Jackknife-Based Selection of Gram−Schmidt Orthogonalized Descriptors in QSAR

Mohsen Kompany-Zareh*,†,‡ and Nematollah Omidikia†

Department of Chemistry, Institute for Advanced Studies in Basic Sciences (IASBS), Zanjan 45137-66731, Iran and Department of Food Science, Faculty of Life Sciences, University of Copenhagen, Rolighedsvej 30, 1958 Frederiksberg C, Denmark

Received May 2, 2010

This study is an implementation of a robust jackknife-based descriptor selection procedure assisted with Gram−Schmidt orthogonalization. Selwood data including 31 molecules and 53 descriptors was considered in this study. Both multiple linear regression (MLR) and partial least squares (PLS) regression methods were applied during the jackknife procedures, and the desired results were obtained when using PLS regression on both autoscaled and orthogonalized data sets. Having used the Gram−Schmidt technique, descriptors were all orthogonalized, and their number was reduced to 30. A reproducible set of descriptors was obtained when PLS-jackknife was applied to the Gram−Schmidt orthogonalized data. The simple statistical *t*-test was applied to determine the significance of the obtained regression coefficients from jackknife resampling. Increasing the sample size, descriptors, based on their information content, were introduced into the model one by one and were sorted. The number of validated descriptors was in proportion with the sample size in the jackknife. The PLS-jackknife parameters, such as sample size and number and number of latent variables in PLS, and the starting descriptor in Gram−Schmidt orthogonalization were investigated and optimized. Applying PLS-jackknife to orthogonalized data in the optimized condition, five descriptors were validated with $q^2_{\mathrm{TOT}} = 0.693$ and $R^2 = 0.811$. Compared to the previous reports, the obtained results are satisfactory.

1. INTRODUCTION

Quantitative structure−activity relationships (QSARs) are models which describe a mathematical relationship between the related molecular descriptors that encode information on the molecular structure and a property of a set of chemicals. For model building, a variety of descriptors have been defined including spatial, information content, topological, thermodynamic, conformational, electronic, quantum mechanical, and shape descriptors. The number of defined descriptors may amount to thousands for a model. Hence, determining the optimal analytical form of the QSAR model and dealing with a small but potentially useful portion of all possible subsets of descriptors make a real challenge. When the number of the independent variables is much greater than the dependent variables (molecules):

(1) The regression coefficient of simple multiple linear regressions cannot be calculated because the variance−covariance matrix of the data is singular. Although some factor-based regression methods, such as partial least squares (PLS) and principal component regression (PCR), are able to deal with this type of data sets and can be used for calculating regression coefficients, there are many reports indicating that highly predictive models were obtained if the variable selections were used before such factor-based approaches.[1,2]

(2) Factor-based regression analyses become more complex and less reproducible as the number of descriptors increases. For n independent variables, there are $2^n - 1$ possible regression equations, and therefore, it led to highly unstable regression coefficients.[3]

(3) The numerosity of the variables enhances the risk of chance correlation as it was pointed out by Topliss et al.[4] This numerosity will lead to substantial overstatement of the statistical significance of the parameters picked up in the final model.[5] Various aspects of large sets of descriptors are considered by Hawkins.[6]

(4) Some variables enclose redundant information and noise. Besides, they have no considerable effects on target property. Irrelevant variables deteriorate the accuracy of the model and cloud the meaningful relationship that exists between important variables.[7]

(5) Due to the existence of correlation between variables, the overfitting is encountered and degrades the prediction ability of the model.

As a result, nowadays, there is considerable interest in variable selection, and this subject has been interestingly studied especially in QSAR. Variable selection is a method of building the stable, interpretable, highly predictive models. Furthermore, it attempts to make a model whose variables are less collinear and are truly relevant to the output. Generally, there are two main categories of variable selection methods. The first group is elaborate artificial intelligent-based methodologies, such as genetic algorithm,[8] ant colony optimization,[9] automatic relevance determination,[10] generalized simulated annealing,[11] evolutionary algorithm,[12] and binary particle swarms.[13] All of the aforementioned strategies preferentially find the best set of variables. However, they are very difficult to use, and their theories and procedures are not so easy to understand for many users. The second group of variable selection methods is those categories that are based on simple statistical rules and are easy to

* Corresponding author. E-mail: kompanym@iasbs.ac.ir. Telephone: +98-241-415-312.
† Institute for Advanced Studies in Basic Sciences (IASBS).
‡ University of Copenhagen.

10.1021/ci100169p
Published on Web 11/11/2010

Table 1. EC_{50} and Physical Properties of 31 Antifilarial Antimycin Analogues

molecules	R'	R''	mp, °C	analysis	recryst solvent	% yield	EC_{50}
1	3-NHCHO	$NHC_{14}H_{29}$	81−83	C,H,N	ethanol/water	38	7.0
2	3-NHCHO	NH-3-Cl-4-(4-ClC_6H_4O)C_6H_3	182−184	C,H,N	acetic acid	38	2.4
3	5-NO_2	NH-3-Cl-4-(4-ClC_6H_4O)C_6H_3	205−207	C,H,N	acetic acid/water	38	0.04
4	5-SCH_3	NH-3-Cl-4-(4-ClC_6H_4O)C_6H_3	143−145	C,H,N,Cl,S	ethanol	27	0.48
5	5-$SOCH_3$	NH-3-Cl-4-(4-ClC_6H_4O)C_6H_3	165−168	C,H,N	hexane/ethyl acetate	43	7.5
6	3-NO_2	NH-3-Cl-4-(4-ClC_6H_4O)C_6H_3	142−146	C,H,N	acetic acid	51	0.15
7	5-CN	NH-3-Cl-4-(4-ClC_6H_4O)C_6H_3	256−258	C,H,N	acetic acid	9	0.0145
8	5-NO_2	NH-4-(4-$CF_3C_6H_4O$)C_6H_4	199−202	C,H,N	acetic acid	46	0.095
9	3-SCH_3	NH-3-Cl-4-(4-ClC_6H_4O)C_6H_3	151−153	C,H,N	acetic acid/water	12	0.38
10	5-SO_2CH_3	NH-3-Cl-4-(4-ClC_6H_4O)C_6H_3	195−196	C,H,N,S	ethanol/water	95	1.0
11	5-NO_2	NH-4-(C_6H_5O)C_6H_4	212−215	C,H,N	acetic acid	5	0.8
12	5-NO_2	NH-3-Cl-4-(4-ClC_6H_4CO)C_6H_3	246−248	C,H,N	acetic acid	35	0.074
13	5-NO_2	NH-4-(2-Cl-4-$NO_2C_6H_3O$)C_6H_4	208−211	C,H,N	acetic acid	22	0.12
14	5-NO_2	NH-3-Cl-4-(4-$CH_3OC_6H_4O$)C_6H_3	156−161	C,H,N	acetic acid/water	41	0.17
15	3-SO_2CH_3	NH-3-Cl-4-(4-ClC_6H_4O)C_6H_3	178−180	C,H,N	ethanol/water	65	0.5
16	5-NO_2	NH-3-Cl-4-(4-ClC_6H_4S)C_6H_3	203−206	C,H,N	acetic acid	28	0.044
17	3-NHCHO	NHC_6H_{13}	62−63	C,H,N	ether/pentane	60	>10
18	3-NHCHO	NHC_8H_{17}	78−80	C,H,N	ethanol/water	38	2.6
19	3-NHCHO	$NHC_{14}H_{29}$	71−72	C,H,N	ethanol/water	25	8.0
20	5-NO_2	$NHC_{14}H_{29}$	90−92	C,H,N	acetic acid	44	0.128
21	3-NO_2	$NHC_{14}H_{29}$	67−68	C,H,N	acetic acid	49	0.152
22	3-NO_2-5-cl	$NHC_{14}H_{29}$	81−83	C,H,N	acetic acid	61	0.0433
23	5-NO_2	NH-4-C(CH_3)C_6H_4	227−229	C,H,N	acetic acid	51	0.59
24	5-NO_2	$NHC_{12}H_{25}$	85−87	C,H,N	acetic acid/water	33	0.039
25	3-NO_2	$NHC_{16}H_{33}$	79−80	C,H,N	acetic acid/water	55	1.1
26	5-NO_2	NH-3-Cl-4-(4-ClC_6H_4NH)C_6H_3	176−175	C,H,N	acetic acid/water	10	0.37
27	5-NO_2	NH-4-(3-$CF_3C_6H_4O$)C_6H_4	176−178	C,H,N	ethanol/water	43	0.094
28	5-NO_2	NH-3-Cl-4-(4-$SCF_3C_6H_4O$)C_6H_3	195−197	C,H,N	ethanol/water	52	0.028
29	5-NO_2	NH-3-Cl-4-(3-$CF_3C_6H_4O$)C_6H_3	192−194	C,H,N	ethanol	43	0.085
30	5-NO_2	NH-4-(C_6H_5CHOH)C_6H_4	170−180	C,H,N	ethanol/water	57	>10
31	5-NO_2	4-ClC_6H_4	170−172	C,H,N	ethanol	23	0.33

understand.[14] Stepwise regression is one of these methods and can be classified into forward selection, backward elimination, and stepwise methods that are based on greedy search. Other techniques, such as successive projection algorithm (SPA), use a simple projection operator in data spaces in order to select the final set of descriptors and can be used for selecting the optimum number of variables with minimum collinearity.[15,16] In our previous study, a modified SPA named correlated weighted successive projection algorithm[17] (CWSPA) was used for variable selection in QSAR. Uninformative variable elimination-SPA[18] (UVE-SPA) is the other technique related to SPA with better prediction ability and is reported to avoid overfitting. UVE is a method of variable selection based on stability analysis of regression coefficient and can eliminate the variables which have no more information but noise for modeling. Novel methods in feature selection are reported recently.[19,20]

