Journal of CHEMOMETRICS

Received: 5 November 2013, Revised: 20 January 2014, Accepted: 23 January 2014, Published online in Wiley Online Library: 17 February 2014

(wileyonlinelibrary.com) DOI: 10.1002/cem.2603

Reshaped Sequential Replacement for variable selection in QSPR: comparison with other reference methods

F. Grisoni, M. Cassotti and R. Todeschini*

The objective of the present work was to compare the Reshaped Sequential Replacement (RSR) algorithm with other well-known variable selection techniques in the field of Quantitative Structure–Property Relationship (QSPR) modelling. RSR algorithm is based on a simple sequential replacement procedure with the addition of several 'reshaping' functions that aimed to (i) ensure a faster convergence upon optimal subsets of variables and (ii) reject models affected by chance correlation, overfitting and other pathologies. In particular, three reference variable selection methods were chosen for the comparison (stepwise forward selection, genetic algorithms and particle swarm optimization), aiming to identify benefits and drawbacks of RSR with respect to these methods. To this end, several QSPR datasets regarding different physical–chemical properties and characterized by different objects/variables ratios were used to build ordinary least squares models; in addition, some well-known (Y-scrambling) and more recent (*R*-based functions) statistical tools were used to analyse and compare the results. The study highlighted the good capability of RSR to find optimal subsets of variables in QSPR modelling, comparable or better than those found by the other reference variable selection methods. Moreover, RSR resulted to be faster than some of the analysed variable selection techniques, despite its extensive exploration of the variables space. Copyright © 2014 John Wiley & Sons, Ltd.

Keywords: variable selection; Reshaped Sequential Replacement; QSPR; QSAR

1. INTRODUCTION

Variable selection (VS) is a key step in multivariate analysis for modelling purposes. It consists in the selection of optimal subsets of variables, in order to obtain parsimonious models, with maximum predictive power and increased interpretability. VS plays a crucial role in scientific fields that deal with a large number of variables, such as Quantitative Structure–Property/Activity Relationship (QSPR/QSAR). QSPR and QSAR are based on the assumption that the structure of a molecule is responsible for its physical, chemical and biological properties. The QSPR (and QSAR) approach can be generally described as an application of statistical and mathematical methods to the issue of finding empirical relationships expressed in the form

$$Y_i = f(x_1, x_2, \ldots, x_p)_i$$

where Y_i is the property of interest of the i-th compound, x_1, x_2, …, x_p are the p predictors of the i-th molecule and f represents the mathematical relationship between independent variables and the property. Molecular descriptors are used as predictors. They can be defined as 'the final result of a logic and mathematical procedure that transforms chemical information of a molecule, such as structural features, into useful numbers or the result of standardized experiments' [1].

Nowadays it is possible to calculate thousands of different descriptors. However, according to Occam's razor principle [2], it is reasonable to assume that only a small number of them are correlated to the experimental response and are, therefore, relevant for building the mathematical model of interest. Furthermore, one fundamental aspect is to find the good trade-off between bias and complexity of the model. The increase of the complexity due to a larger number of descriptors included in the model is able to improve the fitness to the training data, but the inclusion of too many variables can often cause a reduction in the predictive ability, leading to overfitting. On the other hand, if the model is too simple, the bias will increase, and the model will not be able to capture important relationships between predictors and response, leading to underfitting. The optimal subset of variables is reflected in a good predictive ability, more robustness and stability of the model [3].

In this scenario, VS plays a key role in QSAR/QSPR, allowing to select the optimal subset of molecular descriptors for modelling the activity/property of interest and to obtain robust and predictive models. In this way, also the interpretability of the models increases, and non-significant effects can be neglected.

Throughout the years, many different methods and techniques have been proposed to address the problem of VS (e.g. [4,5]). From the classical approaches (e.g. stepwise Backward Elimination (BE) and Forward Selection (FS) [6]) to more sophisticated VS methods. In recent years, the so-called nature-inspired methods [7–9], such as Genetic Algorithms (GA) [10], Particle

* Correspondence to: R. Todeschini, Milano Chemometrics and QSAR Research Group—Department of Earth and Environmental Sciences, University of Milan–Bicocca, P.za della Scienza 1, 20126 Milan, Italy.
E-mail: roberto.todeschini@unimib.it

F. Grisoni, M. Cassotti, R. Todeschini
Milano Chemometrics and QSAR Research Group—Department of Earth and Environmental Sciences, University of Milan–Bicocca, P.za della Scienza 1, 20126 Milan, Italy

Swarm Optimization (PSO) [11], Ant Colony Optimization [12] and Evolutionary Programming [13], have progressively increased in importance. Moreover, the group of penalization techniques, such as Least Absolute Shrinkage and Selection Operator (LASSO) [14] and elastic net [15], recently gained interest from the scientific community to address the issue of VS: These methods were initially aimed at improving the problems of ordinary least squares (OLS) regression and are able to select variables via the shrinkage of the regression coefficients towards zero.

Recently, we proposed a VS method based on Miller's Sequential Replacement (SR) [16], the Reshaped Sequential Replacement (RSR) [17]. Being based on the same replacement procedure, the two methods share a good exploration capability. Some new reshaping features were included in the RSR algorithm in order to (i) decrease the computational time (ensuring a faster convergence towards optimal solutions) and (ii) identify models suffering from different types of pathologies. Our previous study highlighted the capability of the method to perform a good exploration of the space of the variables and of the new reshaping functions to significantly speed up the modelling time and discard models that suffer from different pathologies, such as overfitting or chance correlation.

In the present study, we compared RSR with other widely used VS methods: (i) stepwise FS; (ii) GA; and (iii) PSO. RSR and the reference methods were applied to four QSPR datasets in regression with different properties and objects/variables ratios. The primary objectives were to compare the performances of these VS methods in the field of QSPR modelling and identify benefits and drawbacks of RSR method with respect to the others.

After a brief introduction about reference VS methods (Section 2), the theory of RSR algorithm (Section 3) and details about the materials and methods (Section 4) are presented. Results and discussion can be found in Section 5.

2. REFERENCE VARIABLE SELECTION METHODS

2.1. Stepwise regression

Stepwise regression (SWR) methods [6] are among the most known feature selection methods. SWR is based on two different strategies, namely Forward Selection (FS) and Backward Elimination (BE). FS starts with a model of size 0 and proceeds by adding variables that fulfil a pre-defined criterion. BE method proceeds in the opposite way with respect to FS: It starts from a model of size p (p being the total number of variables), and non-relevant variables are eliminated in a step-by-step procedure. Typically, the inclusion (or exclusion) criterion is the residual sum of squares (RSS): At each step, the variable to be added (or eliminated) is the one that leads to the maximum decrease (or minimum increase) of the RSS.

2.2. Genetic Algorithms

Genetic Algorithms are a nature-inspired method [10,18,19] that takes inspiration from Darwin's theory of evolution. In analogy with biological systems, each gene represents a variable, and each chromosome (sequence of genes/variables) can be seen as a potential model. The evolution of the population of chromosomes is determined by two processes: (i) crossover, in which pairs of chromosomes generate offspring according to a crossover probability; and (ii) mutation, in which some genes of a chromosome can change according to a mutation probability. Every time a new chromosome with a better fitness function (e.g. Q^2_{cv}) than already existing ones is generated, it enters the population and the worst model is discarded. In this way, chromosomes compete against each other, and only the fittest survive, in analogy with Darwin's concept of 'survival of the fittest'.

