Received: 21 March 2015, Revised: 3 August 2015, Accepted: 18 August 2015, Published online in Wiley Online Library: 23 September 2015

(wileyonlinelibrary.com) DOI: 10.1002/cem.2746

Incorporating PLS model information into particle swarm optimization for descriptor selection in QSAR/QSPR

Yong Wang[a,c], Jing-Jing Huang[a], Neng Zhou[d], Dong-Sheng Cao[b]*, Jie Dong[b] and Han-Xiong Li[c]

As a representative paradigm of evolutionary algorithms, particle swarm optimization (PSO) has been combined with partial least square (PLS) (called PSO-PLS) to select informative descriptors in quantitative structure-activity/property relationship (QSAR/QSPR). However, one of the main limitations of PSO-PLS is that it ignores PLS model information. In this paper, by incorporating the PLS model information into PSO-PLS, we present a novel weighted sampling method (called WS-PSO-PLS) to choose the optimal descriptor subset. Due to the fact that the regression coefficients of the PLS model reflect the importance of descriptors in the model development, we firstly obtain the normalized regression coefficients by establishing the PLS model with all the descriptors. Afterward, weighted sampling is used to generate some individuals according to the aforementioned normalized regression coefficients. Finally, we employ some dimensions of the generated individuals to replace the corresponding dimensions of the individuals with poor quality in the population at each generation. WS-PSO-PLS has been assessed through three QSAR/QSPR datasets and the experimental results suggest that WS-PSO-PLS has the capability to effectively guide the search process by introducing the PLS model coefficients into PSO during the evolution and, therefore, performs better than PSO-PLS. WS-PSO-PLS could be considered as a general and promising mechanism to introduce extra information to improve the performance of PSO for descriptor selection in QSAR/QSPR. Copyright © 2015 John Wiley & Sons, Ltd.

Additional supporting information may be found in the online version of this article at the publisher's web site.

Keywords: descriptor selection; particle swarm optimization; partial least square); regression coefficients; weighted sampling

1. INTRODUCTION

Quantitative structure-activity/property relationship (QSAR/QSPR), an important area in the chemical and biomedical sciences, searches the relationship between compounds and corresponding biological activities or chemical properties [1]. In order to obtain this relationship, a variety of statistical learning methods have been proposed in QSAR/QSPR, including multiple linear regression (MLR), principal component regression, partial least square (PLS) regression [2–4], decision tree [5], support vector machines [6,7], random forest [8,9], boosting [10,11], and so on. In QSAR/QSPR studies, the chemical structure of compounds is represented by several descriptors, such as molecular constitutional, topological, shape, autocorrelation, and charge descriptors. In general, the number of descriptors is relatively larger than the number of compounds. Some redundant, noisy, and irrelevant descriptors have a side effect on the QSAR/QSPR model development. Meanwhile, too many descriptors may result in either over fitting or a low correlation between structures and activities [12]. Therefore, it is necessary to perform descriptor selection before the QSAR/QSPR model development. Actually, descriptor selection in QSAR/QSPR has the following advantages [13]: (1) increasing the prediction accuracy of the model; (2) facilitating the interpretability of relationship between descriptors and activities; (3) balancing the effective number of degrees of freedom for calculating reliable estimates of the model's parameters; and (4) decreasing the time complexity of model development.

As the selection of informative descriptors has become one of the key steps for QSAR/QSPR model development, several descriptor selection methods have been presented [14,15], including the correlation-based method[16], information theory-based method [17], statistical criteria-based method [18], competitive adaptive reweighted sampling (CARS) [19], Monte Carlo tree [20], recursive feature elimination [21], ordered predictors selection [22], uninformative variable elimination [23], artificial intelligence-based methods [24,25], and so on. These methods

* *Correspondence to: Dong-Sheng, Cao, School of Pharmaceutical Sciences, Central South University, Changsha 410013, China.*
E-mail: oriental-cds@163.com

a *Y. Wang, J.-J. Huang*
School of Information Science and Engineering, Central South University, Changsha 410083, China

b *D.-S. Cao, J. Dong*
School of Pharmaceutical Sciences, Central South University, Changsha 410013, China

c *Y. Wang, H.-X. Li*
Department of Systems Engineering and Engineering Management, City University of Hong Kong, Hong Kong

d *N. Zhou*
School of chemistry and materials, Yulin Normal University, Yulin, Guangxi, 537000, China