The use of orthogonalized variables as an approach for selection of variables before nonlinear modeling by artificial neural network (ANN) was reported.[21] Gram−Schmidt as an orthogonalization based variable reduction algorithm was followed by regression analyses using ridge regression (RR), PCR, and PLS.[22] Using orthogonal variables has some interesting features, such as possessing the same correlation coefficient R, the standard error S, and the F-test value as the regression model using nonorthogonal variables and the stability of the regression coefficients. Also, the resolution of ambiguities in structure−property studies by use of orthogonal descriptors is reported.[23] There are many literature reports on applicability of orthogonalization in QSAR.[5,24−26]

Leave-one-out cross-validation (LOO-CV),[3] Monte Carlo cross-validation (MCCV),[27] jackknifing,[28,29] and bootstrapping[30] are reported to be used for variable selection. The important variables for modeling could be selected by evaluating the prediction ability of the model and the validity of the regression coefficients. The use of jackknife as a resampling procedure for estimating the confidence interval for estimated parameters in QSAR is reported.[31] The procedure was used for reducing overoptimism in variable selection in QSAR.[32]

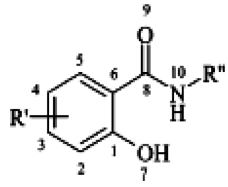
The proposed variable selection procedure used in this study is based on jackknife resampling of orthogonalized data and t-test, which is based on simple statistical rules. This efficient selection procedure provides a sorted set of descriptors, based on their information content, and a considerable improvement in the interpretative and predictive capability of QSAR models.

2. MATERIALS AND METHODS

2.1. Database Set. The Selwood data set[33] was developed by Selwood in a research project aimed at the development/discovery of novel antifilarials (Table 1). Waller[34] employed this data set to test the novel variable selection algorithm, fast random elimination of descriptors (FRED) algorithm, Shu-Shen Liu[35] for variable selection and modeling method based on the prediction (VSMP), and Whitley[36] for unsupervised forward selection. This data set was used as a reference set to derive QSAR models and has become a particular literature benchmark for evaluating variable selection procedures.[37−40]

This data set includes 31 compounds (antifilarial antimycin analogues, Scheme 1), each has the values of 53 physicochemical descriptors and in vitro biological activities,

Scheme 1



$-\log(EC_{50})$ (EC_{50} unit, μm) (Table 2). In EC_{50} assay, the macrofilarial viability was determined after 120 h exposure to a range of drug concentrations, descending from 10 μM.[33−35] The values of the in vitro activities $-\log(EC_{50})$ are widespread and homogeneous. From (Table 1) 8 compounds display pEC_{50} values between −1.0 and 0.0 (low activity), 12 display pEC_{50} values between 0.0 and 1.0 (moderate activity), and 10 show pEC_{50} values between 1.0 and 2.0 (high activity).

2.2. Jackknifing. Jackknifing[41] was coupled with PLS regression by Martens.[42] The method is used for testing the model's robustness[43] and for improving the overall quality of the regression model.[44] The application of jackknifing to variable selection in QSAR is reported by Anderssen et al.[32] Jackknifing is a closely related aspect of the resampling scheme as well as a statistical significance test according to a distribution. A multivariate calibration model consists of regression coefficient estimates whose significance depends on the uncertainty estimation for each regression coefficient. A recently introduced leave-many-out (LMO) method for computing this uncertainty has been utilized to achieve effective descriptors in QSAR (variable selection purpose) with the jackknife method.

For jackknifing, at first, the data set was divided into two subsets of definite size: (a) a training subset (the first subset), including *SS* (sample size) number of molecules, which was used for model development and (b) a validation subset (the second subset) which was used for validation of prediction ability of the model. This division of data set was performed randomly *SN* (sample number) times. Using the calibration subset, from each sn^{th} ($sn = 1$ to SN) random division of data, a regression model was built, and the obtained regression coefficients for the variables are in vector β (sn, 1:nSD), where nSD is the number of introduced descriptors into the jackknife-based models and nVD is the number of significant descriptors with p-value lower than 0.05. The validation sets were predicted SN times to test the prediction ability of SN models. For SN times formation of regression models with randomly selected set of samples, the obtained regression coefficients were recorded in a $SN \times nSD$ matrix (β). The p-value for each sd^{th} ($sd = 1$ to nSD) selected descriptor was calculated by p-value $= p(z(sd))$, where

$$z(sd) = \frac{0 - \text{mean}(\beta(1{:}SN, sd))}{\text{std}(\beta(1{:}SN, sd))} \tag{1}$$

Each descriptor with a p-value lower than 0.05 was considered significant (nVD). It was due to the remarkable difference of the mean value of regression coefficients of the descriptor from zero. MLR and PLS were used as regression methods in jackknifing. After calculating β in each step, the prediction ability of constructed model was determined by the estimation of validation set.

$$Q^2 = 1 - \frac{[\sum_{i=1}^{N_{\text{VALID}}} (y_i - \hat{y}_i)^2]}{[\sum_{i=1}^{N_{\text{VALID}}} (y_i - \bar{y})^2]} \tag{2}$$

where q^2_{TOT} is sum of q^2 for all SN randomly generated regression models.

Table 2. Descriptors of 53 Physicochemical Properties

variable	description	NSDL4	nucleophilic superdelocalizability for atom
ATCH1	partial atomic charge for atom	NSDL5	nucleophilic superdelocalizability for atom
ATCH2	partial atomic charge for atom	NSDL6	nucleophilic superdelocalizability for atom
ATCH3	partial atomic charge for atom	NSDL7	nucleophilic superdelocalizability for atom
ATCH4	partial atomic charge for atom	NSDL8	nucleophilic superdelocalizability for atom
ATCH5	partial atomic charge for atom	NSDL9	nucleophilic superdelocalizability for atom
ATCH6	partial atomic charge for atom	NSDL10	nucleophilic superdelocalizability for atom
ATCH7	partial atomic charge for atom	VDWVOL	van der Waals volume
ATCH8	partial atomic charge for atom	SURF_A	surface area
ATCH9	partial atomic charge for atom	MOFI_X	principal moments of inertia
ATCH10	partial atomic charge for atom	MOFI_Y	principal moments of inertia
DIPV_X	dipole vector	MOFI_Z	principal moments of inertia
DIPV_Y	dipole vector	PEAX_X	principal ellipsoid axes
DIPV_Z	dipole vector	PEAX_Y	principal ellipsoid axes
DIPMOM	dipole moment	PEAX_Z	principal ellipsoid axes
ESDL1	electrophilic superdelocalizability for atom	MOL_WT	molecular weight
ESDL2	electrophilic superdelocalizability for atom	S8_1DX	substituent dimensions
ESDL3	electrophilic superdelocalizability for atom	S8_1DY	substituent dimensions
ESDL4	electrophilic superdelocalizability for atom	S8_1DZ	substituent dimensions
ESDL5	electrophilic superdelocalizability for atom	S8_1CX	substituent centers
ESDL6	electrophilic superdelocalizability for atom	S8_1CY	substituent centers
ESDL7	electrophilic superdelocalizability for atom	S8_1CZ	substituent centers
ESDL8	electrophilic superdelocalizability for atom	LOGP	partition coefficient
ESDL9	electrophilic superdelocalizability for atom	M_PNT	melting point
ESDL10	electrophilic superdelocalizability for atom	SUM_F	sum of F substituent constant
NSDL1	nucleophilic superdelocalizability for atom	SUM_R	sum of R substituent constant
NSDL2	nucleophilic superdelocalizability for atom		
NSDL3	nucleophilic superdelocalizability for atom		

$$q^2_{TOT} = 1 - \frac{[\sum_{sn=1}^{SN}\sum_{i=1}^{N_{VALID}}(y_{i,sn} - \hat{y}_{i,sn})^2]}{[\sum_{sn=1}^{SN}\sum_{i=1}^{N_{VALID}}(y_{i,sn} - \bar{y}_{sn})^2]} \quad (3)$$

2.3. Gram−Schmidt Orthogonalization. Gram−Schmidt orthogonalization (GSO) decorrelates the variables according to their order. GSO steps are described as follows. Assuming the first descriptor $k(0) = \mathbf{x}_j$, ($k(0)$ = starting vector)), a column of $\mathbf{X}_{cal}$ that has maximum correlation with $\mathbf{y}$ (activity) can be chosen as the first $k(0)$ according to Pearson correlation $r = \mathbf{X}_{cal}'\,\mathbf{y}/(I-1)$, and number N is given.

Step 0: Before the first iteration ($n = 1$), let $\mathbf{x}_j = j$th column of $\mathbf{X}_{cal}$; $j = 1, ..., J$.

Step 1: Calculate the projection of $\mathbf{x}_j$ on the subspace orthogonal to $\mathbf{x}_{k(n-1)}$ as:

$$\mathbf{P}\mathbf{x}_j = \mathbf{x}_j - (\mathbf{x}_j^T\mathbf{x}_{k(n-1)})\mathbf{x}_{k(n-1)}(\mathbf{x}_{k(n-1)}^T\mathbf{x}_{k(n-1)})^{-1} \quad (4)$$

For all $\mathbf{x}_j \in \mathbf{X}_{cal}$, where $\mathbf{P}$ is the projection operator.