2.3. Particle Swarm Optimization

Particle Swarm Optimization is an agent-based method inspired by the behaviour of flock of birds [20,21]. Differently from GA, PSO agents do not compete but cooperate in order to find the best solutions. PSO was initially thought of as an optimization method and only later modified in order to address the problem of VS [11]. PSO agents are particles that move in a binary space (in the variant for VS) in which each dimension corresponds to a variable and each position to a model. The particle motion is controlled by a parameter called static probability, which determines the probability of each particle to move to its previous personal best position, to the best global position or to remain in its current position, balancing exploration and exploitation ability of the method.

3. RESHAPED SEQUENTIAL REPLACEMENT ALGORITHM

The RSR method is based on the SR method proposed by Miller in 1984 [16]. The basic idea of Miller's method is to start from a randomly generated model (seed), replace each variable at a time with all the remaining ones and see whether a better model can be obtained. The best model found in the first replacement procedure becomes the new seed for a further replacement. This procedure goes on until no better models can be found. Miller's method has the advantage of performing a good exploration of the variables space, but with the drawback of being extremely time consuming when the number of variables increases.

The RSR method [17] implements new reshaping functionalities over Miller's algorithm that aim to:

(1) decrease the calculation time, retaining the exploration capability of the method;
(2) increase the probability of convergence upon the optimal models;
(3) identify models that suffer from several pathologies, such as overfitting, chance correlation, variable redundancy and collinearity between predictors.

Moreover, the coefficient of determination in cross-validation (Q^2_{cv}, see Section 4.2) is used as a fitness function instead of the RSS used in the original SR algorithm, the latter not necessarily being related with the predictive ability of the model.

The functions able to 'reshape' the original method are as follows:

(1) Tabu list (TL): Preliminary exclusion of variables not correlated with the response according to their univariate Q^2_{cv} in regression. It aims at decreasing the computational time with respect to SR algorithm. Variables are excluded according to the following criterion:

$$\text{if } Q^2_{cv}(\boldsymbol{y},\boldsymbol{x}) < 0 \Rightarrow \boldsymbol{x} \in TL \qquad (1)$$

When the algorithm reaches convergence, tabu variables are re-included and used for a last replacement procedure starting from the optimal population of models. Tabu variables will be selected in each seed only if they provide an improvement to the model larger than a pre-defined threshold (e.g. 0.01 on Q^2_{cv}). TL resulted to be the principal function able to decrease the computational time with respect to SR algorithm (up to 10 times faster).

(2) Roulette wheel (RW): Used for the initialization of the population. Each variable is given a probability of entering the initial population proportional to a chosen fitness function (univariate Q^2_{cv} in regression). This pre-selection algorithm is supposed to generate models closer to the optimal solution, being biased towards promising variables.

(3) *QUIK* rule: A statistical test [22] used in regression during the replacement procedure, in order to reject *a priori* models affected by high predictor collinearity. The collinearity among variables is one of the main problems when applying multiple linear regression that can lead to undesirable consequences [23,24]. The *QUIK* rule is based on the *K* multivariate correlation index [25] and the comparison between the internal correlation of the *X*-block (K_X) and the correlation of the *X*-block plus the *y* response (K_{Xy}):

$$\text{if } K_{xy} - K_x < \delta K \quad \Rightarrow \quad \text{reject the model} \tag{2}$$

The basic assumption is that the total correlation of the independent variables (***X***) selected in the model plus the response (***y***) should be larger than the total correlation calculated on the selected independent variables only. If this criterion is not fulfilled, the model is rejected before being statistically evaluated.

(4) Evaluation functions: Implemented to evaluate the final population of models.

(i) *R*-function-based rules [22] to identify: (a) models with redundancy in explanatory variables (R^P index) and (b) models with noisy variables (R^N index).

(ii) Y-scrambling (in regression): A statistical randomization test [26] commonly used to identify the presence of chance correlation between predictors and response.

(iii) Canonical Model Correlation (CMC) and Canonical Model Distance (CMD) [27]: For the comparison of final models. CMC and CMD allow an easy comparison of the final models in order to determine whether models with different variables are actually different in their nature.

(iv) Nested models screening. A model *F* can be defined as 'nested' if there is a model *G* of higher size (i.e. including more variables) that comprises all the variables of *F* and has a very similar performance (i.e. the difference in their Q^2_{cv} is smaller than a pre-defined threshold, e.g. 0.005). If this occurs, model *G* is rejected because its higher complexity is not balanced by a better performance.

A simplified flowchart of the algorithm is depicted in Figure 1.

4. MATERIALS AND METHODS

4.1. Variable selection strategies

4.1.1. *Stepwise regression*

Stepwise regression was performed with FS using the maximization of the coefficient of determination in cross-validation (Q^2_{cv}) as criterion for the progressive inclusion of each variable up to a maximum model size, chosen as stop criterion.

4.1.2. *Genetic Algorithms*

In addition to the classic GA approach, based on a single run with a randomly initialized population, the version of GA proposed by Leardi and González was used. This version aims to overcome the principal limitations of GA, i.e. the tendency to overfit data and to model noise if the response is noisy and/or a limited number of objects is present (or the ratio objects/variables is small). The approach is based on (i) execution of a large number of runs with different randomly generated initial populations; (ii) optimization of the number of evaluations for each run; and (iii) final stepwise selection approach, based on the frequency of selection of each variable over all the runs. Two further features characterize the algorithm by Leardi and González: (i) for the principle of parsimony and to prevent overfitting, a chromosome *M* cannot enter the population if another chromosome *F* exists that has a higher fitness and is a subset of the variables of *M*; and (ii) GA can be hybridized with a BE procedure that is carried out during or at the end of each run. In order to distinguish the classic approach from the approach of Leardi and González, in this work, they were identified as GA and GA-SW, respectively.

4.1.3. *Particle Swarm Optimization*

In the modified PSO of Shen *et al.* [11], the balance between exploration and exploitation ability of the method is intended to change during the motion of the particles; therefore, the static probability starts with a value equal to 0.5 and decreases to a final value equal to 0.33. According to PSO approach for VS, variables that are not included in the initial random population cannot be included during the run, thus being the exploration capability of this method limited. For this reason, the strategy of Leardi and González (execution of a large number of runs, optimized number of evaluations, check for nested models and BE, and final stepwise) was also adopted for PSO. This version was referred to as PSO-SW.