can be briefly divided into two categories [26]: filtering methods and wrapper methods. The filtering methods assess the relevance of descriptors by only using the intrinsic properties of the data. In most cases, a descriptor relevance score is calculated, and low scoring descriptors are removed. Afterward, the resulting subset of descriptors is presented as input of the modeling algorithm. The disadvantages of the filtering approaches are threefold: (1) they ignore the interaction with the algorithms, that is, the search in the descriptor subset space is separated from the search in the hypothesis space; (2) most proposed methods are univariate; and (3) in these approaches, each descriptor is considered separately, thereby the descriptor dependencies have been ignored [27]. On the other hand, the wrapper methods consist of two components: the objective function and the optimization algorithm. The latter is used to select the optimal descriptor subset for the former. The advantages of the wrapper approaches include the interaction between descriptor subset search and model selection and the ability to take into account descriptor dependencies. The current popular optimization algorithms include simulated annealing [28], genetic algorithm [29–31], and other evolutionary algorithms (EAs) [32,33]. Among them, EAs have become more and more popular to deal with the large space of descriptor subsets. Note, however, that most EAs do not exploit the model information to guide the evolution. As a result, it is still an open issue to incorporate the model information or prior information into EAs to search for the optimal descriptor subset quickly.

In the present study, we attempt to address the aforementioned issue by considering the interact information between the model and EAs. Particle swarm optimization (PSO), one of the most representative paradigms of EAs, has been widely applied to select descriptors in QSAR/QSPR studies [34–36]. Taking PSO-PLS [37] as an example, we propose to use the normalized coefficients, obtained by developing the PLS model with all the descriptors, to guide the search process. Afterward, some individuals are produced via weighted sampling, which is based on the obtained normalized coefficients. Subsequently, these individuals are stored into the population through replacing some dimensions of the inferior individuals at each generation. By the aforementioned process, the model information has been effectively utilized to guide the search process and a weighted sampling PSO-PLS (called WS-PSO-PLS) has been developed to select the optimal descriptor subset in QSAR/QSPR model development. Furthermore, we propose a mutation strategy in WS-PSO-PLS to prevent the population from getting trapped into a local optimum. We have carefully analyzed two significant features of WS-PSO-PLS. The experimental results show that WS-PSO-PLS exhibits better performance than the original version and the other improved versions of PSO-PLS on three QSAR/QSPR datasets. To the authors' best knowledge, this paper is the first attempt to utilize the PLS model information to bias the evolution of PSO.

The rest of this paper is organized as follows. Section 2 gives a detailed description of the proposed WS-PSO-PLS. In Section 3, we briefly introduce three datasets, which are used to validate the effectiveness of WS-PSO-PLS, and the platform for carrying out the experiments. In Section 4, the experimental results of WS-PSO-PLS have been compared with those of PLS, PSO-PLS, and the other improved versions of PSO-PLS on the chosen datasets by taking advantage of three performance metrics. Section 5 discusses the experimental results. Finally, Section 6 provides some concluding remarks.

2. THEORY AND METHODS

2.1. Modified particle swarm optimization

Particle swarm optimization is a population-based stochastic optimization technique proposed by Eberhart and Kenndy in 1995 [38,39]. PSO simulates the social behavior of organisms, such as bird flocking and fish schooling. In PSO, every particle in the swarm is a potential solution to an optimization problem. All particles "fly" through a *D*-dimension search space by learning their own experiences and the experiences of the entire swarm. PSO is initialized with a group of random particles, and each particle (also called an individual) has a velocity, a position, and a corresponding fitness evaluated by the fitness function. The velocity and the position of the *i*th particle are represented as $v_i = (v_{i,1}, v_{i,2}, \ldots, v_{i,D})$ and $\overrightarrow{x_i} = (x_{i,1}, x_{i,2}, \ldots, x_{i,D})$, respectively. In addition, the best previous position of the *i*th particle is called the personal best and represented as $\overrightarrow{p_i} = (p_{i,1}, p_{i,2}, \ldots, p_{i,D})$, and the best previous position of all the particles in the swarm is called the global best and represented as $\overrightarrow{p_g} = (p_{g,1}, p_{g,2}, \ldots, p_{g,D})$. At each generation, the velocity of each particle is updated by making use of $\overrightarrow{p_i}$ and $\overrightarrow{p_g}$. Afterward, it is necessary to update the position of each particle.

For a discrete optimization problem expressed in a binary notation, a particle moves in the search space, each dimension of which is restricted to "0" or "1." Under this condition, the *j*th dimension of the position of the *i*th particle (i.e., $x_{i,j}$) should be in either state "1" or state "0," and the corresponding velocity (i.e., $v_{i,j}$) represents the probability of $x_{i,j}$ being equal to "1." For descriptor selection in QSAR/QSPR, if there are *D* descriptors in the model development, then an individual in PSO will have *D* bits (i.e., *D* dimensions) correspondingly. If a bit of an individual is equal to "1," then the corresponding descriptor will be selected, otherwise the corresponding descriptor will not be selected. In our study, the discrete PSO developed by Yu [37], which is proposed to select descriptors in MLR and PLS modeling for QSAR/QSPR, has been adopted. As in [37], the *j*th dimension of the velocity of the *i*th particle (i.e., $v_{i,j}$) is a random number between 0 and 1, and $x_{i,j}$ is updated by the following rules:

$$\text{if } \left(0 < v_{i,j} \le a\right), \text{ then } x_{i,j}^{G+1} = x_{i,j}^{G} \tag{1}$$

$$\text{if } \left(a < v_{i,j} \le \frac{1+a}{2}\right), \text{ then } x_{i,j}^{G+1} = p_{i,j}^{G} \tag{2}$$

$$\text{if } \left(\frac{1+a}{2} < v_{i,j} \le 1\right), \text{ then } x_{i,j}^{G+1} = p_{g,j}^{G} \tag{3}$$

where *a* is a constant between 0 and 1, and *G* denotes the generation number.

