Application of L_1- Regularization Approach in QSAR Problem. Linear Regression and Artificial Neural Networks

M.I. Berdnyk, A.B. Zakharov, V.V. Ivanov*

V.N. Karazin Kharkiv National University, Faculty of Chemistry, Department of Materials Chemistry, 4 Svobody Sq., Kharkiv, 61022, Ukraine; *e-mail: vivanov@karazin.ua

Recieved: June 10, 2019; Accepted: August 07, 2019
DOI: 10.17721/moca.2019.79-90

One of the primary tasks of analytical chemistry and QSAR/QSPR researches is building of prognostic regression equations based on descriptors sets. The one of the most important problems here is to decrease the number of descriptors in the initial descriptor set which is usually way too big. In current investigation the descriptor set is proposed to be reduced employing the least absolute shrinkage and selection operator (LASSO) approach. Decreased descriptor sets were used for calculations with application of the following QSAR/QSPR methods: ordinary least squares (OLS), the least absolute deviation (LAD) regressions and artificial neural networks (ANN). Contrary to aforementioned methods principal component regression (PCR) and partial least squares (PLS) approaches can produce solutions containing numerous descriptors. In this article we compared the viability of these two different descriptor handling ideologies in application to molecular chemical and physical properties prediction. From the obtained results it is possible to see that there are tasks for which PCR and PLS approaches can fail to produce accurate regression equations. At the same time, methods OLS and LAD that use small amount of descriptors can provide viable solutions for the same cases. It was shown that these small sets of descriptors selected with LASSO approach can be used in ANN to obtain models with even better internal validation characteristics.

Keywords: QSAR/QSPR, regression, OLS, LAD, PCR, PLS, LASSO, LARS, boiling points, pKa, artificial neural networks

Використання L_1-регуляризації в проблемі QSAR. Лінійна регресія та штучні нейронні мережі

М. І. Бердник, А.Б. Захаров, В.В.Іванов*

Кафедра хімічного матеріалознавства, Харківський національний універсиситет імені В. Н. Каразіна, Майдан Свободи, 4, Харків, 61022, Україна; *e-mail: vivanov@karazin.ua

Надійшла: 10 червня 2019 р; Прийнята: 07 серпня 2019 р
DOI: 10.17721/moca.2019.79-90

Побудова прогностичних регресійних рівнянь з набору дескрипторів – одне з основних завдань аналітичної хімії а також QSAR/QSPR досліджень. При цьому однією з найважливіших проблем є скорочення первинного, зазвичай, надто широкого дескрипторного набору. В запропонованій роботі дескрипторні набори скорочуються за допомогою методу LASSO (the least absolute shrinkage and selection operator). Зменшені набори дескрипторів були застосовані для розрахунків з використанням наступних методів QSAR/QSPR: метод найменших квадратів (OLS), найменших абсолютних відхилень (LAD) а також метод штучних нейронних мереж (ANN). У протилежність вищевказаним підходам, методи головних компонент (PCR) та часткової регресії найменших квадратів (PLS) можуть давати регресійні рівняння побудовані на великій кількості дескрипторів. У представленій роботі було проведено порівняння цих двох ідеологій у проблемі прогнозування фізико-хімічних властивостей органічних сполук. Було показано, що для деяких задач методи PCR та PLS не дають задовільних за точністю регресійних рівняннь. В той же час методи OLS та LAD, які реалізуються в просторі невеликої кількості дескрипторів, можуть дати надійні рішення для тих самих задач. Показано, що невеликі дескрипторні набори отримані за допомогою методу LASSO можуть з успіхом бути використані у ANN для отримання моделей з кращими характеристиками внутрішньої валідації.

Ключові слова: QSAR/QSPR, регресія, OLS, LAD, PCR, PLS, LASSO, LARS, температури кипіння, pKa, штучні нейронні мережі

A search of quantitative structure–activity relationship (QSAR) models is an essential step in numerous scientific and technological applications. Among chemical sciences QSAR models are applied in various fields: in medical chemistry (e.g. targeted development of compounds with desired medical properties), toxicology (estimation of toxicity, carcenogenity and mutagenicity of chemical compounds), for estimation of physicochemical properties (lipophilicity, solubility, octane numbers, boiling points, melting points, etc.) and many other fields.

Despite the wide distribution of QSAR models in science the set of mathematical tools that are used for QSAR model building is not so diverse. The regression analysis is the most popular one. Multiple regression equations obtained for biological activity are based on Hansch theory (see for instance [1-5]. These models were reported to have high predictive ability. However, for QSAR problems it is typical that the number of descriptors exceeds the number of observations [6]. In this scenario the usage of such approaches may lead to the number of possible problems due to having much bigger amount of descriptors than observations which one should calibrate equation from. Moreover, given data sets may reveal multicollinearity of input data.

In the aforementioned cases the standard method of ordinary least squares (OLS) is not applicable [6, 7]. Still, there are some approaches that make it possible to overcome these difficulties. Among them there are methods that examine the factor structure of the problem: Principle Component Regression (PCR) and the Partial Least Squares, PLS (another term that could be met in literature is Projection on Latent Structures) [7, 8]. Among robust approaches one should first mention the least absolute deviation (LAD) method [9, 10] and ridge regression [11]. It is remarkable that despite the fact that LAD approach had been developed 50 years before the OLS (in 1757), the latter is overwhelmingly used for QSAR purposes. Thus, for most cases applicability of aforementioned approaches (LAD, PCR and PLS) still remains insufficiently studied in comparison with OLS.

An important group of regression approaches is based on the exclusion of the biggest part of descriptors from the model. Among these approaches – the simplest one is a stepwise regression. This method is almost identical to OLS and is based on ranking of descriptors according to some criteria, usually correlation coefficients. In this method one consequently includes the most important (or excludes the least important) descriptors in the regression equation. Such an approach hardly can be named universal since it does not take into account possible mutual correlations between descriptors and factor structure of descriptors matrix. Still, there are number of works in which approaches based on stepwise regression and other methods based on selection of descriptors are developed [12].

Among modern regression methods based on the selection of descriptors one of the most perspective methods is the Least Absolute Selection and Shrinkage Operator-(LASSO) [13].

The main goal of the current research is the analysis of regression models according to the chemometric rather than statistical point of view. In this work we are concerned with comparison of different algorithms results rather than with getting regression equations themselves. That is why we did not vary our initial set of data in order to get accurate QSAR equations. Instead we examined our initial (not ideal) set of data on "as is" basis and tried to obtain the best regression equations using different methods. Special attention in this work is paid to the LASSO approach which recently started to appear in literature in application to QSAR studies (see for instance [14]). In this work we propose using LASSO as preliminary selection procedure which is combined with least angle regression (LARS) methodology produces highly computationally efficient method which can handle thousands of descriptors. Selected descriptors can be used in the regression methods which cannot be applied directly to sets of data containing big amount of descriptors in it e.g. OLS, LAD, neural networks.