4.2. Model validation

For all VS methods, the coefficient of determination in cross-validation (Q^2_{cv}) was used as fitness function. Q^2_{cv} is defined as follows:

$$Q^2_{cv} = 1 - \frac{\sum_{i=1}^{n_{train}} \left(y_i - \hat{y}_{i/i} \right)^2}{TSS} \tag{3}$$

where n_{train} is the number of training objects, y_i is the real response value of the *i*-th object and $\hat{y}_{i/i}$ is the value of the *i*-th object predicted by the model in which the *i*-th object was not taken into consideration; *TSS* (total sum of squares) is the sum of squared deviations from the dataset mean. In this work, a leave-more-out strategy was used with a 'venetian blind' resampling technique that makes the values of Q^2_{cv} calculated on different models comparable and consistent. Furthermore, the predictive power of the models was assessed also by means of external validation on a test set. This was expressed by the coefficient of determination on the external test set (Q^2_{ext}), calculated as follows [28]:

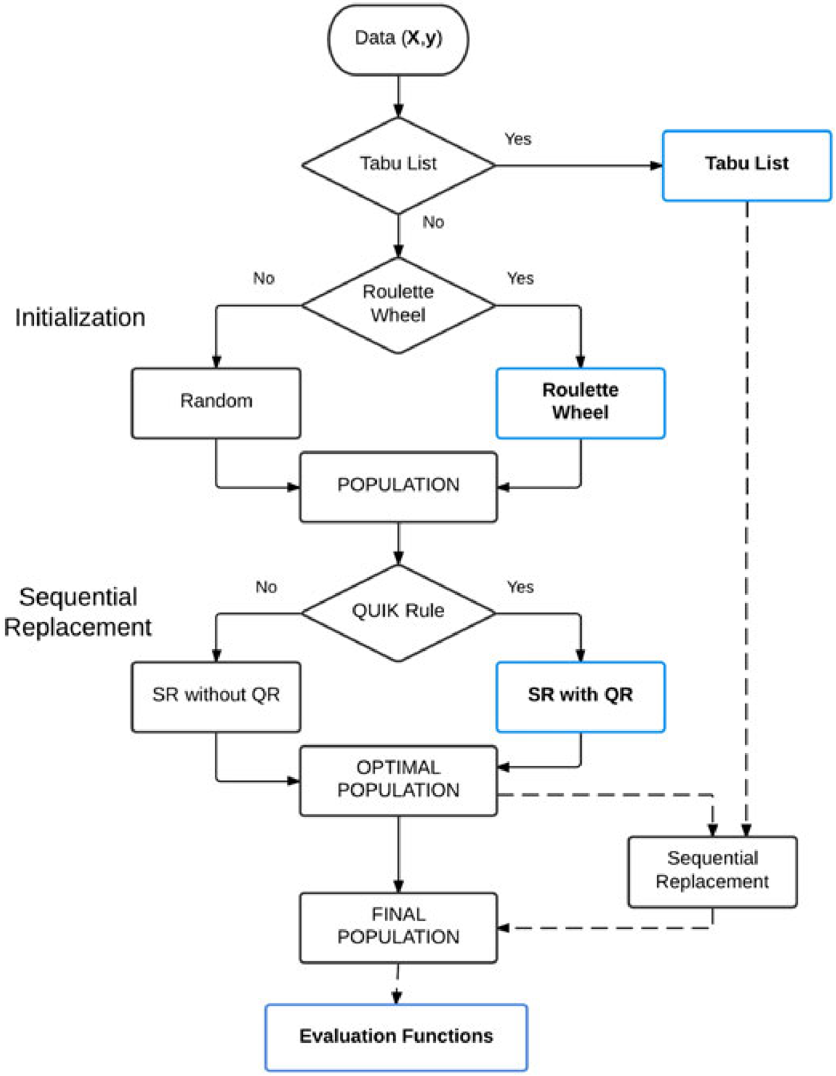

Figure 1. RSR method: simplified flowchart of the algorithm. The new 'reshaping' functions are highlighted in boldface.

$$Q^2_{ext} = 1 - \frac{\left(\sum_{i=1}^{n_{ext}} (y_i - \hat{y}_i)^2\right) / n_{ext}}{TSS/n_{train}} \quad (4)$$

where n_{ext} is the number of objects in the external test set; $\hat{y}_i$ and y_i are the predicted response and the real response of the *i*-th test object, respectively; n_{train} is the number of objects in the training set; and *TSS* is the total sum of squares calculated on the training set.

Finally, in order to represent the ability of the model to fit the training data, the coefficient of determination was also calculated:

$$R^2 = 1 - \frac{\sum_{i=1}^{n_{train}} (y_i - \hat{y}_i)^2}{TSS} \quad (5)$$

where $\hat{y}_i$ and y_i represent the calculated response and the real response of the *i*-th object, respectively. R^2 represents the percentage of the variance explained by the model.

4.3. Datasets

In the present work, comparisons were made on four QSPR datasets that were retrieved from US EPA website and had been used in T.E.S.T. software to develop models [29]. The chosen properties are (i) boiling point (BP), (ii) vapour pressure (VP), (iii) thermal conductivity (TC) and (iv) flash point (FP).

For each dataset, molecular descriptors from 0D to 2D were calculated by means of Dragon 6 [30]. Constant, near constant and descriptors having a standard deviation lower than 0.1 were deleted. Variables showing a pair correlation larger than 0.95 with other descriptors were deleted. The original random splitting between training and external test set used in T.E.S.T. software was retained. The characteristics of the analysed datasets are reported in Table I.

4.4. Software and codes

In the present work, calculations were performed using MATLAB R2012b [31] on an Intel Xeon CPU E5-2620 0 at 2.00 GHz with 16 GB RAM.

Table I. QSPR datasets: name, property, number of objects in training (n_{train}) and external test set (n_{ext}) and number of variables (p) are reported

Dataset	Property	n_{train}	n_{ext}	p
BP	Normal boiling point	4607	1151	823
VP	Vapour pressure at 25 °C	2006	504	937
TC	Thermal conductivity at 25 °C	352	90	566
FP	Flash point	6690	1672	1008

All the MATLAB toolboxes and functions used in this study were written by our research group. The RSR toolbox will be soon available for free download on Milano Chemometrics website [32].

5. RESULTS AND DISCUSSION

The aim of the present work was to compare our method, RSR algorithm, with some reference VS methods, i.e. Stepwise Forward Selection (SWR), Genetic Algorithms (GA and GA-SW) and Particle Swarm Optimization (PSO-SW). To this end, the methods were applied to four QSPR datasets in regression.

For the sake of comparison, for all the VS techniques, OLS was always used as regression method and the internal validation was carried out by means of a fivefold cross-validation (Q^2_{cv}). The maximum dimension of the generated models was arbitrarily set to six for all the methods and all the datasets. This allows (i) an 'internal' comparison of all the VS methods on the same dataset and (ii) a 'global' comparison between the results of the same method on different datasets.

For RSR, three seeds were generated for each model size. RW, TL and the evaluation functions were enabled with the default thresholds; *QUIK* rule was disabled in order to compare the exploration capability of each method based only on Q^2_{cv}.

For both the GA and GA-SW approaches, crossover and mutation probabilities were set to 0.5 and 0.01, respectively, and the number of chromosomes was set to 30. For the classic GA approach, a single run was performed with a different number of evaluations depending on the model size: 500 (from two to three variables), 700 (four variables) and 1000 (from five to six variables). For GA-SW and PSO-SW, 100 runs were performed with an optimized number of evaluations for each dataset.

For PSO-SW, the number of particles was set to 10, and the initial static probability to 0.5 (decreasing to a final value equal to 0.33).

The RSR, GA, GA-SW and PSO-SW algorithms were run three times on each dataset, being meta-heuristic methods. Being our implementation of SWR deterministic, it was run only once on each dataset.

Moreover, in order to have a better understanding of the results, *R*-function based rules and Y-scrambling were also applied *a posteriori* to the final population of models found by each reference method. Only the models that fulfilled Y-scrambling test and R^P and R^N rules were taken into account. Finally, for each dataset, the best model (based on Q^2_{cv}) provided by each method for each model size was used for the comparison.

All the models fulfilled Y-scrambling test, while different percentages of rejection by the *R*-function-based rules were observed on each dataset.

