boundary significance level α, the following equation must be solved:

$$F_{1,n-2,\alpha} = F, \tag{10}$$

where $F_{1,n-2,\alpha}$ is the F-distribution function with 1 and $n-2$ degrees of freedom. The higher the α, the better is the model. We used the MATLAB [44] fcdf function to obtain α-values for our models. Since, α is very close to 1 for good models, P-values were used instead; P-value is defined as $1-\alpha$.

For each example, the variable selection/optimization procedure described above was performed 10 times for each value of the descriptor subset D. Thus, for each experimental dataset, 160 QSAR models were built. Models were considered acceptable, if they satisfied all of the following conditions: $q^2 > 0.5$, $R^2 > 0.6$, R_0^2 or $R_0'^2$ close to R^2, i.e. $[(R^2 - R_0^2)/R^2] < 0.1$ or $[(R^2 - R_0'^2)/R^2] < 0.1$, and the corresponding $0.85 \le k \le 1.15$ or $0.85 \le k' \le 1.15$.

3. Results and discussion

3.1. Steroids dataset

Binding affinities of steroids to corticosteroid binding globulin were taken from [15]. The same two distributions of compounds into the training and test sets as in [15] were used. The first training set included compounds 1–21 and the corresponding test set included compounds 22–31. All 160 kNN QSAR models for the training set had $q^2 > 0.5$. They were used to predict activities of the test set compounds. Thus, for each of these models R^2 values were obtained. The plot of R^2 versus q^2 is shown in Fig. 2a. Obviously, this plot indicates no correlation between q^2 and R^2 but does show several models with high value of R^2. The best predictive model was characterized by $q^2 = 0.74$, $R = 0.93$ ($R^2 = 0.86$), $R_0^2 = 0.82$, RMSE $= 0.04$, $F = 49.3$, $k = 0.98$ and $P = 1.1 \times 10^{-4}$ (see Fig. 2b).

The second training set included compounds 1–12 and 23–31, while the corresponding test set included compounds 13–22. All 160 models obtained with these descriptors had $q^2 > 0.5$. They were used to predict activities of the test

set compounds. For each of these models, R^2 values were obtained. Plot of R^2 versus q^2 is shown in Fig. 2c. Similarly to the previous case, no relationship between q^2 and R^2 could be found as well. The best model obtained with *k*NN QSAR approach and non-chiral as well as some chiral descriptors had $q^2 = 0.82$, $R = 0.91$ ($R^2 = 0.83$), $R_0^2 = 0.79$, RMSE = 0.13, $F = 39.9$, $k = 1.01$ and $P = 2.29 \times 10^{-4}$ (see Fig. 2d).

Additional calculations were performed with the randomized binding affinities of compounds from the training sets. A total of 160 QSAR models were built for each of the two cases. In the first case, only 18 models had $q^2 > 0.5$. The highest q^2 was 0.72. The highest R^2 was 0.49. In the second case, 154 models had $q^2 > 0.5$. The highest q^2 was 0.75, and the highest R^2 was 0.51. These calculations demonstrate that q^2 could be high even if the affinities of compounds from the training set are randomized. Only the use of external test sets made it possible to establish that all models based on the training set with randomized affinities were useless, since they had low predictive power. Actually, in the absence of the test set calculations, we could have come to a wrong conclusion. The fact that some models based on the training sets with randomized affinities appeared to have high q^2 values could be explained by a chance correlation or structural redundancy [45]. Another reason of high q^2 values for randomized data may be the correlation between the real and randomized activities [45]. It is particularly actual for small datasets. In our example, this was not observed. In the first case, the correlation coefficient between real and randomized activity was 0.16, and in the second case, it was 0.01. In summary, the majority of the models with high value of q^2 had no predictive power as assessed by the (low) predictive R^2-value. Only 17 models in the first case and only five models in the second case had predictive $R^2 > 0.6$.