As the test systems we have chosen two experimental sets: boiling points (BP) of fluoroalkanes [15] and pKa of organic compounds [16]. The choice of characteristics under study is explained by the vastness and reliability of available experimental data.

Regression models

OLS regression is the well-known method that was described in multiple works and we do not discuss it in details, however, we still give a short insight into it here in order to complete the picture of methods used in this article.

In the OLS method one needs to minimize the function W(β).

$$W(\beta) = \left\| Y - X\beta \right\|_2^2 . \tag{1}$$

In this equation Y is the vector of property (response), X is the matrix of descriptors (predictors), *β* is the vector of regression coefficients to be found. Here index «2» means the Euclidean norm of matrix. The coefficients *β* can be found from the equation $\partial W / \partial \beta^{+} = 0$:

$$\beta = (X^{+}X)^{-1} X^{+} Y_{.} \tag{2}$$

In the LAD method one minimizes the following function:

$$U(\beta) = \left\| Y - X\beta \right\|_1 , \tag{3}$$

where subscript «1» denotes absolute value of the expression.

There are several LAD algorithms described in [17, 18]. In this work we formulated the matrix version of the algorithm, which corresponds to "variationally weighted" least squares approaches [9, 10]. In this method equation (3) can be transformed to weighted LS:

$$U(\beta) = \left(Y^{+} - \beta^{+}X^{+}\right)S(\beta)^{-1}\left(Y - X\beta\right), \tag{4}$$

where S(β) is the diagonal matrix with elements equal to:

$$S(\beta)_{ii} = \left|\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots - y_i\right|. \tag{5}$$

When the value of $S(\beta)_{ii}$ is small it is taken to be equal to1.

From equations (4) and (5) the expression for gradient ($\partial U/\partial\beta^{+}$) can be obtained. It is supposed that diagonal elements of matrix S(β) remains constant for every specific iteration and are not transformed in the expression for $\partial U/\partial\beta^{+}$. From the condition $\partial U/\partial\beta^{+} = 0$ the following equation for *β* can be obtained:

$$\beta = \left(X^{+}S(\beta)^{-1}X\right)^{-1}X^{+}S(\beta)^{-1}Y. \tag{6}$$

In this equation the vector of coefficients depends on current values of *β*. Thus, the problem has to be solved in iteration manner. Initial values of β_0 were found from expression (2). Small value (~10^{-6}) of difference between *β* obtained on subsequent iterations was assumed to be termination criterion (self-consistency of the solution).

It is obvious that in the LAD method different observations have different influence on the solution formation. In equation (6) matrix $S(\beta)^{-1}$ can be considered as the weight matrix. In case of matrix $S(\beta)^{-1}$ being an identity, the equation (6) coincides with that for OLS method.

Ridge regression can be considered as a generalization of the OLS method where minimizing function includes Euclidian norm (ℓ_2-norm) of vector *β* (also known as Tikhonov regularization [11, 19]):

$$W_{\lambda}^{(2)}(\beta) = \left\|Y - X\beta\right\|_2^2 + \lambda\left\|\beta\right\|_2^2, \tag{7}$$

where λ is the parameter of regularization by Tikhonov. For the *β* norm one can write corresponding representation $\left\|\beta\right\|_2^2 = \beta^{+}\beta$.

Usually λ is taken to be "small". More information on how to choose λ can be found in [19].

Proceeding the minimization of expression (7) leads to the following equation:

$$\beta = (X^{+}X + \lambda I)^{-1}X^{+}Y, \tag{8}$$

where I – is an identity matrix.

The LASSO method can also be considered as a generalization of the OLS method. In the LASSO approach one needs to minimize objective function similar to (1):

$$W_{\lambda}^{(1)}(\beta) = \left\|Y - X\beta\right\|_2^2, \tag{9}$$

with constraints:

$$\left\|\beta\right\|_1 = \sum_i \left|\beta_i\right| \le \tau. \tag{10}$$

Where $\left\|\beta\right\|_1$ is the ℓ_1-norm defined as the sum of the absolute values of the β matrix elements. It is obvious that when τ is small enough all the coefficients turn to zero.

The alternative form of the LASSO problem can be formulated as a minimization of the following function:

$$W_{\lambda}^{(1)}(\beta) = \left\|Y - X\beta\right\|_2^2 + \lambda\left\|\beta\right\|_1. \tag{11}$$

As in the case of Ridge regularization (7), λ is a measure of penalty function contribution. Despite the fact of the visual similarity of expressions (7) and (11), results obtained through minimization of expression (11) lead to different solutions for β. In case of the ℓ_1– penalty for each β_k element there exists λ_k such that contribution of corresponding descriptor k in (11) vanishes. In case of ℓ_2– penalty increase of λ parameter also leads to decrease of β elements. However, with increase of ℓ_2–penalty all elements of β decrease simultaneously. From statistical point of view in expression (11) more "important" descriptors require higher values of λ to disappear from the regression equation than less important one. Thus, with usage of ℓ_1–penalty one can range descriptors according to their importance in statistical models. Review of different approaches based on ℓ_1–penalty can be found in [20].

There are several LASSO algorithms. Some of them were studied in [21]. For example from the expression $\partial W/\partial\beta^{+} = 0$ and representation for $\left\|\beta\right\|_1$:

$$\left\|\beta\right\|_1 = \sum_i \left|\beta_i\right| = \beta^{+}s, \tag{12}$$

where vector s_i is defined as:

$$s_i = \begin{cases} 1, \, if \;\; \beta_i > 0 \\ -1, \, if \;\; \beta_i < 0, \\ 0, \, if \;\; \beta_i = 0 \end{cases} \tag{13}$$

self-consistent expression for β can be obtained:

$$\beta = \left(X^{+}X\right)^{-1}\left(X^{+}Y - \lambda s\right). \tag{14}$$

Our calculations, however, showed that this approach is numerically instable. It is connected with the fact that in (14) matrix ($X^{+}X$) often appears to be ill-conditioned for QSAR problems.

A simple but efficient method is based on the approach similar to variationally-weighted least squares method. In this method we use the following representation for $\left\|\beta\right\|_1$:

$$\left\|\beta\right\|_1 = \sum_i \frac{\beta_i\beta_i}{\left|\beta_i\right|} = \beta^{+}V^{-1}\beta, \tag{15}$$

where diagonal matrix V^{-1} is defined as shown in [21]

$$V^{-1} = \begin{pmatrix} 1/\left|\beta_1\right| & 0 & 0 & 0 \\ 0 & 1/\left|\beta_2\right| & 0 & 0 \\ 0 & 0 & 1/\left|\beta_3\right| & 0 \\ 0 & 0 & 0 & \ldots \end{pmatrix}. \tag{16}$$

In the case if $|\beta_i| \sim 0$ the corresponding element is taken equal to zero $V_{ii}^{-1} = 0$. Thus, V^{-1} is a pseudo inverse matrix. From equations (11) and (15) one can obtain the expression for gradient:

$$\frac{\partial W}{\partial \beta^+} = X^+X\beta - X^+Y + \lambda V^{-1}\beta\,. \tag{17}$$

From equation (17) and taking $\partial W / \partial \beta^+ = 0$ we can obtain following expression for β similar to expression (8):

$$\beta = \left(X^+X + \lambda V^{-1}\right)^{-1} X^+Y\,. \tag{18}$$

This equation is also should be solved with self-consistent approach since V depends on β. This approach appeared to be much more effective than procedure (14).