For BP dataset, 53% of the models were discarded by *R*-functions-based rules: all the discarded models did not fulfil R^N rule (presence of noisy variables) with the exception PSO-SW and GA-SW models that did not fulfil R^P rule (excess of explanatory variables). No models of GA-SW were accepted for this dataset.

Table II reports the best model (based on Q^2_{cv}) for each dimension found by each VS method. The models with largest Q^2 values, both in cross-validation and external validation, had five variables, thus showing that the increase in model complexity is not always balanced by an increase in predictive power. RSR provided the model with the largest Q^2_{cv} (81.8%). GA found a model giving very similar Q^2_{cv} (81.7%) and slightly larger Q^2_{ext} (81.6%) compared with that of RSR model (81.1%). RSR and SWR found the same model with three variables. This model could be regarded as the best one because of its simplicity (it comprises only three descriptors) and its very good performance (Q^2_{cv} and Q^2_{ext} only, respectively, 2.8% and 2.2% lower than those of GA model with five variables).

Table II. Results on BP dataset sorted by $\mathbf{Q^2_{cv}}$: statistics, size and descriptors are reported for each method

Method	R^2 (%)	Q^2_{cv} (%)	Q^2_{ext} (%)	Size	Descriptors					
RSR	82.0	81.8	81.1	5	ATS3p	GATS1v	JGT	CATS2D_05_LL	TPSA(NO)	
GA	81.8	81.7	81.6	5	J_D	SM1_B(p)	ATSC1p	B02[F-F]	TPSA(NO)	
GA	79.4	79.3	79.4	6	PCR	J_Dt	AVS_B(p)	SM1_B(p)	F-083	TPSA(NO)
RSR	79.2	79.0	79.4	3	SPI	SM1_B(p)	TPSA(NO)			
SWR	79.2	79.0	79.4	3	SPI	SM1_B(p)	TPSA(NO)			
RSR	78.4	78.3	78.0	4	WiA_Dt	SM3_D/Dt	GATS1v	B02[F-F]		
GA	75.2	75.1	71.4	4	nF	SpMaxA_L	HyWi_B(m)	TPSA(NO)		
GA	73.7	73.6	72.5	3	pilD	SM1_B(p)	ATSC1i			
RSR	70.0	69.9	68.5	2	SM1_B(p)	TPSA(NO)				
SWR	70.0	69.9	68.5	2	SM1_B(p)	TPSA(NO)				
GA	69.2	69.1	69.6	2	pilD	IAC				
PSO-SW	66.8	66.6	67.2	3	piPC07	pilD	X3sol			
PSO-SW	61.9	61.8	62.2	2	SCBO	Xt				

RSR, Reshaped Sequential Replacement; GA, genetic algorithms; GA-SW, genetic algorithms of Leardi and González; PSO-SW, particle swarm optimization, Leardi and González approach; SWR, stepwise regression with forward selection.

The most frequently selected descriptor (in eight out of 13 models) is ‘TPSA(NO)’, which represents the topological polar surface area calculated from polar fragments with nitrogen and oxygen contributions [33]. The 33% of the selected descriptors (14 out of 43) are 2D matrix-based, i.e. topological indices calculated from different graph-theoretical matrices.

Similar results were found for VP dataset. Approximately 43% of the models did not fulfil *R*-functions-based rules. Once again, most of the models found by GA and PSO with the approach of Leardi and González (GA-SW and PSO-SW) did not fulfil the R^P rules. On this dataset, the model that provided the largest statistics, both in cross-validation and external validation, comprised six variables (RSR algorithm), even though the difference in Q^2_{cv} and Q^2_{ext} with its best model with five variables is very small. RSR gave the best model for each dimension, and its best model with three and four variables was also found by SWR (Table III). The most frequently selected descriptors are ‘piID’ [34] (selected in 12 out of 17 models) and ‘TPSA(NO)’ [33] (11 out of 17 models), representing the conventional bond-order ID number and the topological surface area, respectively. The majority of the selected descriptors belong to the group of walk and path counts (24%) and molecular properties (22%).

On TC dataset, the percentage of rejected models was significantly lower (28%) than the previous cases. RSR gave the model with the largest Q^2_{cv} (78.8%) and Q^2_{ext} (77.5%) (Table IV). RSR model with six variables comprises all the descriptors of RSR model with five variables, plus the descriptor ‘O-056’. It is possible to notice that the addition of the descriptor ‘O-056’ leads to only a modest increase in the Q^2_{cv} (2.8% larger) but to a significant increase in the Q^2_{ext} (10.5% larger). The best model provided by GA comprises five variables and had slightly lower Q^2_{cv} than the RSR model of same size, but larger Q^2_{ext}. Unlike the previous datasets, most of the models provided by GA-SW and PSO-SW were accepted by the *R*-functions. The most frequently selected descriptors are (i) ‘O %’ (20 out of 22 models), the percentage of oxygen atoms; (ii) ‘JGT’ [35] (12 models), a global topological charge index; and (iii) ‘X%’ (10 models), the percentage of halogen atoms. The descriptor ‘O-056’ [36,37] (leading to a significant increase in the Q^2_{ext} of RSR model with six variables) represents the number of alcohol fragments in the molecule. In 47% of the cases, the selected descriptors belong to the class of constitutional indices, i.e. the most simple and commonly used descriptors, which reflect the chemical composition of a compound without any information about its molecular geometry or atom connectivity [1].

Regarding FP dataset, R^N and R^P rules rejected approximately 51% of the models. RSR provided the model with the best performance in both cross-validation and external validation (Table V). The best model of RSR had a CMC of 0.85 with the best of SWR (five variables) and of 0.64 with the best of GA (six variables). In other words, even if RSR and SWR models result to be correlated, RSR provides a model with significantly larger Q^2_{cv} (3.2% larger) and Q^2_{ext} (3.5% larger). No models by PSO-SW and three out of six models by GA-SW were accepted by the *R*-based rules. The most frequently selected descriptors are ‘TPSA(NO)’ (12 out of 16 models) and ‘SCBO’ (eight models), the latter representing the sum of conventional bond orders [1]. Molecular properties and constitutional indices are the most frequently selected classes of descriptors, in 20% and 19% of the cases, respectively.

In order to make a global comparison of the results, a ranking of the best models for each dataset (Tables II–5) was made. For each dataset, the final population of models was tailed-ranked according to Q^2_{cv} and Q^2_{ext}, obtaining two matrices. If for a certain model size the method did not provide accepted models, it was given the last position in the ranking. RSR models occupied high positions in the ranking both regarding the performance on training and test sets with the exception of two models (RSR6 and RSR5 on, respectively, BP and FP datasets) that did not fulfil the *R*-functions (Figure 2). GA also provided models with high positions in the ranking, and it was the only method that provided models accepted by the *R*-based rules for each model size on all datasets. On the contrary, GA-SW and PSO-SW, as noticed earlier, often gave models that did not fulfil the *R* rules and, in

Table III. Results on VP dataset sorted by $\mathbf{Q^2_{cv}}$: statistics, size and descriptors are reported for each method