3.2. Ecdysteroids

This dataset contained 78 analogs of ecdyson [19]. The $-\log(ED_{50})$ values for ecdysteroids and division into the training and test set were taken from [19]. Thorough 2D QSAR studies were performed for this dataset earlier, as the first example to apply our novel chirality descriptors [38,46]. The training set contained 71 compounds. The test set contained seven compounds. One of the best models obtained had the following statistics: $q^2 = 0.71$, $R = 0.98$ ($R^2 = 0.96$), $R_0^2 = 0.95$, RMSE = 0.34, $F = 110.5$, $k = 0.85$ and $P = 1.35 \times 10^{-4}$ (Fig. 3a). The total number of descriptors was 300; 158 out of 160 models built using these descriptors had $q^2 > 0.5$. Plot of q^2 versus R^2 for this example is presented in Fig. 3b.

Additional calculations were performed using randomized training set, and 160 QSAR models were built. The training set for this example was much larger than for Cramer's steroids. Therefore, the chance correlation and large q^2 values were less likely for the same number D of descriptors selected from the whole descriptor set. Indeed, the highest q^2 for the training set with randomized activities was 0.30.

Again, as discussed above for the steroid dataset, without the validation of the models by using the external test set, a wrong conclusion could be made that all 158 models with $q^2 > 0.5$ are good. However, only nine models had $R^2 > 0.6$.

3.3. Histamine H_1 receptor ligands

This dataset included 35 analogs of 1-phenyl-3-amino-1,2,3,4-tetrahydronaphtalenes, 1-phenyl-3-aminotetralins (PATs) and 31 compounds with other structures (non-PATs) [20]. Binding affinities of the compounds to histamine H_1 receptor were taken from [20]. The same training set (50 compounds) and test set (16 compounds) as in [20] were used in our calculations. One of the best QSAR models had the following statistics: $q^2 = 0.69$, $R = 0.85$, ($R^2 = 0.72$), $R_0^2 = 0.72$, RMSE = 0.53, $F = 35.4$, $k = 1.02$, and $P = 3.5 \times 10^{-5}$ (Fig. 4a). All 160 models built with these descriptors appeared to have $q^2 > 0.5$. Therefore, all of them were used to predict binding affinities of compounds from the test set. Plot of q^2 versus R^2 for Histamine H_1

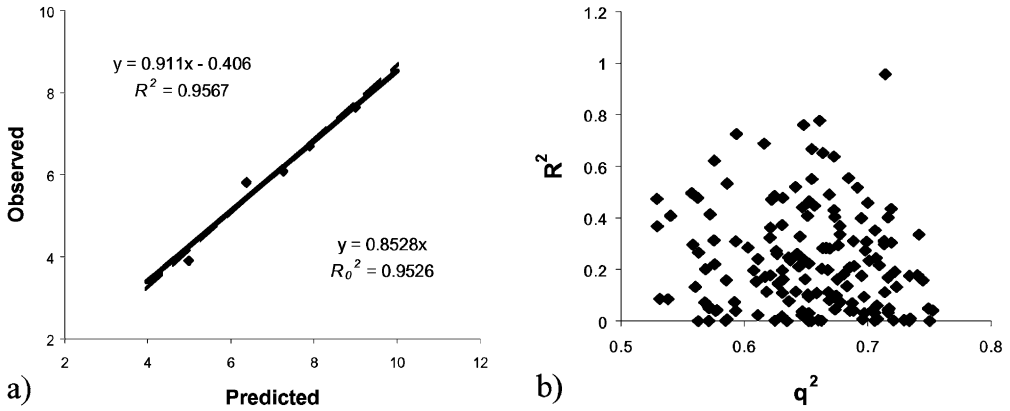

Fig. 3. QSAR modeling of ecdysteroids [19]. (a) Correlation between observed and predicted activity values for the best predictive model for the external test set containing seven compounds. (b) Predictive R^2 vs. q^2 for all models with $q^2 > 0.5$.