Gradient-iteration schemes (steepest descent, conjugated gradient, etc.), obtained from expression (17) appeared to be not effective enough in application to QSAR problems in our calculations. Instead, in this work we used iterative shrinkage-thresholding algorithm (ISTA) [22].

To discuss ISTA used in this work, it is necessary to define the shrinkage operator:

$$T_\lambda(x_i) = (|x_i| - \lambda)_+ \operatorname{sign}(x_i)\,, \tag{19}$$

where

$$(c)_+ = \begin{cases} c, \; c > 0 \\ 0, \; \text{otherwise} \end{cases}\,. \tag{20}$$

The general step of ISTA looks as follows:

$$\beta_{k+1} = T_{\lambda\mu}(\beta_k - \mu X^+(X\beta_k - Y)), \tag{21}$$

where μ is an appropriate stepsize.

However, the most effective approach in application to QSAR problems appeared to be the least angle regression (LARS) approach. It is known that LARS with minor modification to the original scheme guarantees obtaining of all LASSO solutions [23, 24]. Hereinafter we will call LARS with such modification as LARS-LASSO.

For the LARS algorithm initial data was mean-centered and variance-scaled i.e. each variable both dependent and independent was transformed in such way to have the average value equal to zero and the variance equal to unity. In the LARS algorithm one subsequently adds one descriptor per iteration to the model. On the first iteration from a given collection of possible descriptors one having the largest absolute correlation with the response is chosen. In the direction of this descriptor we take largest step possible until another descriptor has as much correlation with the current residual:

$$C = X^T(Y - X\beta_{LASSO})\,, \tag{22}$$

where C is called a vector of current correlations [23, 24], β_{LASSO} is a vector of current LASSO estimations. For descriptors not included to the model corresponding β elements are equal to zero. At this point we add this second descriptor to the model and proceed in the direction equiangular between these two descriptors until third descriptor appears to have the same correlation with the residual. This process continues until all the descriptors are included to the model.

In LARS-LASSO modification, while moving in equiangular direction between included descriptors, it is also necessary to check when coefficient signs of the aforementioned descriptors change. When the coefficient path of variable included into the model crosses through zero before another descriptor “earns” its place in “equicorrelation” set, we stop moving in this direction and exclude descriptor which corresponds to zero coefficient from the model. More information on LARS-LASSO algorithm can be found in [23, 24]. In our work we have used algorithm proposed by Tibshirani [23].

The LARS-LASSO is the most effective approach when studying problems which have higher amount of descriptors rather than number of molecules. In these cases all other LASSO approaches fail to find the solution because of numerical problems. It should be also mentioned here that, to the best of our knowledge, LARS-LASSO is also both the fastest and the least demanding to computer resources approach among all the LASSO methods.

Methods of principal component regression (PCR) and partial least squares (PLS) regression are well-known in QSAR studies. In our work we used the NIPALS algorithm of PCR method [25] and original NIPALS algorithm by Wold [7] for PLS.

In the PCR method principal components are produced from initial data X. After that simple OLS method with principal components as independent data is used to approximate experimental data. The benefits of this method are:

- principal components are made in such a way that obtained new set of latent variables is always orthonormal, thus, in equation (2) the corresponding matrix of new variables can always be inverted;
- usually in obtained set of principal components only first components are important while the last components usually depict errors in initial data.

However there are also drawbacks of this method:

- Obtained principal components may loose important data through the transformation and might be even not eligible for predicting properties.
- In PCR and PLS calculations all descriptors are used.

The first problem of the method partially can be solved with the usage of the PLS approach. In PLS approach dependant and independent data are accounted simultaneously to obtain latent variables.

To study regression methods for this research all the aforementioned algorithms were computationally implemented in program package.

Results and discussion

The experimental data for Boiling points of fluoroalkanes (BP) and pK_a of organic acids and bases were taken from the literature [15, 16]. The molecular geometry was optimized using semi-empirical AM1 method available in GAMESS [26]. After that, for the set of organic acids and bases we used PaDEL-descriptor software [27] to compute all available 1D&2D descriptors and E-Dragon 1.0 program [28, 29] to compute descriptors for fluoroalkanes.

In OLS and LAD calculations we used sets of descriptors which were preliminary chosen from LARS-LASSO calculation. Obtained models were tested with usage of leave-one-out cross-validation (LOOCV) procedure. Predictive abilities of obtained models were calculated with well-known equations:

$$R^2 = 1 - \sum_i (y_i - \hat{y}_i)^2 \Big/ \sum_i (y_i - \bar{y})^2 , \qquad (23)$$

$$Q^2 = 1 - \sum_i (y_i - \hat{y}_{i/i})^2 \Big/ \sum_i (y_i - \bar{y})^2 , \qquad (24)$$

$$\theta = R^2 - Q^2 , \qquad (25)$$

where y_i – experimental property values, $\bar{y}$ – average value of the corresponding dataset $\{y_i\}$, $\hat{y}_i$ – calculated property values, $\hat{y}_{i/i}$ – «predicted» property values from LOOCV procedure.

Coefficient of determination of LOOCV procedure (Q^2), and θ are important values to test quality of predictive ability of obtained models. Models were considered to be successful if $Q^2 > 0.5$ and θ is small enough [30].

Estimation of models quality and validation of QSAR equations are known to be a not solved issue. During the past decades there were multiple works which stated necessity of further research of significance of different validation characteristics [31-33]. Multiple alternative expressions to (23) and (24) were proposed to evaluate quality of models. However, none of them, to the best of our knowledge, are considered to be universal so far. At the same time (23) and (24) validation characteristics are still widely applied to evaluate quality of the models and considered to be the necessary validation characteristics to study that is why in current research we evaluate quality of the obtained models with these expressions.

Boiling points of fluoroalkanes

A dataset of 82 molecules was used for the study of BP of fluoroalkanes (herein is given in °C) [15]. 919 descriptors were calculated for this set with E-Dragon 1.0 program [28, 29]. 24 molecules out of 82 (30 % of the molecules) were used as the test set for validation of models obtained with usage of the training set.

With usage of LARS-LASSO approach we obtained subsequent set of descriptors that according to statistical point of view are the most important for prediction of boiling points of fluoroalkanes (26).

For description of the mentioned parameters see documentation of DRAGON [34].