Step 2: Let $k(n) = \arg(\max ||\mathbf{P}\mathbf{x}_j||, \mathbf{x}_j \in \mathbf{X}_{cal})$.

Step 3: Let $\mathbf{x}_j = \mathbf{P}\mathbf{x}_j$, $\mathbf{x}_j \in \mathbf{X}_{cal}$.

Step 4: Let $n = n + 1$; if $n < N$, then go back to Step1.

End: The resulting descriptors are $\{k(n);\ n = 0, ..., N-1\}$.

In this modified (improved) algorithm, there is no need to define any ensemble containing the remaining descriptors that have not been selected yet. It is because the norm value of projection vectors for selected descriptors is zero in the subspace of the next step and may not be selected again. Contrary to this algorithm, in the original algorithm, an ensemble, such as S, was introduced that included the set of descriptors which had not been selected yet.[15] Totally the proposed algorithm in this work was performed using a smaller number of variables and steps.

Both GSO and SPA are obtained from a similar algorithm. However, the purposes for their applications are different. GSO manipulates the data in order to generate a new set of orthogonal vectors, which in general does not have physical meaning. SPA, on the contrary, does not modify the original data vectors, and the projection is used only for a selection purpose.[15] SPA is a variable selection technique designed to minimize the collinearity problem. It is a forward selection method that starts with one descriptor and incorporates a new descriptor in each iteration, until a specified number of variables (N) is obtained.

2.4. Q^2_{F3}. Todeschini et al.[45] introduced a new Q^2, as Q^2_{F3}, for evaluating the predictive ability of models:

$$Q^2_{F3} = 1 - \frac{\sum_{i}^{N_{VALID}}(y_i - \hat{y}_i)^2/N_{VALID}}{\sum_{i}^{N_{TRAIN}}(y_i - \bar{y}_{TRAIN})^2/N_{TRAIN}} \quad (5)$$

The different mathematical behaviors of eqs 2 and 5 were investigated, and it was shown that Q^2 in eq 2 is related to the number and the distribution of objects in an external test (validation) set, but Q^2_{F3} is reported to be independent of size and distribution of the external test (validation) set. It was shown that Q^2_{F3} and root-mean-square error (rmse) are always completely correlated ($r = 1$), and their behavior agrees with each other.[45]

$$\text{rmse} = \sqrt{\frac{\sum_{i=1}^{N_{VALID}}(y_i - \hat{y}_i)^2}{N_{VALID}}} \quad (6)$$

This is due to the fact that the numerator in eq 5 is rmse square and the denominator of the equation is a constant value when the members of the training set do not change. In this way, contrary to Q^2, Q^2_{F3} can be applied to determine the overfitting of the model, in the same way as the rmse can be applied for this purpose.

In this study, the size of training set (SS) was varied from 5 to $N_{TOTAL} - 1$, and in each SS value, a $Q^2_{TOT(F3)}$ was calculated and summed for all different sample numbers.

$$Q^2_{TOT(F3)} = 1 - \frac{[\sum_{sn=1}^{SN}\sum_{i=1}^{N_{VALID}}(y_{i,sn} - \hat{y}_{i,sn})^2/N_{VALID}]}{[\sum_{sn=1}^{SN}\sum_{i=1}^{N_{TOTAL}}(y_{i,sn} - \bar{y}_{TOTAL})^2/N_{TOTAL}]} \quad (7)$$

$$\text{rmsecv} = \sqrt{\frac{\sum_{sn=1}^{SN}\sum_{i=1}^{N_{VALID}}(y_i - \hat{y}_i)^2}{SN \times N_{VALID}}} \quad (8)$$

where rmsecv and $Q^2_{TOT(F3)}$ have rank correlation, and their results are in agreement with each other.

3. RESULTS AND DISCUSSION

3.1. Variable Selection. The great interest in a QSAR model establishment is the determination of the relevant variables that properly describe dependencies between activity and chemical structures of compounds. More often than not interrelation between variables skews the results, and therefore, nonpredictive QSAR models may be obtained.[16] To eliminate the interrelation between descriptors and to perform an initial selection of descriptors, GSO was applied to all 53 descriptors. Changing the first starting vector ($k(0)$) and assigning the number of orthogonalized descriptors equal to the rank of data matrix (N), we obtained a $\mathbf{X}_{GSO}(31 \times N)$ matrix of orthogonalized descriptors. In fact, by orthogonalization we can remove uninformative descriptors and parts of selected descriptors (collinearity among descriptors) and avoid model overfitting.

The maximum number of selectable orthogonalized descriptors is equal to the rank of data matrix, and in the absence of rank deficiency, the maximum number is equal to the minimum of number of rows and columns. Hence, 30 orthogonal descriptors were obtained from the application of GSO on Selwood data set of 31 molecules. The norm of successively selected projected descriptors vs the selected descriptor number shows the end point of the operation. The 30th projected descriptor has the norm = 0.38, whereas the 31st has the norm = 6.8×10^{-14} and the rank of $\mathbf{X}_{GSO,6}(31 \times 31)$(selected descriptors) is 30. In addition, many of correlation values of the 31st projected descriptor with other descriptors are nonzero. It indicates that there is no remaining

Table 3. Selected (Ranked) Descriptors Using GSO Technique and Jackknife Procedure on Orthogonalized Data from GSO[a]

starting vector	method	sorted descriptors
$k(0) = 6$	GSO	6, 20, 46, 13, 29, 48, 44, 4, 41, 49, 8, 11, 50, 19, 2, 52, 53, 43, 42, 45, 16, 31, 51, 17, 14, 12, 35, 30, 38, 22
$k(0) = 6$	jackknife	6, 4, 11, 46, 50, 16, 52, 22, 20, 41, 43, 42, 49, 38, 13, 19, 14, 35, 12, 53, 2, 51, 48, 17, 9, 8, 30, 31, 29, 44
$k(0) = 50$	GSO	50, 29, 13, 8, 4, 49, 20, 41, 52, 48, 44, 31, 19, 2, 53, 46, 42, 35, 16, 51, 43, 6, 17, 14, 11, 12, 45, 30, 38, 22
$k(0) = 50$	jackknife	50, 52, 31, 48, 35, 22, 16, 49, 6, 42, 29, 4, 51, 45, 11, 38, 30, 2, 43, 8, 9, 12, 13, 14, 17, 19, 20, 41, 44, 46

[a] Descriptor selection (ranking) was performed using both 6th and 50th descriptor, as the starting vectors in GSO.

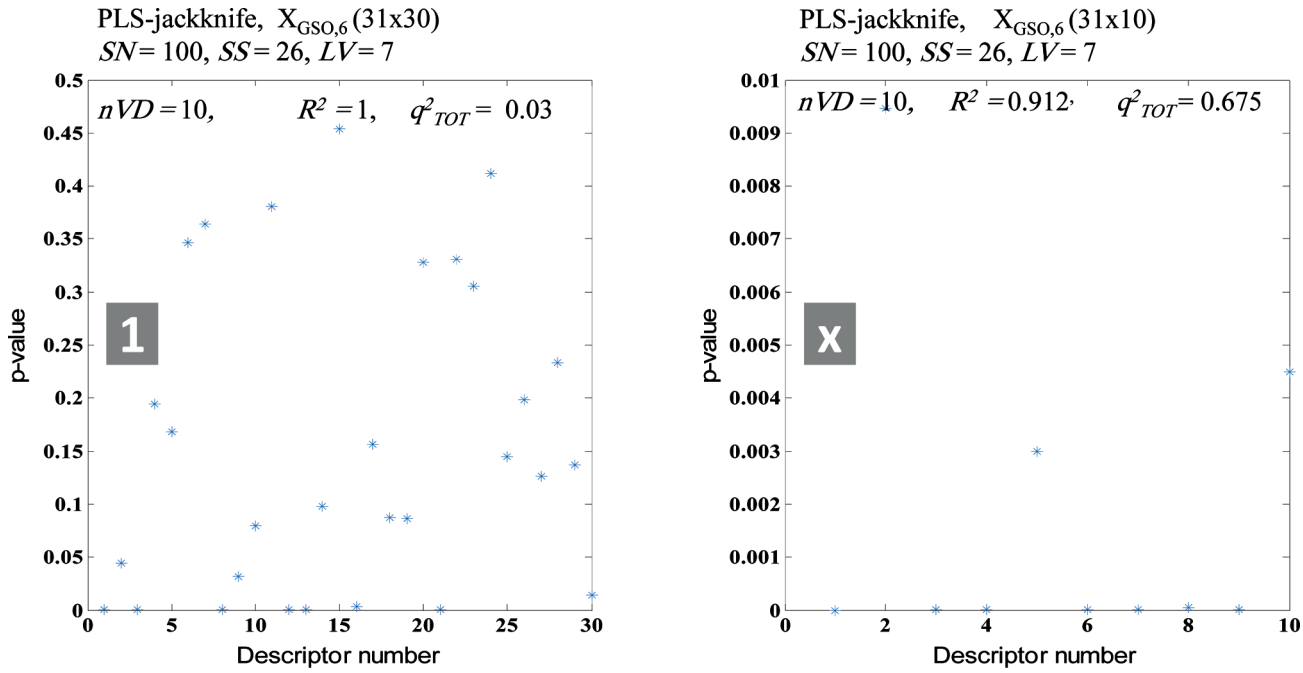


Figure 1. Calculated *p*-values for 30 and 10 descriptors applying PLS-jackknife on GSO orthogonalized Selwood data in one step. The last step (the 2nd) was not accounted for because there was no reduction in the number of descriptors in that stage.

independent information in the 31st descriptor, and it only includes unwanted fluctuations. The 30 orthogonalized descriptors after applying GSO are listed in Table 3.