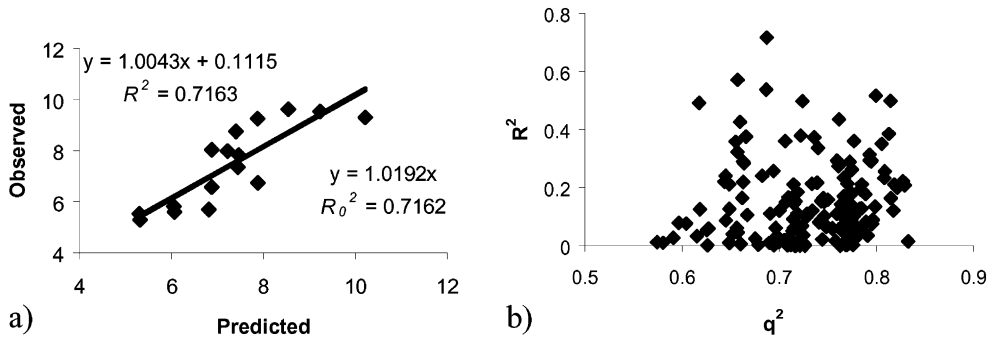

Fig. 4. QSAR modeling of histamine H_1 receptor ligands [20]. (a) Correlation between observed and predicted activity values for the best predictive model for the external test set containing 16 compounds. (b) Plot of predictive R^2 vs. q^2 for all models with $q^2 > 0.5$.

receptor ligands is presented in Fig. 4b. Despite high q^2 for all models, only one of them had predictive $R^2 > 0.6$.

Affinities of compounds in the training set were randomized. QSAR models based on the randomized training set were built. The same parameter values as for calculations with the real training set were used for these calculations and 160 QSAR models were built. For all these models, maximum q^2 was 0.43.

As in the previous example, without the validation of our QSAR models by using the external test set, we could have come to a wrong conclusion about high predictive ability of all our models. In fact, only one of them appeared to have a relatively high predictive ability.

4. Conclusions

This paper emphasizes that the predictive ability of a QSAR model can only be estimated using an external test set of compounds that was not used for building the model. We formulate the following criteria for a QSAR model to have high predictive power.

1. High value of cross-validated R^2 (q^2).
2. Correlation coefficient R between the predicted and observed activities of compounds from an external test set close to 1. At least one (but better both) of the correlation coefficients for regressions through the origin (predicted versus observed activities, or observed versus predicted activities), i.e. R_0^2 or $R_0'^2$ should be close to R^2.
3. At least one slope of regression lines through the origin should be close to 1. (It will correspond to R_0^2 or $R_0'^2$ that is closer to R^2.)

Based on the calculations and analysis presented in this paper, we conclude that despite its wide acceptance, a high value of q^2 alone is insufficient criterion for a QSAR model to be highly predictive. This conclusion appears to be a common feature of all QSAR approaches in which the number of descriptors is close to or higher than the number of compounds. Indeed, any procedure for QSAR model development is aimed at the enhancement of q^2 rather than predictive power (external R^2) of the model. The higher the number of descriptors relative to the number of compounds, the higher is a chance to select those of them (or latent variables in the case of PLS analysis) that give high q^2 values. Another reason for overestimating the q^2 can be structural redundancy of the training set [45]. In the case of non-linear methods, such as *k*NN applied in this work, the existence of multiple minima could present an additional problem. Many models could be characterized by high q^2, but only few of them are really highly predictive as judged by external validation. That is why the use of an external set of compounds for the model validation is always necessary. We summarize this conclusion in a simple statement: "*beware of* q^2!".

References

[1] C. Hansch, A. Leo, Substituent Constants for Correlation Analysis in Chemistry and Biology, Wiley/Interscience, New York, 1979.
[2] A. Tropsha, S.J. Cho, W. Zheng, New tricks for an old dog: development and application of novel QSAR methods for rational design of combinatorial chemical libraries and database mining, in: A.L. Parrill, M.R. Reddy (Eds.), Rational Drug Design: Novel Methodology and Practical Applications, ACS Symposium Series No. 719, 1999, pp. 198–211.
[3] W. Zheng, S.J. Cho, A. Tropsha, Rational combinatorial library design. 1. Focus-2D: a new approach to the design of targeted combinatorial chemical libraries, J. Chem. Inform. Comput. Sci. 38 (1998) 251–258.
[4] S.J. Cho, W. Zheng, A. Tropsha, Rational combinatorial library design. 2. Rational design of targeted combinatorial peptide libraries using chemical similarity probe and the inverse QSAR approaches, J. Chem. Inform. Comput. Sci. 38 (1998) 259–268.
[5] S. Clementi, S. Wold, How to choose the proper statistical method, in: H. van de Waterbeemd (Ed.), Chemometrics Methods in Molecular Design, VCH, Weinheim, 1995, pp. 319–338.
[6] S. Wold, PLS for multivariate linear modeling, in: H. van de Waterbeemd (Ed.), Chemometrics Methods in Molecular Design, VCH, Weinheim, 1995, pp. 195–218.
[7] B. Hoffman, S.J. Cho, W. Zheng, S. Wyrick, D.E. Nichols, R.B. Mailman, A. Tropsha, Quantitative structure–activity relationship modeling of dopamine D(1) antagonists using comparative molecular field analysis, genetic algorithms-partial least-squares, and *k* nearest neighbor methods, J. Med. Chem. 42 (1999) 3217–3226.