To build regression equations with higher predictive ability one should subsequently add to the equation descriptors from (26) from the left side to the right side. General LASSO path of these descriptors under L_1-regularization conditions is shown in Fig. 1.

$$\text{IAC} > \text{TIC2} > \text{X0A} > \text{RDF020v} > \text{Mor08m} > \text{HATS2v} > \text{Mor30u} > \text{HATS2e} \qquad (26)$$

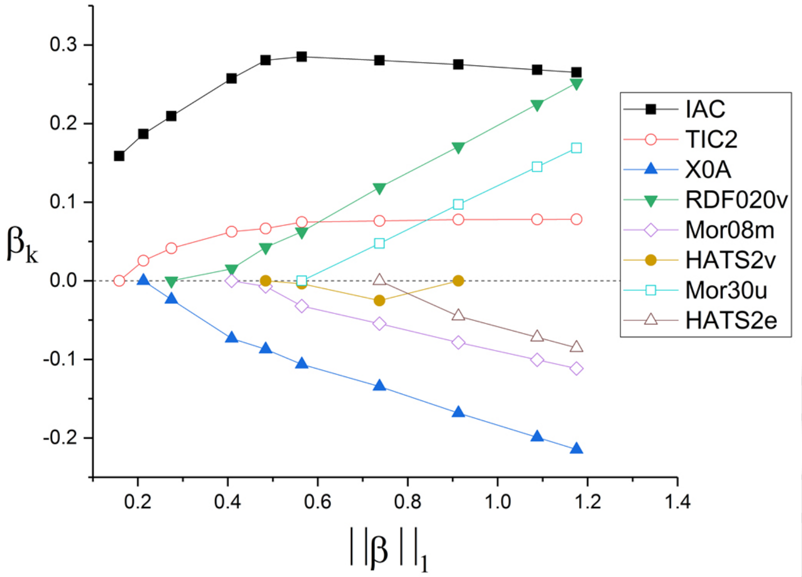

Fig. 1. General LASSO path for descriptors included in QSAR modeling of BP of fluoroalkanes.

One can see from Fig. 1 that almost all of the coefficients tend to shrink to zero with decrease of $\|\beta\|_1$, however there are still some exceptions. Descriptor HATS2V appears in equation as soon as descriptor HATS2e is excluded. It should be noted here that these two descriptors are highly correlated (R = 0.986). Thus, interchange of descriptor HATS2e with HATS2V almost does not affect quality of the regression equation. It is an important property of the LASSO approach that it does not tend to leave couples of highly correlated descriptors in the equation.

In this example we obtained multiple regression equations with application of different methods. Table 1 shows the comparison of quality of the equations which were obtained in PCR, PLS. Also in Table 1 the OLS and LAD regression equations which were obtained with descriptors selected by L_1-regularization are presented.

From Table 1 one can see that despite the fact that for building of regression equations in PCR all descriptors are used, the quality of such approximations is generally much worse than quality of equations which were obtained with OLS approach. Thus, obtained equations in LARS-LASSO-OLS two-step procedure are both more efficient and easily interpretable than those obtained in PCR. However, the quality of approximations which were obtained in PLS is similar to those obtained in LARS-LASSO-OLS procedure. The benefit of OLS and LAD approaches is the simplicity of obtained equations which do not use the whole descriptors set.

The regression equations for BP of fluoroalkanes obtained within OLS and LAD approaches are:

OLS:

$$BP=516.881+4.126IAC+0.718TIC2-692.497X0A, \quad R^2 = 0.861,\ Q^2 = 0.828,\ R^2_{TEST} = 0.774, \tag{27}$$

LAD:

$$BP=884.796+3.128LAC+0.752TIC2-1107.716X0A, \quad R^2 = 0.821,\ Q^2 = 0.811,\ R^2_{TEST} = 0.583, \tag{28}$$

Table 1. Comparison of quality of the models which were obtained in PCR, PLS, LAD and OLS for BP of fluoroalkanes.

	OLS			LAD			PCR			PLS		
N	R^2	Q^2	R^2_{TEST}	R^2	Q^2	R^2_{TEST}	R^2	Q^2	R^2_{TEST}	R^2	Q^2	R^2_{TEST}
1	0.74	0.72	0.75	0.74	0.74	0.75	0.15	0.08	0.19	0.66	0.59	0.68
2	0.77	0.75	0.73	0.77	0.77	0.75	0.67	0.63	0.74	0.82	0.78	0.85
3	0.86	0.83	0.77	0.82	0.81	***0.58***	0.68	0.61	0.76	0.94	0.90	0.91
4	0.92	0.89	0.92	0.90	0.85	0.89	0.84	0.81	0.80	0.96	0.92	0.94
5	0.94	0.92	0.91	0.94	0.93	0.92	0.86	0.82	0.82	0.97	0.93	0.96
6	0.95	0.93	0.94	0.95	0.93	0.94	0.87	0.83	0.83	0.98	0.94	0.96

In Table 1: N — is the amount of descriptors included in the calculation from set (26) in OLS/LAD methods and the amount of latent variables in PLS/PCR methods. R^2 and R^2_{TEST} are coefficients calculated according to equation (23) for the training and test set correspondingly. Q^2 was calculated according to equation (24).

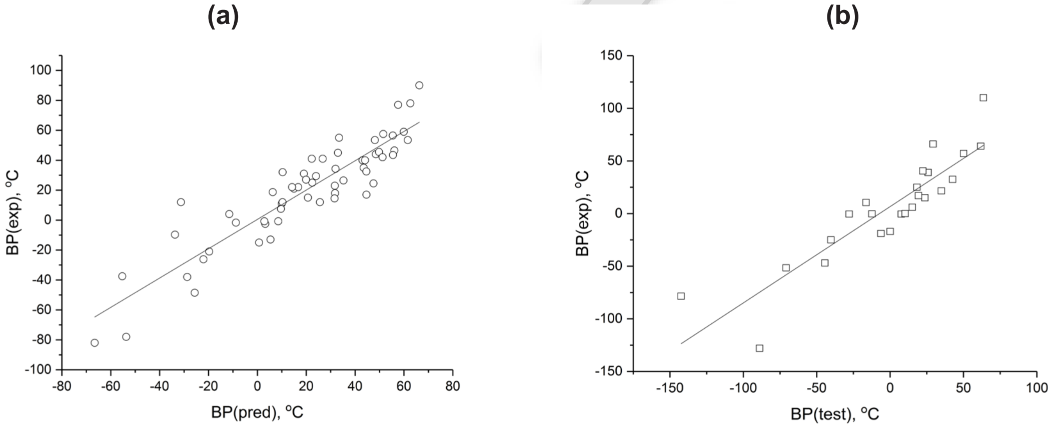

Fig. 2. Plot of predicted BP (a) obtained through LOOCV-procedure and calculated for the test set values (b) plotted against the experimental BP of fluoroalkanes (The results were obtained in OLS with three descriptors in equations).

In the following Figs. (2,3,4,5) one can see theory-experiment dependences (predicted with LOOCV and for the test set) calculated in approaches OLS, LAD, PCR and PLS correspondingly. The number of descriptors used in the model and the number of latent variables is equal to 3.