One remarkable point about orthogonalization is that in selecting almost any of descriptors in the first row of Table 3 as $k(0)$, the chosen set of descriptors from GSO are almost the same (less than two descriptors would be different). It is also the case with the descriptors in the third row of the table. Selecting a few number of descriptors that do not exist in that table, as $k(0)$, resulted in completely different sets of descriptors. For instance, by using both $k(0) = 6$ and 17 the same set of descriptors would be obtained from GSO, whereas using $k(0) = 1$ it led to 7 different descriptors.

The other point is that, if PCA is applied to autoscaled data, then pure PCs cannot be achieved. This is due to the fact that the criterion for selecting PCs is the direction of maximum variance. By pure PCs, it is meant that each PC is only due to changes in one descriptor. On the other hand, applying PCA to orthogonalized data resulted in pure PCs. One acceptable reason is that the values of singular values are equal to the norm of selected projected descriptor in each stage, which is descending.

3.2. MLR-Jackknife with Orthogonal Data. MLR-jackknife (a jackknifing procedure including MLR as a regression method) was applied to the autoscaled Selwood data $\mathbf{X}_{\text{auts}}$ (31×53) to make a proper regression model between $\mathbf{X}_{\text{auts}}$ and the vector of activities, $\mathbf{y}$ (31×1). Applying MLR-jackknife, *p*-values were computed for each descriptor. No significant descriptor was found, and all *p*-values were higher than 0.05. So, the resulted model did not include any satisfactory (validated) parameters and overfitting occurred. Even when the number of variables (columns) in the matrix of orthogonalized descriptors was less than or equal to the number of molecules (rows), none of the descriptors was selected by applying MLR-jackknife because none had a *p*-value lower than 0.05.

In the case of PLS-jackknife on autoscaled data, after three steps of backward elimination, a set of descriptors with *p*-values lower than 0.05 was obtained. However, the final set of validated descriptors was not reproducible. Running the third step led into 18 significant descriptors with $q^2_{\text{TOT}} = 0.55$. Totally considering both theory and practice, variable elimination steps seemed necessary after both MLR- and PLS-jackknife when utilizing the autoscaled data.

In the next stage, MLR-jackknife was applied to $\mathbf{X}_{\text{GSO},6}(31 \times 26)$. The applied conditions were $SS = 26$ and $SN = 100$. $\mathbf{X}_{\text{GSO},6}(31 \times 26)$ means that orthogonalized 31×26 data was used and that $k(0)$ was 6. To estimate the Moore−Penrose pseudoinverse without rank-deficiency problem during the MLR procedure, the number of applied orthogonalized

descriptors (by GSO) was adjusted to a value lower than or equal to *SS*. For instance, in the case of $SS = 26$ for the regression coefficients to be calculated properly, the data containing 26 descriptors were considered. Using different sets of a limited number of descriptors, proper models can result from MLR. However, in a number of different proper models, the regression coefficient of a significant descriptor can be negative, and for the other sets, it could be positive. Hence the totally observed distribution will not be different from zero, and the descriptors will be distinguished as nonsignificant. This is the case for most of the descriptors when applying MLR-jackknife to a large number of descriptors (orthogonalized or not orthogonalized autoscaled data).

3.3. PLS-Jackknife with Orthogonal Data. On orthogonalized data $\mathbf{X}_{GSO,6}(31 \times nSD)$ in condition, $nSD = 30$, $LV = 7$ (the number of latent variables was optimized as $LV = 7$ by applying leave-10-out cross validation during application of PLS on all samples in autoscaled data), $SS = 26$, $SN = 100$, and $k(0) = 6$, PLS-jackknife was utilized. Ten descriptors were found as significant with $q^2_{TOT} = 0.030$ and $R^2 = 1.000$ in one step that shows overfitting as a result of large numbers of descriptors. Doing PLS-jackknife on the $nSD = nVD = 10$ selected descriptors in a second step, all of the 10 were significant with $q^2_{TOT} = 0.675$ and $R^2 = 0.912$, as shown in Figure 1. Results show that PLS-jackknife had a better final q^2_{TOT} value and was more reproducible. Estimation of q^2_{TOT} from the set of validated descriptors resulted from the first step of PLS-jackknife, in a second step, will be performed in the next sections of this report.

3.4. Optimization of Parameters for Variable Selection. *3.4.1. Determination of Sample Size.* For determination of the optimum sample size (*SS*) when applying PLS-jackknife to the orthogonalized data, $\mathbf{X}_{GSO,6}(31 \times 30)$, it was varied from 5 to 30. The number of validated descriptors (*nVD*) in each *SS* value was determined. At all different values of *SS*, the condition $nSD = 30$, $LV = \min(7, nSD)$, and $SN = 100$ was applied. Table 4 contains the list of validated descriptors in each sample size. The table illustrates the successive addition of descriptors into the list of validated descriptors, as the *SS* increases. It is due to the narrower distribution of the estimated regression coefficients during the resampling in jackknife. As the *SS* increases, it causes a less extent of overlap of the distribution with zero. Proportionality of *nVD* to *SS* is illustrated in Figure 2a. In some *SS* values, one unit increase in *SS* results in no addition of more descriptors to the model, but in some other values of *SS*, it causes introduction of more than one descriptor into the model. This shows the similarity of distribution of regression coefficients for a number of descriptors.

As *SS* increases, the *nVD* increases, and the optimum number of selected descriptors in the final model (as the result of selection of the optimum *SS* value) is a main question in variable selection. The effect of *SS* on q^2_{TOT} and R^2 is shown in plots Figure 2b and c. To obtain these two plots, as *SS* increased from 5 to 29, the applied *nSD* was adjusted on the *nVD* that corresponded to the considered *SS* value in each step. To have information about the distribution of the parameters, box plots were drawn for 10 time repetitions. At each specified value of *SS*, a PLS-jackknife was applied to the validated descriptors—$\mathbf{X}_{GSO,6}(31 \times nSD)$ and $nSD = nVD$—to estimate the q^2_{TOT} and R^2 values. Up to $SS = 19$, the q^2_{TOT} value was proportional to *SS* and as a result proportional to *nVD*. For the *SS* values higher than 19, there was a level-off in q^2_{TOT} value. It shows that there was no improvement in the prediction ability as a result of an increase in *nVD* over 19. Hence, $SS = 19$ was selected as the optimum value, with $q^2_{TOT} = 0.68$ and $nSD = nVD = 6$. In all different values of *SS*, R^2 that was the representative of calibration fitting was increasing with an increase in *SS* (and *nVD*). Considering that q^2_{TOT} is also a function of *nSD* (= *nVD*), that is increasing in proportion to *SS*, the net and individual functionality of q^2_{TOT} to *SS* and *nSD* is not distinguished.

Table 4. Selected Descriptors by Increasing Sample Size (*SS*) in Applied One Step PLS-Jackknife on $\mathbf{X}_{GSO,6}$ (31×30)[a]

sample size	valid descriptors
5	6
6	6
7	6
8	6
9	6
10	6
11	6, 4
12	6, 4, 11
13	6, 4, 11
14	6, 4, 11
15	6, 4, 11
16	6, 4, 11, 46
17	6, 4, 11, 46, 50
18	6, 4, 11, 46, 50
19	6, 4, 11, 46, 50, 16
20	6, 4, 11, 46, 50, 16
21	6, 4, 11, 46, 50, 16
22	6, 4, 11, 46, 50, 16
23	6, 4, 11, 46, 50, 16, 52
24	6, 4, 11, 46, 50, 16, 52, 22
25	6, 4, 11, 46, 50, 16, 52, 22, 20
26	6, 4, 11, 46, 50, 16, 52, 22, 20, 41
27	6, 4, 11, 46, 50, 16, 52, 22, 20, 41, 43, 42, 49
28	6, 4, 11, 46, 50, 16, 52, 22, 20, 41, 43, 42, 49, 38, 13, 19, 14
29	6, 4, 11, 46, 50, 16, 52, 22, 20, 41, 43, 42, 49, 38, 13, 19, 14, 35, 12, 53
30	6, 4, 11, 46, 50, 16, 52, 22, 20, 41, 43, 42, 49, 38, 13, 19, 14, 35, 12, 53, 2, 8, 9, 17, 29, 30, 31, 44, 45, 48

[a] The procedure could be utilized to sort the descriptors according to the order of their introduction into the model. The applied settings in PLS-jackknife was $SN = 100$ and $LV = \min(7, nVD)$.

Therefore, descriptors were sorted according to their stability and order of their introduction into the model as a function of *SS*. In order to eliminate the effect of sample size on q^2_{TOT} and to find the net effect of increase in *nSD,* the sorted validated descriptors were entered one by one into the model. By increasing *nSD* at fixed values of *SS*, q^2_{TOT} vs *nSD* was drawn in Figure 3a. According to the information given in the previous parts, $SS = 19$ was considered as optimum. To see the net effect of *nSD* on the value of q^2_{TOT}, sorted variables by PLS-jackknife on $\mathbf{X}_{GSO,50}(31 \times nSD)$, Table 3, were manually introduced into the model one by one and at fixed values of *SS* and q^2_{TOT} vs *nSD* drawn in Figure 3b. The figure shows the value of q^2_{TOT} as a function of *nSD* at different fixed values of *SS* from 15 to 26. It shows that overfitting takes place as a result of introducing extra variables. At $SS = 15$, the maximum q^2_{TOT} is about 0.64 that shows the optimum number of selected variables equal to 5. And the optimum number of selected variables at $SS = 26$ was 17 (with $q^2_{TOT} = 0.88$). The results show the net effect of *nSD* on the obtained q^2_{TOT}, and effect of *SS* on q^2_{TOT} is separated.