[8] Ajay, A unified framework for using neural networks to build QSARs, J. Med. Chem. 36 (1993) 3565–3571.
[9] S. Wold, L. Eriksson, Statistical validation of QSAR results, in: H. van de Waterbeemd (Ed.), Chemometrics Methods in Molecular Design, VCH, Weinheim, 1995, pp. 309–318.
[10] M. Recanatini, A. Cavalli, F. Belluti, L. Piazzi, A. Rampa, A. Bisi, S. Gobbi, P. Valenti, V. Andrisano, M. Bartolini, V. Cavrini, SAR of 9-amino-1,2,3,4-tetrahydroacridine-based acetylcholinesterase inhibitors: synthesis, enzyme inhibitory activity, QSAR, and structure-based CoMFA of tacrine analogues, J. Med. Chem. 43 (2000) 2007–2018.
[11] J.A. Morón, M. Campillo, V. Perez, M. Unzeta, L. Pardo, Molecular determinants of MAO selectivity in a series of indolylmethylamine derivatives: biological activities, 3D-QSAR/CoMFA analysis, and computational simulation of ligand recognition, J. Med. Chem. 43 (2000) 1684–1691.
[12] V. Austel, Experimental design in synthesis planning and structure–property correlations, in: H. van de Waterbeemd (Ed.), Chemometrics Methods in Molecular Design, Vol. 49, VCH, Weinheim, 1995, p. 62.
[13] E. Novellino, C. Fattorusso, G. Greco, Use of comparative molecular field analysis and cluster analysis in series design, Pharm. Acta Helv. 70 (1995) 149–154.
[14] U. Norinder, Single and domain made variable selection in 3D QSAR applications, J. Chemomet. 10 (1996) 95–105.
[15] H. Kubinyi, F.A. Hamprecht, T. Mietzner, Three-dimensional quantitative similarity–activity relationships (3D QSAR) from SEAL similarity matrices, J. Med. Chem. 41 (1998) 2553–2564.
[16] R.D. Cramer III, D.E. Patterson, J.D. Bunce, Comparative molecular field analysis (CoMFA). 1. Effect of shape on binding of steroids to carrier proteins, J. Am. Chem. Soc. 110 (1988) 5959–5967.
[17] E.A. Coats, The CoMFA steroids as a benchmark dataset for development of 3D QSAR methods, in: H. Kubinyi, G. Folkers, Y.C. Martin (Eds.), 3D QSAR in drug design, Vol. 3, Kluwer Academic Publishers, ESCOM, Dordrecht, 1998, pp. 199–213.
[18] W. Zheng, A. Tropsha, Novel variable selection quantitative structure–property relationship approach based on the *k* nearest neighbor principle, J. Chem. Inform. Comput. Sci. 40 (2000) 185–194.
[19] L. Dinan, R.E. Hormann, T. Fujimoto, An extensive ecdysteroid CoMFA, J. Comput.-Aided Mol. Des.13 (2) (1999) 185–207.
[20] E. Bucholz, R.L. Brown, A. Tropsha, R.G. Booth, S.D. Wyrick, Synthesis, evolution, and comparative molecular field analysis of 1-phenyl-3-amino-1,2,3,4-tetrahydronaphtalenes as ligands for histamine H1 receptors, J. Med. Chem. 42 (1999) 3041–3054.
[21] Molconn-Z, http://www.eslc.vabiotech.com/.
[22] M. Randić, On characterization on molecular branching, J. Am. Chem. Soc. 97 (1975) 6609–6615.
[23] L.B. Kier, L.H. Hall, Molecular Connectivity in Chemistry and Drug Research, Academic Press, New York, 1976.
[24] L.B. Kier, L.H. Hall, Molecular Connectivity in Structure–Activity Analysis, Wiley, New York, 1986.
[25] L.B. Kier, A shape index from molecular graphs, Quant. Struct. Act. Relat. 4 (1985) 109–116.