The following equations for trend lines of corresponding dependences 'BP(exp) — BP(pred/test)' were obtained in OLS:

$$\begin{aligned} &BP(exp) = 0.979BP(pred) + 0.471, \\ &BP(exp) = 0.875BP(test) + 6.613. \end{aligned} \qquad (29)$$

here in after we assume Bp(exp) are the experimental values for fluoroalkanes, Bp(pred) are values obtained with LOOCV procedure, and Bp(test) - are the values calculated for the test set.

$$\begin{aligned} &BP(exp) = 0.960BP(pred) + 0.994, \\ &BP(exp) = 0.768BP(test) + 13.579. \end{aligned} \qquad (30)$$

The point with Bp(exp) = - 78.5 from the test sample which lies sufficiently away from linear fit for the test data (see Fig. 3) corresponds to molecule CH_3F and most likely caused problems with regression equation (28) in LAD approach. It is most likely connected with the simplicity of the molecule which made it different from other molecules in the test set. This results in "robust" approach LAD not taking this molecule into account when building model with 3 parameters. In PCR and PLS studies, however, this molecule does not cause any problems.

For the PCR approach we obtained the next results:

$$\begin{aligned} &BP(exp) = 0.980BP(pred) - 0.074;\ R^2 = 0.615, \\ &BP(exp) = 1.203BP(test) + 2.016;\ R^2 = 0.789; \end{aligned} \qquad (31)$$

while for PLS:

$$\begin{aligned} &BP(exp) = 1.019BP(pred) - 0.550;\ R^2 = 0.900, \\ &BP(exp) = 1.004BP(test) + 3.643;\ R^2 = 0.912. \end{aligned} \qquad (32)$$

It is worth noting that the structures of molecules taken for this example were very similar. That is why predictive ability Q^2 in all methods appeared to be very similar to R^2 for the training set. In our next example it is not the case for some approaches.

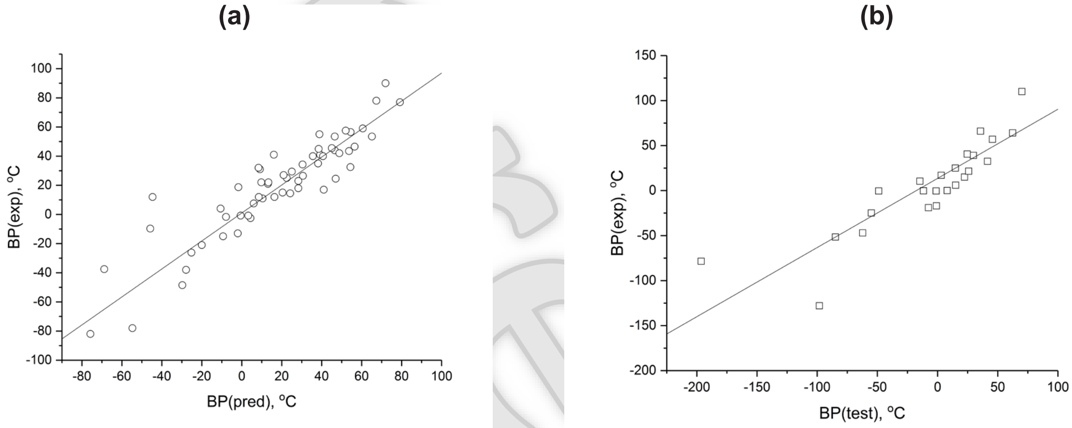

Fig. 3. Plot of predicted BP (a) obtained through LOOCV procedure and calculated for the test set values (b) plotted against the experimental BP. The results were obtained in LAD with three descriptors in equations.

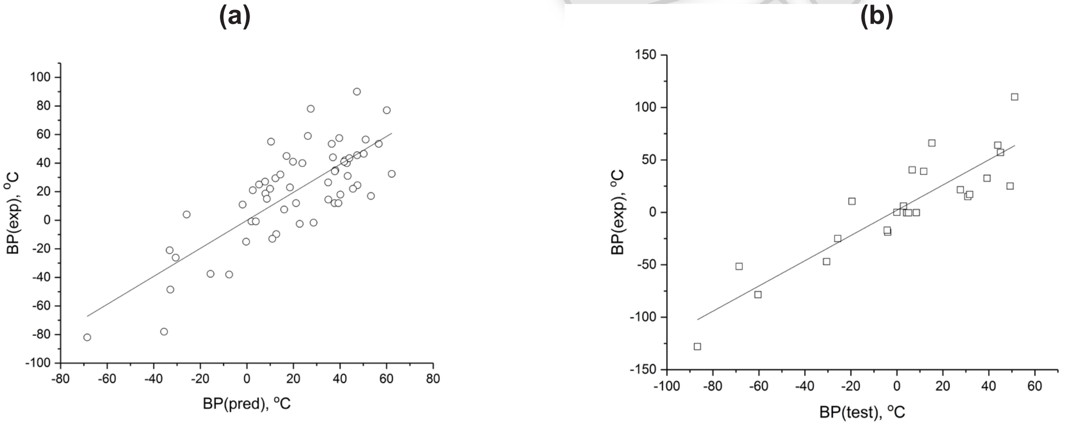

Fig. 4. Plot of predicted BP (a) obtained through LOOCV procedure and calculated for the test set values (b) plotted against the experimental BP of fluoroalkanes. The results were obtained in PCR with three latent variables in equations.

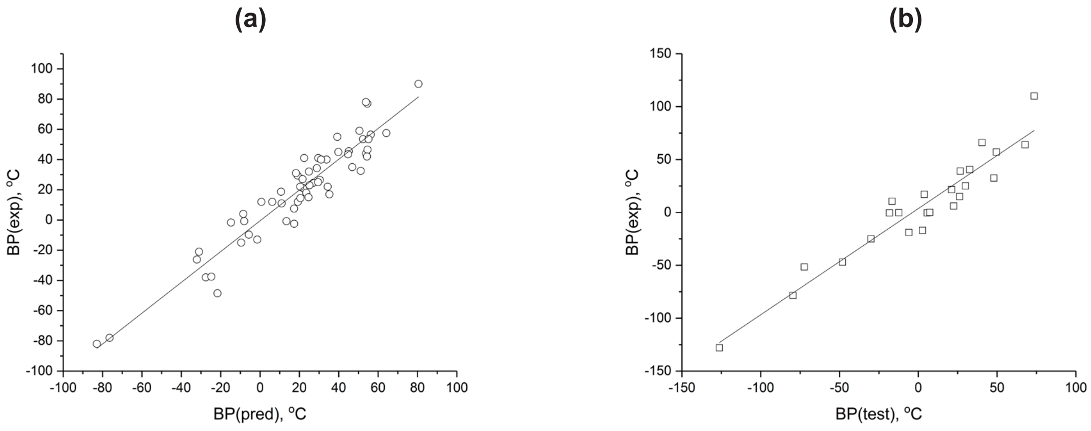

Fig. 5. Plot of predicted BP (a) obtained through LOOCV procedure and calculated for the test set values (b) plotted against the experimental BP of fluoroalkanes. The results were obtained in PLS with three latent variables in equations.