Method	R^2 (%)	Q^2_{cv} (%)	Q^2_{ext} (%)	Size	Descriptors					
RSR	87.1	86.8	88.3	6	S3K	piID	X3sol	GGI3	NssO	TPSA(NO)
RSR	86.5	86.3	88.1	5	nHM	ICR	WiA_Dt	NssO	TPSA(NO)	
RSR	83.9	83.7	85.3	4	piID	Eta_F_A	TPSA(NO)	MLOGP2		
SWR	83.9	83.7	85.3	4	piID	Eta_F_A	TPSA(NO)	MLOGP2		
GA	83.8	83.6	84.6	6	WiA_Dt	SpPosA_B(p)	ATS8m	F03[C-C]	F08[C-O]	TPSA(NO)
GA	83.6	83.4	84.1	5	WiA_Dt	SpPosA_B(p)	ATS8m	F08[C-O]	TPSA(NO)	
RSR	81.6	81.4	82.6	3	piID	Eta_F_A	TPSA(NO)			
SWR	81.6	81.4	82.6	3	piID	Eta_F_A	TPSA(NO)			
GA	81.1	80.7	80.3	4	ICR	piID	CATS2D_02_DA	TPSA(NO)		
GA	76.2	76.1	73.6	3	piID	ATS8m	B04[C-O]			
RSR	76.1	75.9	78.2	2	WiA_Dt	TPSA(NO)				
SWR	75.6	75.4	75.4	2	piID	TPSA(NO)				
GA-SW	74.7	74.5	73.0	3	ICR	piID	X5			
PSO-SW	73.3	73.2	71.7	3	ECC	piPC03	piID			
GA-SW	72.9	72.8	71.4	2	ICR	piID				
GA	71.7	71.4	70.4	2	piID	CATS2D_02_DA				
PSO-SW	68.6	68.5	67.5	2	ECC	piPC03				

RSR, Reshaped Sequential Replacement; GA, genetic algorithms; GA-SW, genetic algorithms of Leardi and González; PSO-SW, particle swarm optimization—Leardi and González approach; SWR, stepwise regression with forward selection.

Table IV. Results on TC dataset, sorted by $\mathbf{Q^2_{cv}}$: statistics, size and descriptors are reported for each method

Method	R^2 (%)	Q^2_{cv} (%)	Q^2_{ext} (%)	Size	Descriptors					
RSR	81.5	78.8	77.5	6	N%	O%	J_Dz(p)	EE_B(m)	JGT	O-056
RSR	78.7	76.0	67.0	5	N%	O%	J_Dz(p)	EE_B(m)	JGT	
GA	77.8	74.8	73.2	5	N%	O%	MAXDN	JGT	nROH	
RSR	75.9	73.4	63.0	4	N%	O%	EE_B(m)	JGT		
GA	72.6	70.9	63.1	4	nN	O%	SM1_Dz(Z)	JGT		
RSR	72.0	69.5	56.6	3	N%	O%	JGT			
SWR	72.0	69.5	56.6	3	N%	O%	JGT			
GA-SW	67.2	66.2	62.3	4	O%	X%	JGT	B02[F-F]		
GA-SW	67.6	65.8	63.4	6	O%	X%	JGT	B01[C-O]	B02[F-F]	MLOGP
GA-SW	67.5	65.8	62.5	5	nO	O%	X%	JGT	B02[F-F]	
GA	68.0	65.2	73.2	6	nN	nO	ATS2m	SpMax4_Bh(s)	Eta_sh_p	O-056
PSO-SW	65.0	62.9	61.6	6	O%	X%	SIC1	SpMax_L	B01[C-O]	B02[F-F]
GA	63.5	62.6	72.1	3	O%	X%	nOHp			
PSO-SW	64.3	62.5	60.2	5	O%	IC1	SpMax_L	JGI1	B01[C-O]	
SWR	63.0	62.2	55.2	2	O%	JGT				
RSR	63.0	62.2	55.2	2	O%	JGT				
PSO-SW	62.5	61.5	62.5	3	O%	X%	SpMax_L			
GA-SW	62.1	61.1	58.7	3	O%	X%	B02[F-F]			
PSO-SW	62.5	60.9	62.6	4	O%	X%	SpMax_L	MLOGP		
GA-SW	59.5	58.7	60.0	2	O%	X%				
PSO-SW	59.5	58.7	60.0	2	O%	X%				
GA	44.4	43.2	38.2	2	B01[C-O]	B02[C-F]				

RSR, Reshaped Sequential Replacement; GA, genetic algorithms; GA-SW, genetic algorithms of Leardi and González; PSO-SW, particle swarm optimization—Leardi and González approach; SWR, stepwise regression with forward selection.

Table V. Results on FP dataset, sorted by $\mathbf{Q^2_{cv}}$: statistics, size and descriptors are reported for each method

Method	R^2 (%)	Q^2_{cv} (%)	Q^2_{ext} (%)	Size	Descriptors					
RSR	81.3	81.1	81.5	6	MW	ZM1Kup	piID	CATS2D_02_DA	T(O..O)	TPSA(NO)
SWR	78.2	77.9	78.0	5	SCBO	MWC02	SM1_Dz(p)	T(O..O)	TPSA(NO)	
GA	77.9	77.7	79.0	6	J_D	AVS_B(m)	SM1_B(p)	nImidazoles	F04[C-O]	TPSA(NO)
RSR	77.7	77.5	78.6	4	J_D	SM1_B(p)	F02[C-O]	TPSA(NO)		
SWR	77.0	76.4	77.2	4	SCBO	SM1_Dz(p)	T(O..O)	TPSA(NO)		
GA	76.5	76.3	78.5	5	piID	X3sol	Eig13_AEA(ri)	F02[C-O]	TPSA(NO)	
RSR	75.1	74.9	77.2	3	piID	SM1_B(p)	TPSA(NO)			
GA	74.3	74.1	75.7	4	SCBO	CATS2D_02_DA	F06[C-O]	TPSA(NO)		
SWR	74.0	73.7	74.8	3	SCBO	T(O..O)	TPSA(NO)			
GA	73.0	72.7	74.4	3	SCBO	F08[C-O]	TPSA(NO)			
RSR	71.4	71.2	73.3	2	SM1_B(p)	TPSA(NO)				
SWR	70.4	70.2	73.7	2	SCBO	TPSA(NO)				
GA-SW	68.5	68.5	71.8	3	SCBO	nN	MWC02			
GA-SW	67.8	67.7	71.2	2	SCBO	nN				
GA-SW	64.1	63.9	67.8	5	MPC07	Eig04_AEA(dm)	Eig11_AEA(ri)	CATS2D_08_DL	B07[C-N]	
GA	62.0	62.0	66.9	2	MWC02	X3sol				

RSR, Reshaped Sequential Replacement; GA, genetic algorithms; GA-SW, genetic algorithms of Leardi and González; PSO-SW, particle swarm optimization—Leardi and González approach; SWR, stepwise regression with forward selection.

particular, the R^P rule (99% of the cases). This could be related to the final stepwise based on the selection frequency of each variable over all the runs. In fact, if two or more relevant variables are correlated, as it is often the case of molecular descriptors, their frequency of selection over all the runs is likely to be similar. This reflects on the inclusion of both variables in the final stepwise model even if they carry the same information, thus causing redundancy in explanatory predictors. These limitations could be connected to the fact that the method was originally proposed for PLS modelling of spectral and