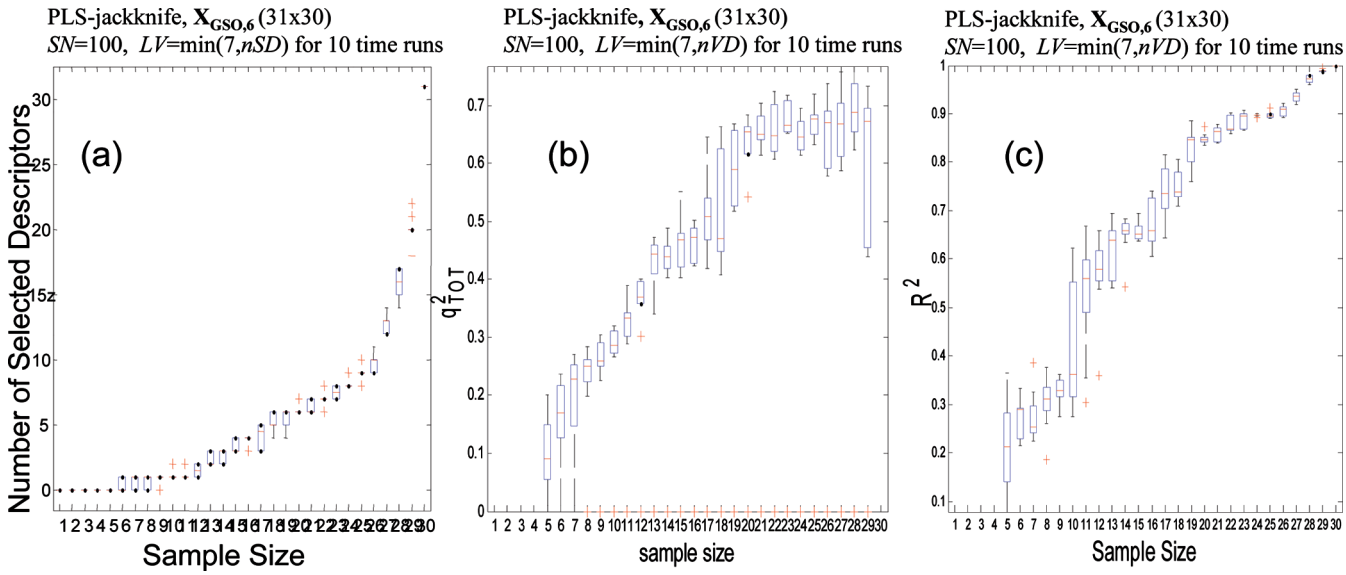


Figure 2. Boxplots for 10 time runs of PLS-jackknife on GSO orthogonalized Selwood data with 30 variables at different sample sizes. (a) *nSD* vs *SS;* (b) q^2_{TOT} vs *SS,* and (c) R^2 vs *SS. SN* was equal to 100 for all plots and $LV = \min(7,nVD)$.

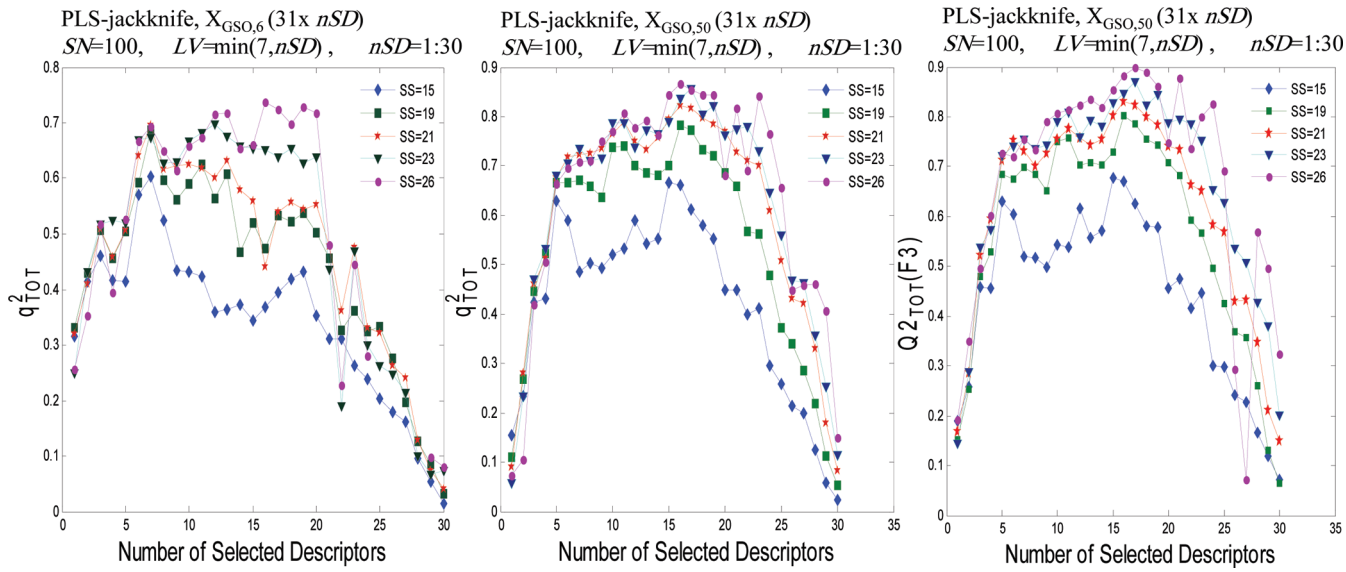


Figure 3. Plots of q^2_{TOT} (a and b) and $Q^2_{TOT(F3)}$ (c) as a function of the number of selected descriptors (from sorted descriptors in 2nd or 4th row of Table 3), at different fixed values of sample size. In all plots PLS-jackknife was applied to GSO orthgonalized data. Starting variable in GSO was the sixth for (a) and the 50th for (b) and (c).

It is reported[45] that an alternative way to eliminate the effect of *SS* on the estimated q^2_{TOT} value is the utilization of $Q^2_{TOT(F3)}$, which is not related to sample size. The plot of $Q^2_{TOT(F3)}$ as a function of number of the introduced variables, Figure 3c, was very similar to that of q^2_{TOT}, Figure 3b. As shown in this figure, the obtained $Q^2_{TOT(F3)}$ plots at different values of *SS* are completely different. It seems that rmsecv and $Q^2_{TOT(F3)}$ are functions of *SS* as well as *nSD*. It is worth noting that as the selected descriptors are from the same set of sorted valid descriptors, they are the same for each value of *nSD*, although the *SS* is changed. It shows that the change in $Q^2_{TOT(F3)}$ in different *SS* values is not due to change in selected descriptors. So the effect of the number of selected variables on rmsecv and $Q^2_{TOT(F3)}$ was investigated at different fixed sample sizes. So, even when the number of objects is small (like in this work), the results obtained from q^2_{TOT} and $Q^2_{TOT(F3)}$ are similar.

Descriptors sorted by GSO with different $k(0)$ values and those sorted by order of introduction of orthogonalized variables (by GSO) into model by increasing the sample size in PLS-jackknife procedure are shown in Table 3. As it can be seen, the order of the sorted variables is different in various conditions. When two or more descriptors were introduced simultaneously into the model by an one unit increase in *SS*, they were ranked according to their *p*-values.

3.4.2. Starting Vector k(0). The affecting parameter on the order of orthogonalized descriptors according to GSO algorithm is the starting vector that determines the value of resulting q^2_{TOT}. Selecting any descriptor with high correlation with **y** (activity), as the first vector $k(0)$ selected, results in a high value of q^2_{TOT}. In this work, the sixth descriptor showed the highest absolute *r* value ($r(6) = 0.606$) and the 14th descriptor gave the lowest ($r(14) = 3.6e-4$). At $k(0) = 14$ and after selection of 30 descriptors by Gram−Schmidt, $\mathbf{X}_{GSO}$ (31×30) was obtained. PLS-jackknife was run 20 times and in two steps at conditions $LV = 7$, $SS = 26$, and $SN = 100$, and the result was $q^2_{TOT} = 0.31$ and $R^2 = 0.856$ with a common set of 10 descriptors for all 20 times of running.