Ionization constants of organic compounds

For this example we have taken 43 molecules from [16] of different organic acids and bases, with different groups participating in acid-base equilibrium. In this case the set of the molecules is diverse. It leads to the situation when some of the descriptors may predict pKa well for some molecules while they can even not make sense for others. This circumstance should result in decrease of quality of obtained equations in methods which use the whole set of descriptors, i.e. PCR and PLS. We did not choose a test set from initial data set for this example due to its smallness and high structures variety.

In similar manner to our previous example using LARS-LASSO approach we obtained subsequent set of descriptors that are most important for prediction of pK_a of organic compounds (33).

General LASSO path of these descriptors under L_1-regularization conditions is shown in Fig. 6.

From Fig. 6 can see that the most important descriptors are AATS4e and AATSC5e. Contribution of all other descriptors disappears from the model very fast. Thus, we can assume that inclusion of more than 2 descriptors in model will lead to very small improvements in general quality of the model.

This idea can be confirmed with Fig. 7 which shows the comparison of quality of models which were obtained in PCR, PLS and models which were obtained in OLS and LAD with descriptors selected using L_1-regularized LASSO. One can see from figures for LAD and OLS that further addition of descriptors to current model almost does not change the overall quality of the model.

$$\text{AATS4e} > \text{AATSC5e} > \text{AATS1s} > \text{C2SP3} > \text{AATSC5s} > \text{GATS5s} > \text{maxHother} > \text{JGI3} \qquad (33)$$

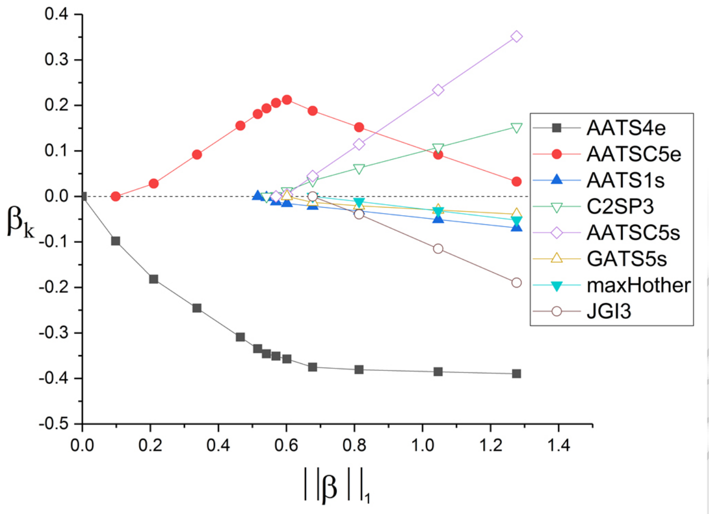

Fig. 6. General LASSO path for descriptors included in QSAR modeling of pK_a's of organic compounds.

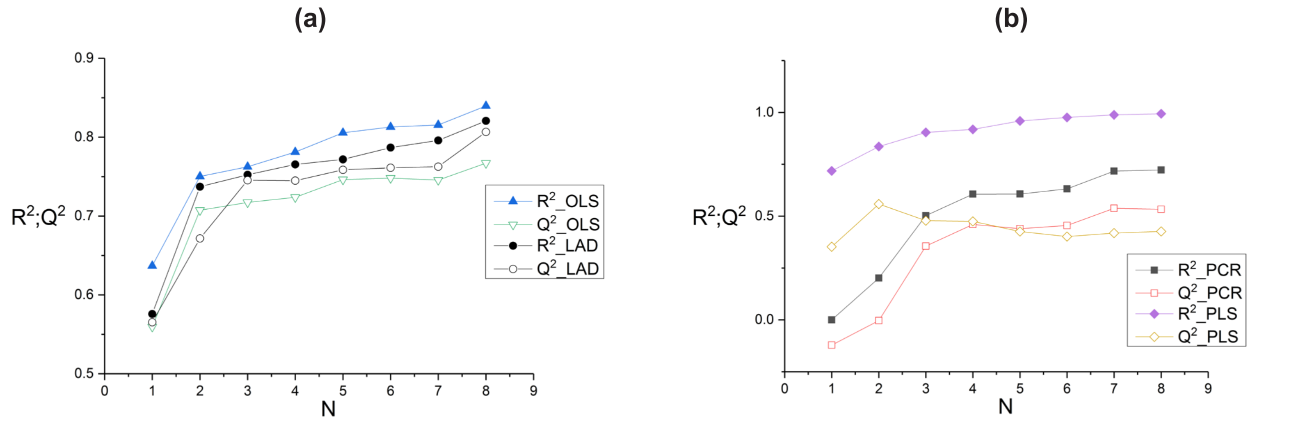

Fig. 7. Comparison of internal validation characteristics for equations which were obtained in (a) OLS, LAD and (b) PCR, PLS for pK_a of organic compounds.

As it was expected the quality of the equations obtained in PCR and PLS methods appeared to be very low. While the R^2 criteria for PLS calculation is still high, the predictive ability of the equation according to procedure LOOCV Q^2 shows that obtained models are of bad quality. Equations with two descriptors obtained in OLS and LAD approaches for pK_a in this case are the following:

OLS:

$$pK_a = 42.189 - 4.478AATS4e + 112.5AATSC5e, \quad (34)$$

R^2=0.750, Q^2=0.707;

LAD:

$$pK_a = 48.096 - 5.282AATS4e + 112.0AATSC5e, \quad (35)$$

R^2=0.737, Q^2=0.672;

where AATS4e - Average Broto-Moreau autocorrelation - lag 4 / weighted by Sanderson electronegativities and AATSC5e - Average centered Broto-Moreau autocorrelation - lag 5 / weighted by Sanderson electronegativities (for more details see [35]).

For the OLS approach the next equation was obtained:

$$pK_a(exp) = 0.966pK_a(pred) - 0.287, \quad (36)$$

while for the LAD approach:

$$pK_a(exp) = 0.787pK_a(pred) + 1.889. \quad (37)$$

As one can see the equations (34,35) obtained in LAD and OLS methods appeared to be very similar for this problem. See corresponding graphical representation of equations (36, 37) in Fig. 8.