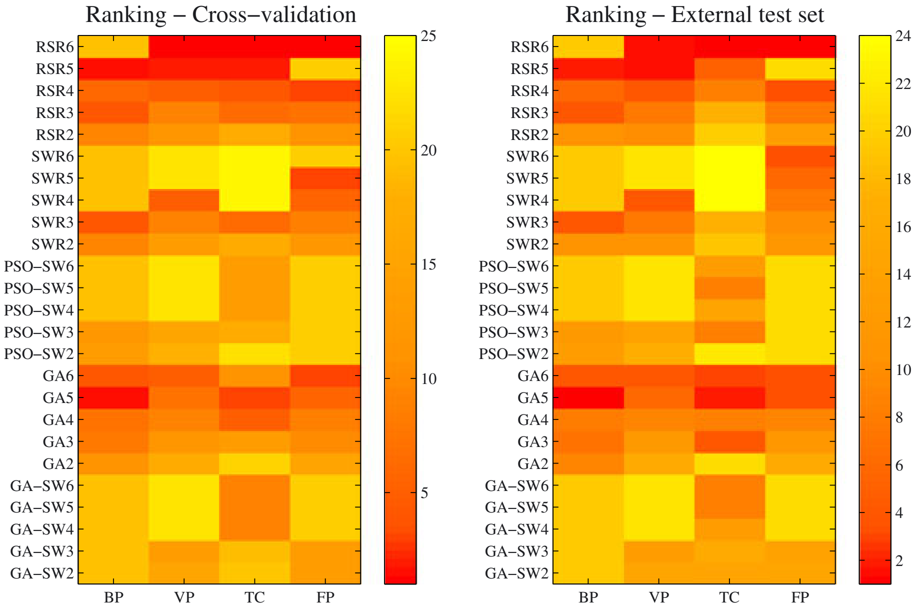

Figure 2. Heat map of the ranking of the best models for each method on training (Q^2_{cv}) and test (Q^2_{ext}) sets. The darker the colour, the higher the ranking position.

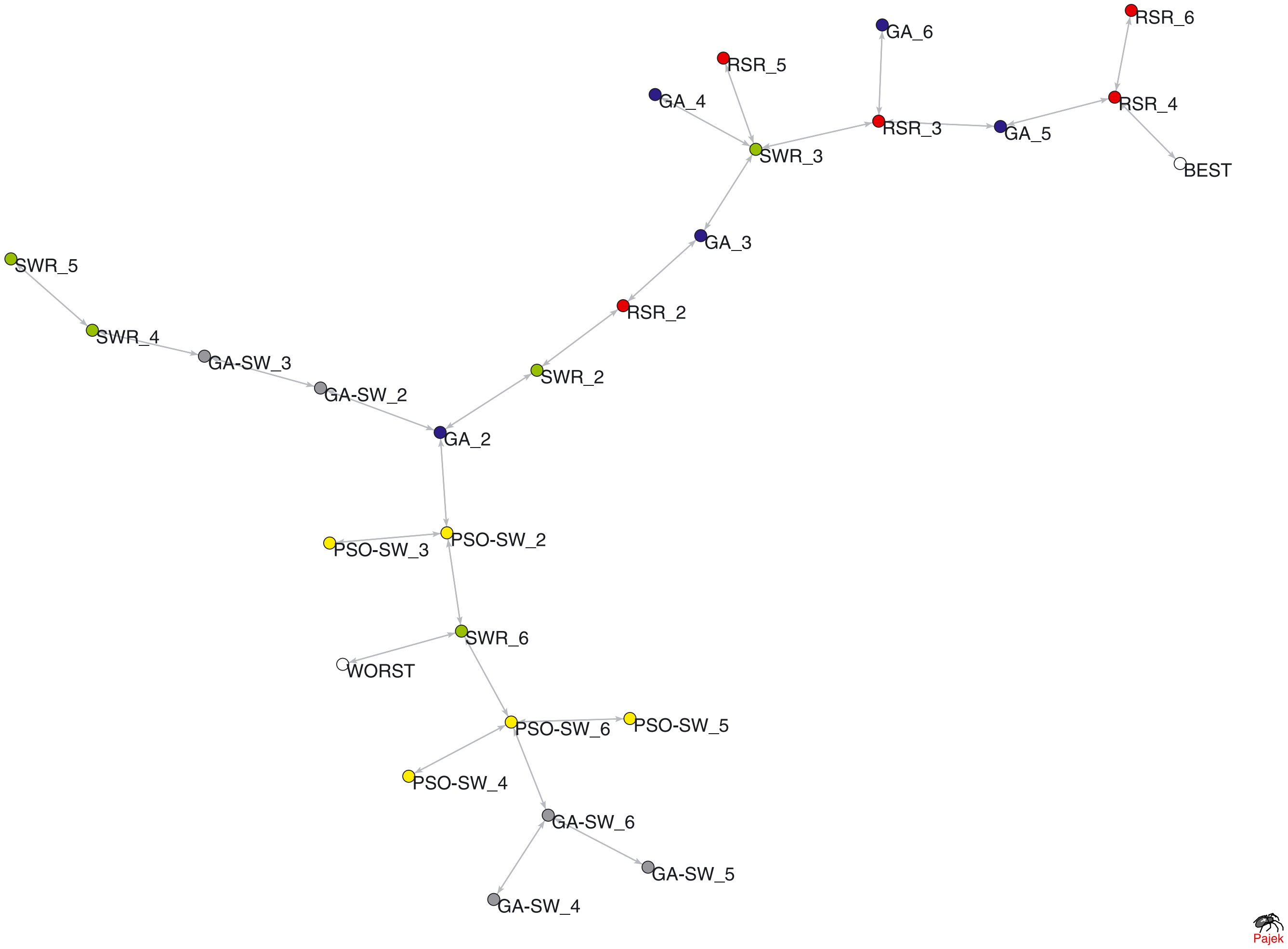

Figure 3. Minimum spanning tree of the ranking of the best model on the basis of $\mathbf{Q^2_{cv}}$. Labels correspond to the method and number of variables.

chromatographic datasets and not thought for OLS regression applied to molecular descriptors.

In order to further compare the results of the methods, the matrices were then range scaled between 1 and 0, and minimum spanning trees (MST) were built (Figures 3 and 4). Two dummy models were also added, BEST (always first position in the ranking) and WORST (always last position), in order to visually locate the optimal and non-optimal regions.

In Figure 3 (MST on the Q^2_{cv}-based ranking), a clear separation occurs between RSR and GA models that lay in the proximity of BEST, in respect to those of GA-SW and PSO-SW, which are close to WORST; SWR models show intermediate behaviour. RSR models with six, four, three and two variables are closer to BEST than those of the same size of the other methods. Finally, PSO-SW and GA-SW lay in the region of WORST because of both their poor performance in cross-validation and their high rejection rate by the *R* rules.

By observing the MST made on the Q^2_{ext} ranking (Figure 4), the situation appears similar. Still a clear separation between GA/RSR and other methods can be seen. In this case, however, GA models are closer to BEST with respect to RSR models. This behaviour is connected to the tendency of GA to give models with lower Q^2_{cv} but slightly larger Q^2_{ext} than RSR for the same size, even if in three out of four cases RSR provided the model with the largest Q^2_{ext}, regardless of model size. As already noticed for the Q^2_{cv}-based MST, PSO-SW and GA-SW are in the region of WORST for the poor performance in prediction with respect to the other VS methods and for the large number of models rejected by the *R* rules. Finally, the simplest model in the proximity of BEST is RSR-4.