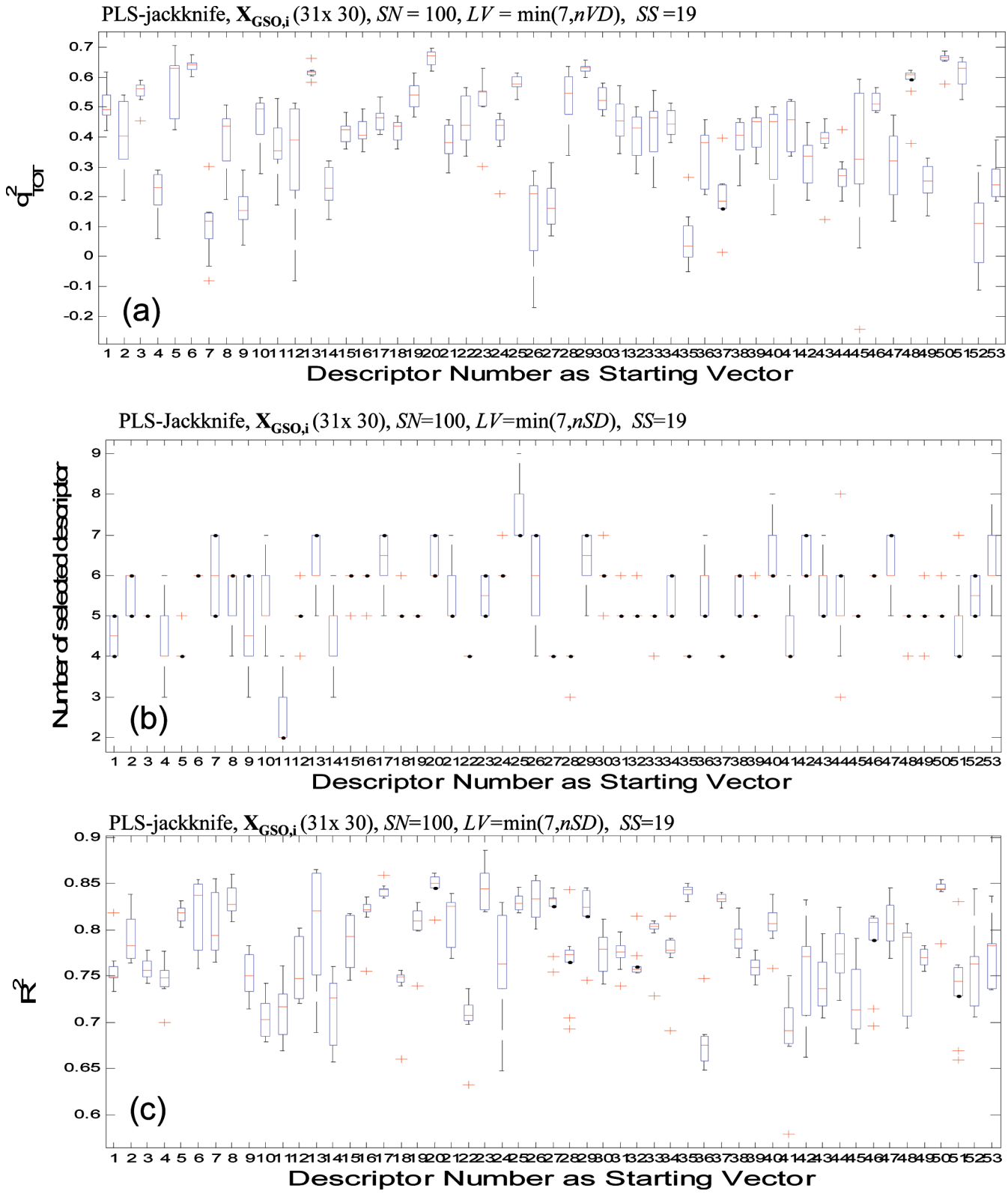


Figure 4. Boxplots from applying a one step PLS-jackknife on GSO orthogonalized data, using different starting descriptor in GSO algorithm, $k(0)$. Ten time runs resulted in boxplots. (a) Boxplot for q^2_{TOT} against $k(0)$; (b) boxplot for nVD vs $k(0)$; and (c) boxplot for R^2 as a function of $k(0)$.

When $k(0) = 6$ was used, at previous conditions, the results obtained were $q^2_{TOT} = 0.65$ and $R^2 = 0.852$, and the number of commonly selected descriptors was 7. Therefore, $k(0) = 6$ was preferred, and the determining effect of correlation of the first descriptor on q^2_{TOT} value and the selected descriptors in the final model was observed.

All 53 descriptors were checked to choose the optimum $k(0)$, according to the amount of q^2_{TOT} and nVD. The value of $k(0)$ was varied between all initial descriptors and by applying GSO; 53 different sets of orthogonalized data matrix were obtained. At $\mathbf{X}_{GSO,i}$ (31 × 30), i = 1, ..., 53, PLS-jackknife was run once, and by using significant descriptors,

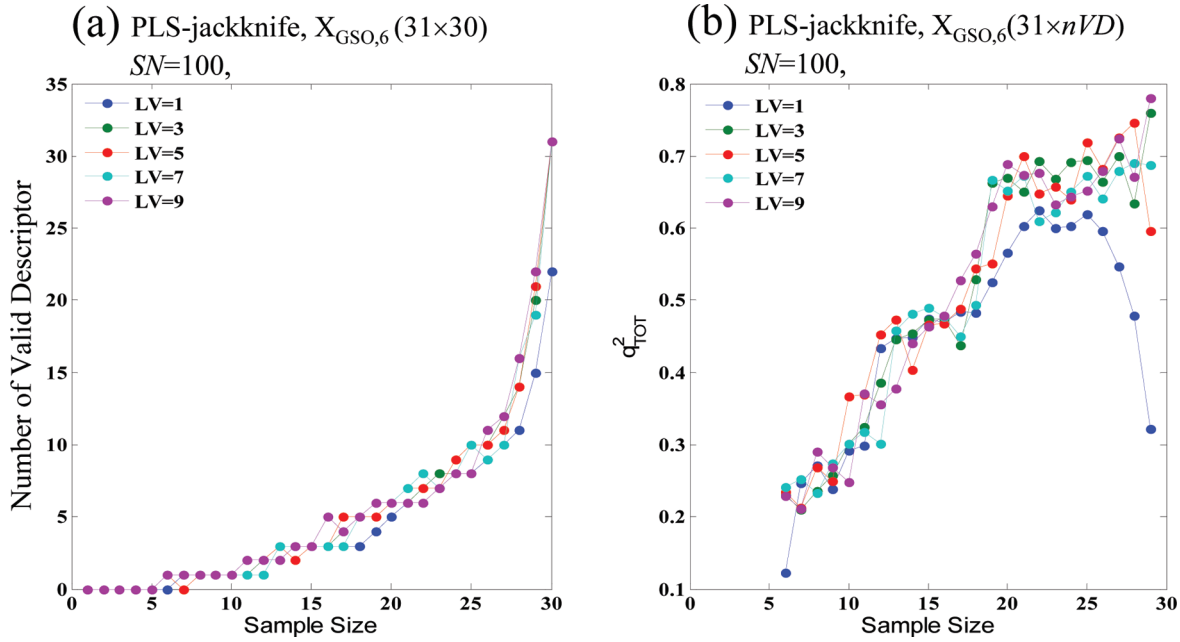


Figure 5. (a) *nVD* vs *SS* for a 1 step PLS-jackknife on $\mathbf{X}_{GSO,6}$ (31 × 30) (with *SN* = 100, *i* = 1:2:9, *LV* = min (*i*,*nVD*)), at different values of *LV*. (b) q^2_{TOT} against *SS* for a 1 step PLS-jackknife on $\mathbf{X}_{GSO,6}$ (31 × 30) (with *SN* = 100, *LV* = min(*i*,*nVD*)). As the *SS* increases, the q^2_{TOT} is increasing and is not remarkably depend on *LV*.

q^2_{TOT} was estimated and the whole process was repeated 10 times. The obtained q^2_{TOT} and final number of selected descriptors and R^2 at different values of *k*(0) are shown in Figure 4a−c.

By using *k*(0) = 6 as the starting vector, 6 descriptors that included [4, 6, 11, 16, 46, 50] were selected. By using *k*(0) = 50 (correlation of 50th descriptor with **y** was lower than sixth), the maximum value of q^2_{TOT} and 5 significant descriptors were obtained ([50, 52, 48, 31, 35] ≡ [LOGP, SUM-F, S8-1CY, NSD-7, VDWVOL]). According to the results in Figure 4a, *k*(0) = 50 was selected because the median q^2_{TOT} for 10 time repetitions for *k*(0) = 6 was equal to 0.63 and for *k*(0) = 50 was equal to 0.683.

3.4.3. Effect of LV. Effect of *LV* on *nVD* and q^2_{TOT} was investigated. At each different value of *LV* from 1 to 10, sample size (*SS*) varied from 5 to 29. And using the obtained descriptors with *p*-values smaller than 0.05 (significant descriptors) in each step, the other PLS-jackknife (as a second step) was run and the obtained number of significant descriptors and q^2_{TOT} values were recorded (Figures 5(a) and 5(b)). Plots show that increasing *LV* higher than 2 has no significant effect on the values of *nVD* and q^2_{TOT}, and that the value of q^2_{TOT} is increasing in proportion to sample size. The curve for the *LV* = 1 is different from that for *LV* ≥ 2, when *SS* is higher than 19.

3.4.4. Determination of Sample Number (SN). Sample number (*SN*) had no effect on the number and the type of validated descriptors as well as on q^2_{TOT} values. It only affected the distribution of estimated regression parameters for descriptors (the width of Gaussian distribution) obtained during running PLS-jackknife on the orthogonalized data. Reduction of the distribution widths, as a result of increase in *SN*, only leads to the introduction of the same variables in lower values of *SS*. By increasing the sample number, one can see a slight decrease in the q^2_{TOT} standard deviation from different resampling runs in PLS-jackknife procedures. Considering the computation time, we can conclude that *SN* = 100 was sufficient to obtain a proper set of descriptors in a short period of time. High values of *SN* resulted in higher repeatability of q^2_{TOT}, despite a longer period of time. The resulted box plot that shows the distribution of q^2_{TOT} values, from 200 times running of a one-step PLS-jackknife on $\mathbf{X}_{GSO,50}$ (31 × 30) at different values of *SN* (in conditions *LV* = 7, validated variables: [50, 52, 31, 35, 48]) is shown in Figure 6.