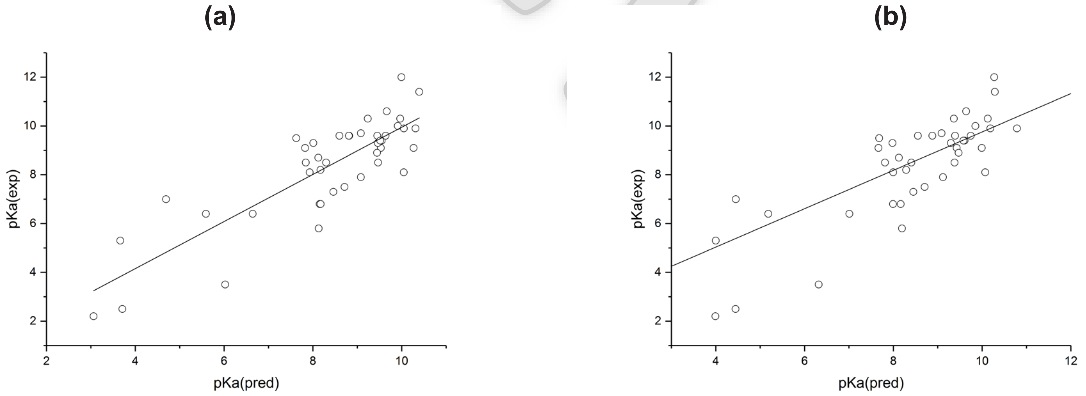

Fig. 8. Plot of predicted pK_a of organic compounds in LOOCV-procedure plotted against the experimental values in (a) OLS approach and (b) in LAD. Corresponding equations contained two descriptors.

High diversity of the molecules in this set and simplicity of descriptors employed made it impossible to take away even small amount of molecules for the testing purposes. Exclusion of molecules from the training set leads to significant decrease of predictive ability of models obtained in some researched methods, for example, in PLS (see Fig. 7b). This conclusion and the conclusion of general bad behavior of obtained models can be made from analysis of internal validation characteristics. For instance, big differences between values of R^2 and Q^2 for PLS models are connected with overfitting, consequently, models obtained for this task in this method are unreliable. Taking away even one molecule leads to big changes in structure of models making obtained values of Q^2 very low while values of R^2 tend to be high. PCR method fails to produce good results even for the test set according to the values of R^2 and Q^2 and thus cannot be used for prediction. OLS and LAD regression in this case gave the most reliable equations being less sensitive to decrease of the number of molecules in the training set. This conclusion was made from the fact of the difference between R^2 and Q^2 characteristics being small for these models. Even with reduced training sample these methods tend to produce acceptable results for pK_a of organic compounds but still with not very good values of internal characteristics.

To obtain models with better characteristics we assumed nonlinear structure of the task which can be modeled with artificial neural networks (ANN) approach. “NeuPy” package [36] was used to build neural networks in this study.

In our calculations we used multilayer feed-forward architecture of the neural network. For our purposes we constructed ANN with three layers: input and output layer with linear activation function and one hidden layer with hyperbolic tangent as activation function.

For the hidden layer we took 4 neurons. Such architecture of ANN appeared to give the best internal validation characteristics for this problem.

The structure of the neural network used in this study is given in Fig. 9.

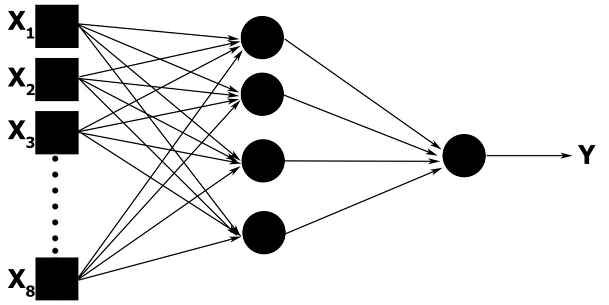

Fig. 9. Architecture of the ANN with 8 descriptors used in calculation. Here squares denote input descriptors while circles denote neurons.

Initialize parameters for the neural networks were sampled from the normal distribution with mean being equal to zero and standard deviation = 0.01. We used conjugate gradients algorithm with golden search modification for ANN training.

For the sake of comparability with regression equations we also used LOOCV procedure for ANNs quality testing. LOOCV procedure for ANN was performed in the following manner: network was trained first with the whole set of molecules; afterwards it was used as the initial approach for training networks obtained for every stage of LOOCV procedure. It appeared that Q^2 was high only for those ANNs which almost did not change during LOOCV procedure i.e. when initial network trained for the whole set of molecules is appropriate for all sets of LOOCV procedure.

In Table 2 the results of internal characteristics for each number of descriptors included in ANNs are presented. The same set of descriptors that was used in OLS, LAD was taken for ANN computations (33).

Table 2. Internal validation coefficients R^2; Q^2 for ANNs with different amount of descriptors included in models.

Number of descriptors	R^2	Q^2
1	0.680	0.529
2	0.833	0.787
3	0.854	0.679
4	0.932	0.709
5	0.967	0.865
6	0.980	0.882
7	0.994	0.943
8	0.994	0.935

As one can see from the Table 2 the ANN with just two descriptors outperforms any other model obtained within OLS, LAD, PCR and PLS approximations. Models with big amount of descriptors are good according to the internal validation. The best ANN results have been obtained by using seven descriptors where $Q^2 = 0.943$. Corresponding figure “theory-experiment” for neural network with 2 descriptors is presented in Fig. 10.

The Following Equation for trend line of corresponding dependence ‘pK_a(exp)-pK_a(pred)’ was obtained for ANN approach:

$$pK_a(\text{exp}) = 0.977pK_a(\text{pred}) + 0.196,\ Q^2 = 0.787. \qquad (38)$$

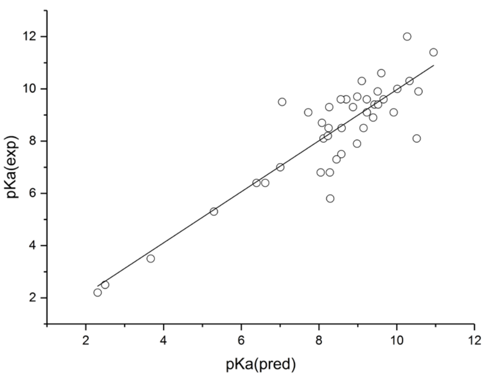

Fig. 10. Plot of predicted pK_a plotted against the experimental pK_a of organic compounds. The results were obtained in ANN approach with 2 descriptors.

Conclusions

Nowadays statistical science offers wide range of different approaches for building prognostic models. However, in the practical calculations especially in QSAR investigations the certain amount of calculations are still performed by using only OLS approach. It is connected with the natural desire to obtain the most compact and interpretable equation when the number of descriptors is small enough. However, it should be recognized that in general case it is difficult to find one, two, or three independent descriptors which perfectly correlated with the response. In this connection, more sophisticated approaches (PCR, PLS, etc.) are used for the case when smaller set of descriptors is not obvious. Both the PCR and PLS methods may produce reliable models in terms of quality, but from the chemical point of view they do not give any nontrivial information about the nature of obtained equations. Moreover, these methods may even fail to produce adequate models in some situations which were discussed in this work.