The same information can be obtained by ranking the models according to the sum of ranking differences [38], using the maximum as reference value. RSR and GA models always occupy high positions in the ranking for all the datasets, while all the methods based on the final stepwise approach (i.e. GA-SW and PSO-SW) occupy the lowest positions in the ranking; SWR shows an intermediate behaviour, depending on the dataset.

In general, RSR appears to perform similar to SWR for what concerns low-dimensional models: in most of the cases, in fact, the two methods find the same solutions with two and three variables. On the other hand, for higher dimensional models, results of RSR diverge from those of SWR and are more similar to those of GA. SWR is a widely used method but known to have several drawbacks [39,40], such as the bias related to the inclusion of one variable at a time without considering other subsets of variables [41] and the tendency to model noise [39]. These problems increase in their relevance with the increase of the number of variables included. In this perspective, the divergence of RSR from SWR when the model dimension increases can be seen as representative of the ability of RSR to extensively explore the possible subsets and combinations of variables. Moreover, the activation of TL allows to temporarily exclude variables not correlated with the response, thus preventing overfitting and noise modelling, which are often related to the extensive exploration of the combinations

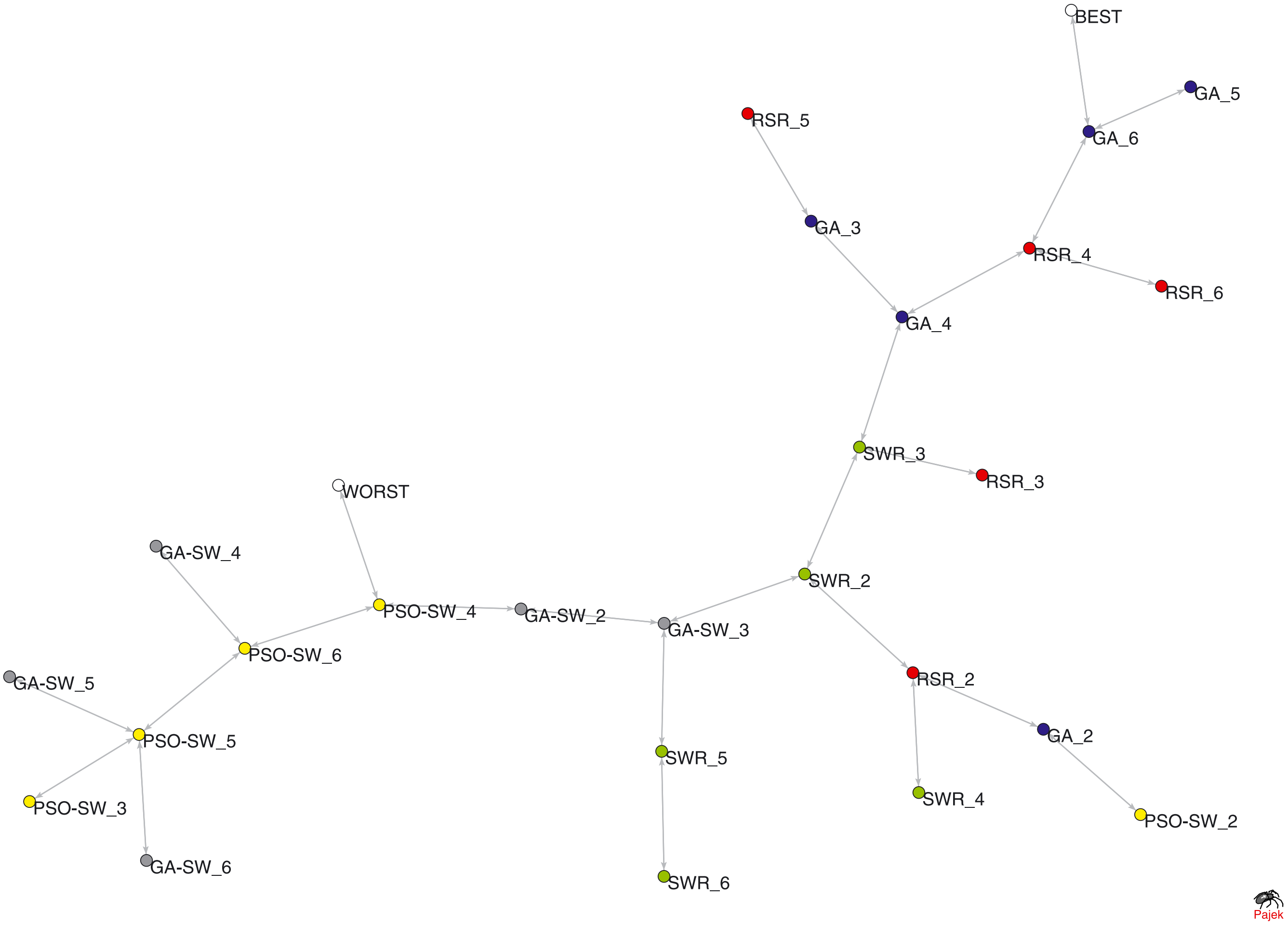

Figure 4. Minimum spanning tree of the ranking of the best model on the basis of $\mathbf{Q^2_{ext}}$. Labels correspond to the method and number of variables.

of variables. The drawbacks of SWR are confirmed by R^N rule, which rejected 100%, 75% and 25% of its models with respectively six, five and four variables.

Finally, for what concerns computational time, the general ranking of the methods is as follows: GA-SW < RSR (about 1.6 times slower than GA-SW) < PSO-SW (about 2.2 times slower than GA-SW) < GA (about 7.1 times slower than GA-SW) < SWR (7.7 times slower than GA-SW). In other words, RSR, despite its extensive exploration of the variables space, has a computational time comparable with that of the other meta-heuristic methods, thanks to the addition of TL and RW functions.

6. CONCLUSIONS

In the present work, the RSR algorithm for VS was compared with other reference methods: Genetic Algorithms (GA and GA-SW), Particle Swarm Optimization (PSO-SW) and Stepwise Forward Selection (SWR).

The methods were applied to four QSPR datasets that differ in their objects/variables ratios and the physical-chemical property to be modelled.

In order to analyse the final populations of models by means of a common procedure, Y-scrambling test and *R*-function-based rules were applied, and only the models that fulfilled these tests were retained.

In three out of four cases, RSR algorithm found the best models, in terms of Q^2_{cv} and Q^2_{ext}, while in the other case, the best model was found by GA.

The GA-SW and PSO-SW often provided models that did not fulfil *R*-based rules, and the accepted models had, in most of the cases, a poor performance in cross-validation and external validation. Moreover, RSR found low-dimensional (less than four variables) models similar to those of SWR, while for higher dimensions, the performance of models found was more similar to that of genetic algorithms.

Computational time of RSR resulted to be comparable with that of GA-SW and PSO-SW and lower than that of GA and SWR.