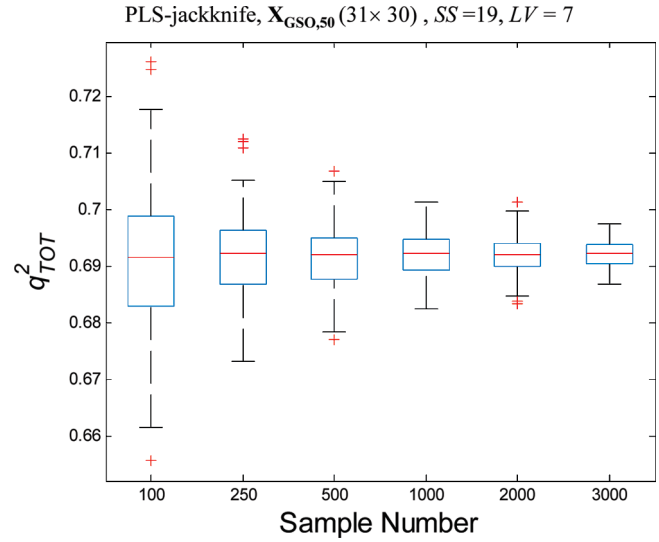


Figure 6. Box plot of q^2_{TOT} from 200 times running of a one-step PLS-jackknife on $\mathbf{X}_{GSO,50}$ (31 × 30) at *SS* = 19, *LV* = 7, and different values of *SN* (from 100−3000).

3.4.5. Final Model. PLS-jackknife was applied to orthogonalized descriptors $\mathbf{X}_{GSO,50}$ (31 × 30) in conditions *LV* = 7, *SN* = 100, and *SS* = 19. The five obtained significant descriptors with *p*-values lower than 0.05 from the first step were investigated using another PLS-jackknife as a next step, and in the same condition and as a consequence, the same five descriptors were confirmed. The one-step procedure was repeated 100 times, and average values of q^2_{TOT} and R^2 were obtained as 0.693 and 0.811, respectively. The obtained significant descriptors, the corresponding multiple regression

Table 5. Comparison between Selected Descriptors and Model Parameters in This Work and Previous Selwood Data Set Models

size	descriptors						R^2	q^2	who[a]
5	ATCH4	DIPV-X	MOFI-Z	LOGP	SUM-F		0.818	0.699	K
5	ATCH4	ATCH5	DIPV-X	MOFI-Y	LOGP		0.826	0.696	KM
5	ATCH4	ATCH5	DIPV-X	MOFI-Z	LOGP		0.826	0.696	K
5	S8-1CY	NSDL-7	VDWVOL	LOGP	SUM-F		0.811	0.693	our work
5	ATCH4	DIPV-X	MOFI-Y	LOGP	SUM-F		0.813	0.692	K
6	ATCH4	ESDL3	VDWVOL	LOGP	SUM-F	MOFI-X	0.827	0.677	WA
4	DIPV-Y	MOFI-Y	LOGP	SUM-F			0.745	0.655	T
3	MOFI-Y	SUM-F	LOGP				0.721	0.647	TRKCL
4	MOFI-Y	MOL-WT	LOGP	SUM-F			0.740	0.646	T
3	ESDL3	SURF-A	LOGP				0.719	0.644	RKCL
4	MOFI-Y	M-PNT	LOGP	SUM-F			0.749	0.644	TR
3	MOFI-Z	SUM-F	LOGP				0.718	0.643	KCL
4	ESDL3	VDWVOL	LOGP	MOFI-X			0.742	0.639	T
4	NSDL2	MOFI-Y	LOGP	SUM-F			0.736	0.639	T
4	ATCH-4	ESDL-3	LOGP	PEAX-X			0.774	0.636	RKL
4	ATCH-4	ATCH-5	LOGP	MOFI-Z			0.772	0.624	KL
4	ATCH-4	ATCH-5	MOFI-Y	LOGP			0.768	0.621	RKL
3	ATCH-4	ATCH-5	DIPV-X				0.688	0.612	M
4	ATCH-4	ESDL3	SURF-A	LOGP			0.747	0.609	R
4	ATCH-4	ATCH-5	PEAX-X	LOGP			0.762	0.606	KL
3	ESDL3	MOFI-Y	LOGP				0.702	0.604	RKCL
3	ESDL3	MOFI-Z	LOGP				0.697	0.601	C
3	PEAX-X	SUM-F	LOGP				0.688	0.598	TRC
4	ATCH-4	ESDL3	MOFI-Y	LOGP			0.750	0.592	R
3	ESDL3	PEAX-X	LOGP				0.683	0.589	TR
2	LOGP	M-PNT					0.531	0.397	S

[a] The authors of the models are indicated by the following labels: C = Cho,[37] K = Kubinyi,[12] L = Luke,[38] M = McFarland,[39] R = Rogers,[8] S = Selwood,[33] T = Todeschini,[40] and WA = Waller.[34]

coefficients and the confidence limits for the parameters are all given in the following equation:

$$-\log(EC_{50}) = 0.5003(\pm 0.0809)\,\mathrm{LOG\,P} + 0.4490(\pm 0.0645)\,\mathrm{SUM} - \mathrm{F} - 0.4293 (\pm 0.0857)\mathrm{S8} - 1\mathrm{CY} - 0.2427(\pm 0.0929)\,\mathrm{NSDL7} - 0.2998(\pm 0.0672)\,\mathrm{VDWVOL}$$

The obtained sets of five descriptors were exactly the same in all the 100 times of repetitions, which shows the robustness and the stability of the final five variable models. In Table 5, the selected descriptors and the model parameters for different Selwood data set models with different variable selection methods are compared. The selected descriptors in this report are more or less similar to the selected descriptors in other papers. It can be seen that the selected descriptors in this study resulted in one of the best values of q^2_{TOT} and R^2 compared to other reports.

Selwood is a small data set, and selecting a small-sized external test set from this data is useless because it will likely give poor estimates of predictive ability and more importantly the final result will be a random estimate. So, the proposed method was applied to anti-HIV data,[16] which is a larger set. It contains 107 molecules of 1-[2-hydroxyethoxy)methyl]-6-(phenylthio)thymine (HEPT) derivations and 160 calculated descriptors for each molecule. Thirty-five molecules were selected randomly and considered as an external validation set. These 35 molecules included: {39, 66, 37, 35, 77, 67, 42, 13, 32, 12, 92, 91, 93, 11, 22, 50, 62, 8, 52, 100, 2, 57, 72, 83, 16, 21, 47, 75, 97, 76, 6, 85, 34, 79, 38} from ref 16. The 90th descriptor that showed maximum correlation with **y** was considered as the starting vector to start the projection operation, and 107 orthogonalized descriptors were obtained. *SS* was changed from 5 until $N_{\mathrm{total,HIV}} - 1$, and descriptors were sorted according to their entrance into the model. Optimum number of descriptors for predicting external set was $nVD = 8$ (90, 91, 36, 56, 111, 76, 63, 116), $q^2_{\mathrm{TOT}} = 0.710$, and $R^2 = 0.819$. Results showed that this variable selection strategy could successfully select a set of variables that led to a high prediction model.

An advantage of this method is the simple statistical base of the procedure which could be understood easily. The other advantage is ability to order the variables according to their importance. The final obtained set of variables is reproducible, which is not the case for many other variable selection procedures. However the method includes an orthogonalization step and needs projection of the test set onto the orthogonalized calibration set. Sometimes, this part of the procedure does not lead to acceptable results for small data sets. Orthogonalization of variables in the subspace, which is vertical to the previously selected variable, prevents similar (coinformative) variables from being selected. This could be another limitation for the present procedure.

CONCLUSION

Using partial least squares (PLS) as a regression method inside the jackknife procedure resulted in a set of selected significant descriptors with acceptable values of q^2_{TOT}. However, when multiple linear regression (MLR) was used, no descriptor was selected as significant. Hence, PLS-jackknife was used to find the significant regression parameters. By using the Gram−Schmidt algorithm as a descriptor-thinning method, before QSAR modeling, one can show that the collinearity and the redundant variables were removed and that the number of selected descriptors was reduced without losing any valuable information.

Applying PLS-jackknife to an orthogonalized data set caused a smaller reproducible set of validated descriptors and a higher prediction ability. Considering the autoscaled data and doing the three steps of PLS-jackknife, 18 descriptors were validated as significant. However, by dealing with orthogonalized data and performing only one step of PLS-jackknife, a reproducible set of 10 descriptors was selected. The sample size in PLS-jackknife was used to rank the descriptors for being selected one by one. Although parameters, such as the starting variable for GSO, the sample size and number, and the number of latent variables were considered and optimized in this report, the only crucial parameter which can be regarded in future works for being optimized is the sample size. Totally the whole procedure seems simple and practical.

REFERENCES AND NOTES

(1) Xu, H.; Liu, Z.; Cai, W.; Shao, X. A wavelength selection method based on randomization test for near-infrared spectral analysis. *Chemom. Intell. Lab. Syst.* **2009**, *97*, 189–193.

(2) Nadler, B.; Coifman, R. R. The prediction error in CLS and PLS: the importance of feature selection prior to multivariate calibration. *J. Chemom.* **2005**, *19*, 107–118.

(3) Baumann, K. Cross-validation as the objective functions for variable-selection techniques. *Trends Anal. Chem.* **2003**, *22*, 395–406.

(4) Topliss, J. G.; Costello, R. J. Chance correlations in structure-activity studies using multiple regression analysis. *J. Med. Chem.* **1972**, *15*, 1066–1068.