As the alternatives to rather straightforward approaches (PCR and PLS) the LASSO and LARS–LASSO methods appeared to be able to choose the descriptor sets for more compact and reliable models. According to the studied examples in this work LASSO-based approaches produce models which may outperform in some cases PLS and PCR methods. Also simple ANN approach that employs the descriptors which were selected using LASSO, despite the fact of generally not interpretable representations, appeared to be able to predict property with good characteristics of internal validation.

With such a pretty “colorful” picture of available approaches the problem of choice of appropriate model arises. In this regard, we assume a pragmatic approach, which is demonstrated in this work. For the contemporary level of computer technology, it is easy to carry out calculations based on different regression approaches. The registration of significant discrepancies in QSAR models (and corresponding prognostic parameters) for a test sample serve as an indicator to involve additional research for choice of appropriate model. On the other hand, identical within statistical significance results obtained in various regression approaches (as well as success of prediction by ANN) is evidence of correctness of the proposed equation.

References

1. Kubinyi H. QSAR: Hansch analysis and related approaches, *Methods and principles in medical chemistry*, VCH Verlagsgesellschaft mbH, 1993.

2. Marini F. Chemometrics in Food Chemistry, *Data Handling in Science and Technology*, Elsevier, 2013, 28(1), 512 p.

3. Roy K., Kar S., Das R.N., A Primer on QSAR/QSPR Modeling Fundamental Concepts, *Springer briefs in molecular science*, 2015.

4. Roy, K.; Advances in QSAR modeling Applications in Pharmaceutical, Chemical, Food, Agricultural and Environmental Sciences, *Challenges and Advances in Computational Chemistry and Physics*, 2017, Vol. 24.

5. Gupta, S.P.; QSAR and Molecular Modeling Studies in Heterocyclic Drugs II, *Topics in Heterocyclic Chemistry*, 2006, Vol. 4.

6. Filzmoser P., Gschwandtner M., Todorov V. Review of sparse methods in regression and classification with application to chemometrics, *J. Chemom*., 2012, 26, 42–51.

7. Wold S., Ruhe A., Wold H., Dunn W.J. The collinearity problem in linear regression. the partial least squares (PLS) approach to generalized inverses, *Siam j. sci. stat. comp.*, 1984, 5(3), 735-743.

8. Wold S., Eriksson L. Trygg J., Kettaneh N. The PLS method - partial least squares projections to latent structures – and its applications in industrial RDP (research, development, and production), PLS in industrial RPD - for Prague, 2004.

9. Mudrov V.I., Kushko V.L. Metod naimen’shih moduley, *Znanie*, 1971. (in Russ.)

10. Mudrov V.I., Kushko V.L. Metody’ obrabotki izmereniy, *Sovetskoe radio*, 1976. (in Russ.)

11. Tikhonov A.N., Arsenin V.Y. Solutions of ill-posed problems, *John Wiley & Sons.*, 1977

12. Miller A. Subset Selection in Regression, *Chapman & Hall CRC*, 2002

13. Tibshirani R. Regression Shrinkage and Selection via the Lasso, *J. Roy. Statist. Soc.*, 1996, 58(1), 267–288.

14. Long J., Li T., Yang M., Hu G., Zhong W. Hybrid strategy integrating variable selection and a neural network for fluid catalytic cracking modeling, *Ind. Eng. Chem. Res.*, 2019, 58(1), 247-258.

15. Rucker C., Meringer M., Kerber A., QSPR Using MOLGEN-QSPR: The Challenge of Fluoroalkane Boiling Points, *J. Chem. Inf. Model*., 2005, 45(1), 74-80.

16. Jensen J.H., Swain C.J., Olsen L. Prediction of pK_a values for drug-like molecules using semiempirical quantum chemical methods, *J. Phys. Chem*. A, 2017, 121(3), 699–707.

17. Wesolowsky G.O. A new descent algorithm for the least absolute value regression problem, *Communications in Statistics-Simulation and Computation*, 1981, 10(5), 479-491.

18. Bloomfield P., Steiger W.L. Least Absolute Deviations: Theory, Applications and Algorithms, progress in probability and statistics. 1983, 349 p.

19. Morozov V.A. Regulation Methods for ill-posed problems, *CRC Press.* 1 edition, 1993, 272 p.

20. Hastie T., Tibshirani R., Wainwright M. Statistical Learning with Sparsity. The Lasso and Generalizations, *CRC Press*, 2015.

21. Schmidt M. Least Squares Optimization with L1-Norm Regularization, CS542B Project Report, 2005.

22. Beck A., Teboulle M., A Fast Iterative Shrinkage-Thresholding Algorithm for Linear Inverse Problems, *Siam J. Imaging Sciences*, 2009, 2(1), 183–202.

23. Tibshirani R.J. The lasso problem and uniqueness, *Electr. J. Statistics*, 2013, 7, 1456–1490.

24. Efron B., Hastie T., Johnstone I., Tibshirani R. Least angle regression, *The Annals of Statistics*, 2004, 32(2), 407–451.

25. Geladi P., Kowalski B.R., Partial Least-Squares Regression: A Tutorial. *Anal. Chim.* Acta, 1986, 185, 1-17.

26. GAMESS official website: https://www.msg.chem.iastate.edu/gamess/.

27. PaDEL-Descriptor software official webpage http://www.yapcwsoft.com/dd/padeldescriptor/.

28. Tetko I.V., Gasteiger J., Todeschini R., Mauri A., Livingstone D., Ertl P., Palyulin V.A., Radchenko E.V., Zefirov N.S., Makarenko A.S., Tanchuk, V.Y., Prokopenko V.V. Virtual computational chemistry laboratory - design and description, *J. Comput. Aid. Mol. Des.*, 2005, 19, 453-463.

29. VCCLAB, Virtual Computational Chemistry Laboratory, 2005, http://www.vcclab.org.

30. Veerasamy R., Rajak H., Jain A., Sivadasan S., Varghese, C.P., Agrawal R.K. Validation of QSAR Models - Strategies and Importance, *Int. J. Drug Design and Discovery*, 2011, 2(3). 511-519.

31. Todeschini R., Beware of unreliable Q2! A comparative study of regression metrics for predictivity assessment of QSAR models, *J. Chem. Inf. Model*., 2016, 56(10), 1905–1913.

32. Golbraikh A., Tropsha A., Beware of Q^2 Journal of Molecular Graphics and Modelling, 2002, 20(4), 269–276.

33. Alexander D.L.J., Tropsha A., Winkler D.A., Beware of R^2: Simple, Unambiguous Assessment of the Prediction Accuracy of QSAR and QSPR Models, *J. Chem. Inf. Model*., 2015, 55(7), 1316-1322.

34. DRAGON molecular descriptor list http://www.talete.mi.it/products/dragon_molecular_descriptor_list.pdf.

35. Todeschini R., Consonni V. Molecular descriptors for chemoinformatics, Wiley VCH Verlag GmbH & Co. KGaA, 2009, 714-726.

36. NEUPY python library home page http://neupy.com/pages/home.html.