REFERENCES

1. Todeschini R, Consonni V. *Molecular Descriptors for Chemoinformatics* (2nd Revised and Enlarged Edition). Wiley-VCH: Weinheim, Germany; 2009.
2. Myung IJ, Pitt MA. Applying Occam's razor in modeling cognition: a Bayesian approach. *Psychon. B. Rev.* 1997; **4**:79–95.
3. Yu L, Lai KK, Wang S, Huang W. A bias–variance–complexity trade-off framework for complex system modeling. In *Computational Science and Its Applications—ICCSA 2006*, Gavrilova M, Gervasi O, Kumar V, Tan CJK, Taniar D, Laganá A, *et al.* (eds.). Springer: Berlin Heidelberg, 2006; 518–27.
4. Gonzalez MP, Teran C, Saiz-Urra L, Teijeira M. Variable selection methods in QSAR: an overview. *Curr. Top. Med. Chem.* 2008; **8**: 1606–27.
5. Guyon I, Elisseeff A. An introduction to variable and feature selection. *J. Mach. Learn. Res.* 2003; **3**: 1157–82.
6. Efroymson M. Multiple regression analysis. *Math. Methods Digital Comput.* 1960; **1**: 191–203.
7. Kim K-J, Cho S-B. A comprehensive overview of the applications of artificial life. *Artif. Life* 2006; **12**(1): 153–82.
8. Leardi R *Nature-inspired Methods in Chemometrics: Genetic Algorithms and Artificial Neural Networks*. Elsevier: Amsterdam, The Netherlands; 2003.
9. Yang X-S. *Nature-Inspired Metaheuristic Algorithms: Second Edition*. Luniver Press: Frome, United Kingdom; 2010.
10. Goldberg DE, Holland JH. Genetic algorithms and machine learning. *Mach. Learn.* 1988; **3**(2-3): 95–9.
11. Shen Q, Jiang J-H, Jiao C-X, Shen G, Yu R-Q. Modified particle swarm optimization algorithm for variable selection in MLR and PLS modelling: QSAR studies of antagonism of angiotensin II antagonists. *Eur. J. Pharm. Sci.* 2004; **22**(2–3): 145–52.
12. Dorigo M, Di Caro G. Ant colony optimization: a new meta-heuristic. *Proceedings of the 1999 Congress on Evolutionary Computation* 1999. p. -1477 Vol. 2.
13. Fogel DB. *Evolutionary Computation: Toward a New Philosophy of Machine Intelligence*. John Wiley & Sons: New Jersey, USA; 2006.
14. Tibshirani R Regression shrinkage and selection via the LASSO. *J. R. Stat. Soc. Ser. B (Methodological)* 1996; **58**(1): 267–88.
15. Zou H, Hastie T. Regularization and variable selection via the elastic net. *J. R. Stat. Soc. Ser. B (Statistical Methodology)* 2005; **67**(2): 301–20.
16. Miller AJ. Selection of subsets of regression variables. *J. R. Stat. Soc. Ser. A (General)*. 1984; **147**(3): 89–425.
17. Cassotti M, Grisoni F, Todeschini R. Reshaped Sequential Replacement algorithm: an efficient approach to variable selection. *Chemom. Intell. Lab.* (in press). 2014. DOI: 10.1016/j.chemolab.2014.01.011
18. Bledsoe W. The use of biological concepts in the analytical study of systems. *Proceedings of the ORSA-TIMS National Meeting* 1961.
19. Holland JH. *Adaptation in Natural and Artificial Systems: An Introductory Analysis with Applications to Biology, Control, and Artificial Intelligence*. University of Michigan: USA, 1975. Oxford, England: U Michigan Press.
20. Eberhart R, Kennedy J. A new optimizer using particle swarm theory. *Proceedings of the Sixth International Symposium on Micro Machine and Human Science* 1995; p. 39–43.
21. Kennedy J, Eberhart RC. A discrete binary version of the particle swarm algorithm. *IEEE International Conference on Systems, Man, and Cybernetics*, 1997; p. 4104–4108 vol.5.
22. Todeschini R, Consonni V, Mauri A, Pavan, M. Detecting 'bad' regression models: multicriteria fitness functions in regression analysis. *Anal. Chim. Acta* 2004; **515**(1): 199–208.
23. Mason CH, Perreault WD. Collinearity, power, and interpretation of multiple regression analysis. *J. Marketing Res.* 1991; **28**(3): 268–80.
24. Stewart GW. Collinearity and least squares regression. *Stat. Sci.* 1987; **2**(1): 68–84.
25. Todeschini R, Consonni V, Maiocchi A. The K correlation index: theory development and its application in chemometrics. *Chemom. Intell. Lab. Syst.* 1999; **46**(1): 13–29.
26. Lindgren F, Hansen B, Karcher W, Sjöström M, Eriksson L. Model validation by permutation tests: applications to variable selection. *J. Chemom.* 1996; **10**(5-6): 521–32.
27. Todeschini R, Ballabio D, Consonni V, Manganaro A, Mauri A. Canonical measure of correlation (CMC) and canonical measure of distance (CMD) between sets of data. Part 1. Theory and simple chemometric applications. *Anal. Chim. Acta* 2009; **648**(1): 45–51.
28. Consonni V, Ballabio D, Todeschini R. Evaluation of model predictive ability by external validation techniques. *J. Chemom.* 2010; **24**(3-4): 194–201.
29. US EPA 2013. Toxicity Estimation Software Tool (T.E.S.T.)—http://www.epa.gov/nrmrl/std/qsar/qsar.html [9 September 2013]
30. Talete srl. Dragon (Software for Molecular Descriptor Calculation) Version 6.0—2012—http://www.talete.mi.it/. 2012.
31. MATLAB. R2012b. *Natick*. The MathWorks Inc.: Massachusetts, 2012.
32. MICHEM website. http://michem.disat.unimib.it/chm/.
33. Ertl P, Rohde B, Selzer P. Fast calculation of molecular polar surface area as a sum of fragment-based contributions and its application to the prediction of drug transport properties. *J. Med. Chem.*. 2000; **43**(20): 3714–7.
34. Randić M, Jurs PC. On a fragment approach to structure–activity correlations. *Quant. Struct.-Act. Relat.* 1989; **8**(1): 39–48.
35. Galvez J, Garcia R, Salabert MT, Soler R. Charge indexes. New topological descriptors. *J. Chem. Inf. Comput. Sci.* 1994; **34**(3): 520–5.
36. Ghose AK, Viswanadhan VN, Wendoloski JJ. Prediction of hydrophobic (lipophilic) properties of small organic molecules using fragmental methods: an analysis of ALOGP and CLOGP methods. *J. Phys. Chem. A* 1998; **102**(21): 3762–72.
37. Viswanadhan VN, Ghose AK, Revankar GR, Robins RK. Atomic physicochemical parameters for three dimensional structure directed quantitative structure–activity relationships. 4. Additional parameters for hydrophobic and dispersive interactions and their application for an automated superposition of certain naturally occurring nucleoside antibiotics. *J. Chem. Inf. Comput. Sci.* 1989; **29**(3): 163–72.

38. Héberger K, Kollár-Hunek K. Sum of ranking differences for method discrimination and its validation: comparison of ranks with random numbers. *J. Chemom.* 2011; **25**(4): 151–8.

39. Derksen S, Keselman HJ. Backward, forward and stepwise automated subset selection algorithms: frequency of obtaining authentic and noise variables. *Br. J. Math. Stat. Psychol.* 1992; **45**(2): 265–82.

40. Steyerberg EW, Eijkemans MJC, Habbema JDF. Stepwise selection in small data sets: a simulation study of bias in logistic regression analysis. *J. Clin. Epidemiol.* 1999; **52**(10): 935–42.

41. Burnham KP, Anderson DR. *Model Selection and Multi-model Inference: A Practical Information-theoretic Approach*. Springer: New York, USA; 2002.