(5) Kraker, J. J.; Hawkins, D. M.; Basak, S. C.; Natarajan, R.; Mills, D. Quantitative Structure-Activity Relationship (QSAR) modeling of juvenile hormone activity: Comparison of validation procedures. *Chemom. Intell. Lab. Syst.* **2007**, *87*, 33–42.

(6) Hawkins, D.; Basak, S.; Shi, C. QSAR with Few Compounds and Many Features X. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 663–670.

(7) Alexandridis, A.; Patrinos, P.; Sarimveis, H.; Tsekouras, G. A two-stage evolutionary algorithm for variable selection in the development of RBF neural network models. *Chemom. Intell. Lab. Syst.* **2005**, *75*, 149–162.

(8) Rogers, D. R.; Hopfinger, A. J. Application of genetic function approximation to quantitative structure-activity relationships and quantitative structure-property relationships. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 854–866.

(9) Shamsipour, M.; Zare-Shahabadi, V.; Hemmateenejad, B.; Akhond, M. An efficient variable selection method based on the use of external memory in ant colony optimization. Application to QSAR/QSPR studies. *Anal. Chim. Acta* **2009**, *646*, 39–46.

(10) Burden, F. R.; Ford, M. G.; Whitley, D. C.; Winkler, D. A. Use of automatic relevance determination in QSAR studies using Bayesian neural networks. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 1423–1430.

(11) Sutter, J. M.; Dixon, S. L.; Jurs, P. C. Automated Descriptor Selection for Quantitative Structure-Activity Relationships Using Generalized Simulated Annealing. *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 77–84.

(12) Kubinyi, H. Variable Selection in QSAR Studies. I. An Evolutionary Algorithm. *Quant. Struct.-Act. Relat.* **1994**, *13*, 285–294.

(13) Agrafiotis, D. K.; Cedeno, W. Feature Selection for Structure-Activity Correlation Using Binary Particle Swarms. *J. Med.Chem.* **2002**, *45*, 1098–1107.

(14) Chong, I1-G.; Jun, C.-H. Performance of some variable selection methods when multicollinearity is present. *Chemom. Intell. Lab. Syst* **2005**, *78*, 103–112.

(15) Araújo, M. C. U.; Saldanha, T. C. B.; Galvaõ, R. K. H.; Yoneyama, T.; Chame, H. C.; Visani, V. The successive projections algorithm for variable selection in spectroscopic multicomponent analysis. *Chemom. Intell. Lab. Syst.* **2001**, *57*, 65–73.

(16) Kompany-Zareh, M.; Akhlaghi, Y. Application of radial basis function networks and successive projection algorithm in a QSAR study of anti-HIV activity for a large group of HEPT derivatives. *J. Chemom.* **2006**, *20*, 1–12.

(17) Kompany-Zareh, M.; Akhlaghi, Y. Correlation Weighted successive projections algorithm as a novel method for variable selection in QSAR studies: investigation of anti-HIV activity of HEPT derivatives. *J. Chemom.* **2007**, *21*, 239–250.

(18) Ye, S.; Wang, D.; Min, S. Successive projections algorithm combined with uninformative variable elimination for spectral variable selection. *Chemom. Intell. Lab. Syst.* **2008**, *91*, 194–199.

(19) Teófiloa, R. F.; Martins, J. P. A.; Ferreira, M. M. C. Sorting variables by using informative vectors as a strategy for feature selection in multivariate regression. *J. Chemom.* **2009**, *23*, 32–48.

(20) Hemmateenejad, B.; Javadnia, K.; Elyasi, M. Quantitative structure-retention relationship for the Kovats retention indices of a large set of terpenes: A combined data splitting-feature selection strategy. *Anal. Chim. Acta* **2007**, *592*, 72–81.

(21) Lučić, B.; Nadramija, D.; Ivan Bašic, I.; Trinajstić, N. Toward Generating Simpler QSAR Models: Nonlinear Multivariate Regression versus Several Neural Network Ensembles and Some Related Methods. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1094–1102.

(22) Basak, S. C.; Mills, D.; Mumtaz, M. M. A quantitative structure-activity relationship (QSAR) study of dermal absorption using theoretical molecular descriptors. *SAR QSAR Environ. Res.* **2007**, *18*, 45–55.

(23) Randić, M. Resolution of Ambiguities in Structure-Property Studies by Use of Orthogonal Descriptors. *J. Chem. Inf. Comput. Sci.* **1991**, *31*, 311–320.

(24) Morales, A. H.; Pérez, M. A. C.; Combes, R. C.; Gonźalez, M. P. Quantitative structure activity relationship for the computational prediction of nitrocompounds carcinogenicity. *J. Toxicol.* **2006**, *220*, 51–62.

(25) Reino, J. L.; Saiz-Urra, L.; Ndez-Galan, R. H.; Aran, V.; Hitchhoock, P. B.; Hanson, J. R.; Gonzalez, M. P.; Collado, I. G. Quantitative Structure-Antifungal Activity Relationships of Some Benzohydrazides against Botrytis cinerea. *J. Agric. Food Chem.* **2007**, *55*, 5171–5179.

(26) Basak, S. C.; Natarajan, R.; Mills, D.; Hawkins, D. H.; Kraker, J. J. Quantitative Structure-Activity Relationship Modeling of Juvenile Hormone Mimetic Compounds for Culex Pipiens Larvae, with a Discussion of Descriptor-Thinning Methods. *J. Chem. Inf. Comput. Sci.* **2006**, *46*, 65–77.

(27) Konovalov, D. A.; Sim, N.; Deconinck, E.; Heyden, Y. V.; Coomans, D. Statistical Confidence for Variable Selection in QSAR Models via Monte Carlo Cross-Validation. *J. Chem. Inf. Model.* **2008**, *48*, 370–383.

(28) Westad, F.; Martens, J. Variable selection in near infrared spectroscopy based on significance testing in partial least square regression. *J. Near Infrared Spectrosc.* **2000**, *8*, 117–124.

(29) Daszykowski, M.; Wrobel, M. S.; Czarnik-Matusewicz, H. Walczak, B.Near-infrared reflectance spectroscopy and multivariate calibration techniques applied to modelling the crude protein, fiber and fat content in rapeseed meal. *Analyst* **2008**, *133*, 1523–1531.

(30) Wehrens, R.; Putter, H.; Buydens, L. M. C. The bootstrap: a tutorial. *Chemom. Intell. Lab. Syst.* **2000**, *54*, 35–52.

(31) Dietrich, S. W.; Dreyer, N. D.; Hansch, C.; Bentley, D. L. Confidence Interval Estimators for Parameters Associated with Quantitative Structure-Activity Relationships. *J. Med. Chem.* **1980**, *23*, 1201–1205.

(32) Anderssen, E.; Dyrstad, K.; Westad, F.; Martens, H. Reducing over-optimism in variable selection by cross-model validation. *Chemom. Intell. Lab. Syst.* **2006**, *84*, 69–74.

(33) Selwood, D. L.; Livingstone, D. J.; Comley, J. C. W.; O'Dowd, A. B.; Hudson, A. T.; Jackson, P.; Jandu, K. S.; Rose, V. S.; Stables, J. N. Structure-Activity Relationships of Antifilarial Antimycin Analogues: A Multivariate Pattern Recognition Study. *J. Med. Chem.* **1990**, *33*, 136–142.

(34) Waller, C. L. Development and Validation of a Novel Variable Selection Technique with Application to Multidimensional Quantitative Structure-Activity Relationship Studies. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 345–355.

(35) Liu, S.-S.; Liu, H.-L.; Yin, C.-S.; Wang, L.-S. VSMP: A Novel Variable Selection and Modeling Method Based on the Prediction. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 964–969.

(36) Whitley, D. C.; Ford, M. G.; Livingstone, D. J. Unsupervised Forward Selection: A Method for Eliminating Redundant Variables. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 1160–1168.

(37) Cho, S. J.; Hermsmeier, M. A. Genetic Algorithm Guided Selection: Variable Selection and Subset Selection. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 927–936.

(38) Luke, B. T. Evolutionary Programming Applied to the Development of Quantitative Structure- Activity Relationships and Quantitative Structure-Property Relationships. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 1279–1287.

(39) McFarland, J. W.; Gans, D. J. On Identifying Likely Determinants of Biological Activity in High-Dimensional QSAR. *Quant. Struct.-Act. Relat* **1994**, *13*, 11–17.

(40) Todeschin, R.; Consonni, V.; Mauri, A.; Pavan, M. Detecting "bad" regression models: multicriteria fitness functions in regression analysis. *Anal. Chim. Acta* **2004**, *515*, 199–208.

(41) Efron, B. Bootstrap methods: Another look at the jackknife. *Ann. Stat.* **1979**, *1*, 1–26.

(42) Martens, H.; Martens, M. modified jack-knife estimation of parameter uncertainty in bilinear modeling by partial least square regression (PLSR). *Food Qual. Prefer.* **2000**, *11*, 5–16.
(43) Sacan, M. T.; Erdem, S. S.; Ozpınar, G. A.; Balcıoglu, I. A. QSPR Study on the Bioconcentration Factors of Nonionic Organic Compounds in Fish by Characteristic Root Index and Semiempirical Molecular Descriptors. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 985–992.
(44) Huuskonen, J. QSAR Modeling with the Electrotopological State: TIBO Derivatives. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 425–429.
(45) Consonni, V.; Ballabio, D.; Todeschini, R. Comments on the Definition of the *Q*2 Parameter for QSAR Validation. *J. Chem. Inf. Model.* **2009**, *49*, 1669–1678.

CI100169P