[26] L.B. Kier, Inclusion of symmetry as a shape attribute in kappa-index analysis, Quant. Struct. Act. Relat. 6 (1987) 8–12.
[27] L.H. Hall, L.B. Kier, Determination of topological equivalence in molecular graphs from the topological state, Quant. Struct. Act. Relat. 9 (1990) 115–131.
[28] L.H. Hall, B.K. Mohney, L.B. Kier, The electrotopological state: an atom index for QSAR, Quant. Struct. Act. Relat. 10 (1991) 43–51.
[29] L.H. Hall, B.K. Mohney, L.B. Kier, The electrotopological state: structure information at the atomic level for molecular graphs, J. Chem. Inform. Comput. Sci. 31 (1991) 76–82.
[30] L.B. Kier, L.H. Hall, Molecular Structure Description: The Electrotopological State, Academic Press, New York, 1999.
[31] G.E. Kellogg, L.B. Kier, P. Gaillard, L.H. Hall, The *E*-state fields. Applications to 3D QSAR, J. Comput. Aid. Mol. Des. 10 (1996) 513–520.
[32] L.B. Kier, L.H. Hall, A differential molecular connectivity index, Quant. Struct. Act. Relat. 10 (1991) 134–140.
[33] M. Petitjean, Applications of the radius–diameter diagram to the classification of topological and geometrical shapes of chemical compounds, J. Chem. Inform. Comput. Sci. 32 (1992) 331–337.
[34] H. Wiener, Structural determination of paraffin boiling points, J. Am. Chem. Soc. 69 (1947) 17.
[35] J.R. Platt, Prediction of isomeric differences in paraffin properties, J. Phys. Chem. 56 (1952) 328.
[36] C. Shannon, W. Weaver, Mathematical Theory of Communication, University of Illinois, Urbana, IL, 1949.
[37] D. Bonchev, O. Mekenyan, N. Trinajstic, Isomer discrimination by topological information approach, J. Comput. Chem. 2 (1981) 127–148.
[38] A. Golbraikh, D. Bonchev, A. Tropsha, Novel chirality descriptors derived from molecular topology, J. Chem. Inform. Comput. Sci. 41 (2001) 147–158.
[39] I. Gutman, B. Ruscic, N. Trinajstic, C.F. Wilcox Jr., Graph theory, J. Chem. Phys. 62 (1975) 3399.
[40] G. Rücker, C. Rücker, Counts of all walks as atomic and molecular descriptors, J. Chem. Inform. Comput. Sci. 33 (1993) 683–695.
[41] D. Bonchev, Overall connectivity and molecular complexity, in: J. Devillers, A.T. Balaban (Eds.), Topological Indices and Related Descriptors, Gordon and Breach, Reading, UK, 1999, pp. 361–401.
[42] D. Bonchev, Novel indices for the topological complexity of molecules, SAR/QSAR Env. Res. 7 (1997) 23–43.
[43] L. Sachs, Applied Statistics. A Handbook of Techniques, Springer, Berlin, 1984.
[44] MATLAB, http://www.mathworks.com/products/matlab/.
[45] R.D. Clark, D.G. Sprous, J.M. Leonard, Validating models based on large dataset, in: H.-D. Höltje, W. Sippl (Eds.), Rational Approaches to Drug Design. Proceedings of the 13th European Symposium on Quantitative Structure–Activity Relationships, Prous Sci., 2001, 475–485.
[46] A. Golbraikh, D. Bonchev, Y.-D. Xiao, A. Tropsha, Novel chiral topological descriptors and their applications to QSAR, in: H.-D. Höltje, W. Sippl (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, 219–223.