In the proposed WS-PSO-PLS, we incorporate the PLS model information into the discrete PSO in [37] to guide the search of the optimal descriptor subset. Firstly, we build the PLS model with all the descriptors and normalize the coefficient corresponding to each descriptor. Note that the importance of a descriptor can be determined by its normalized coefficient in the PLS model, that is, the larger the normalized coefficient, the more important the descriptor in model development [40]. Then, *subsize* individuals (denoted as set $A, A = \left\{\overrightarrow{s_i},\ i = 1, \ldots, subsize\right\}$) are produced via weighted sampling with the normalized coefficients. During the weighted sampling, the value of the normalized coefficient is regarded as the probability of a bit of an individual being "1," and as a result, the more important a descriptor, the higher probability it will be selected. Subsequently, we sort the population according to the fitness in ascending order and select *subsize* individuals (denoted as set *B*, $B = \{\overrightarrow{x_i},\ i = 1, \ldots, subsize\}$) with the worst quality at each generation. Next, each dimension of an individual $\overrightarrow{s_i}$ in set *A* is utilized to replace the corresponding dimension of an individual $\overrightarrow{x_i}$ in set *B* by the following equation:

$$\text{if } rand > b, \text{ then } x_{i,j}^{G+1} = s_{i,j}^{G+1},\ i = 1, \ldots, subsize,\ j = 1, \ldots, D \tag{4}$$

where *rand* stands for a uniformly distributed random number between (0, 1) and *b* is a constant between 0 and 1. By Equation 4, the quality of the individuals in set *B* could be improved by exploiting the information of some important descriptors. Moreover, we can control how much the model information will be incorporated into the inferior individuals in the population through tuning the value of *b*. In this paper, *subsize* is set to $\alpha * popsize$, where α is a constant in (0,1) and *popsize* is the number of individuals in the population. It is necessary to point out that in the initial population, *subsize* individuals are generated based on the weighted sampling, and the remaining individuals are generated randomly.

Next, we use an example to illustrate the weighted sampling proposed in this paper. Suppose that there are four descriptors in QSAR/QSPR model development, which implies that an individual in PSO includes four bits (i.e., four dimensions) correspondingly. Suppose also that the vector of the normalized coefficients, which are obtained by building the PLS model with all the four descriptors, is equal to $p = [0.145, 0.8, 0.05, \text{and } 0.005]$. Under this condition, the probability that the first bit of an individual is equal to "1" and the first descriptor is selected, is 0.145, the probability that the second bit of the individual is equal to "1" and the second descriptor is selected, is 0.8, and so on. Compared with the random sampling, in which each descriptor is selected with the same probability (i.e., 0.25), in the weighted sampling, the important descriptors, which have relatively larger normalized coefficients, are more likely to be selected. As a result, high quality individuals can be generated by the weighted sampling, which is beneficial to improve the overall quality of the whole population. We just incorporate such information into each generation of PSO to improve the quality of individuals, thus guiding the evolution process.

According to the aforementioned introduction, it is clear that the PLS model coefficients can be utilized to improve the quality of the population in PSO. Consequently, WS-PLS-PSO has the capability to find the optimal descriptor subset promptly.

In this paper, a mutation strategy has been proposed, the purpose of which is to prevent the population from getting trapped into a local optimum. The mutation strategy is implemented as follows. Firstly, we choose $\lceil c * popsize \rceil$ individuals from the population randomly, where *c* is the mutation probability between 0 and 1. And then, a randomly selected bit of each individual is flipped from "1" to "0" or from "0" to "1."

2.2. The fitness function

In order to evaluate the performance of each individual, the predictive Q^2 value [41] is used as the fitness function, which is defined as follows:

$$Q^2 = 1 - \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} (y_i - \bar{y}_i)^2} \tag{5}$$

where y_i is the observed value of activity, *i,n* is total number of compounds, $\bar{y}_i$ is the average of y_i, and $\hat{y}_i$ is the value predicted by the PLS model via using the fivefold cross-validation procedure. In the fivefold cross-validation procedure, the original whole dataset is randomly partitioned into five subsets with equal size. Afterward, four of these five subsets are used as training data and the remaining one is used as the validation data for testing the model. The aforementioned procedure repeats five times (i.e., fivefolds), then all the predicted results $\hat{y}_i (i = 1, \ldots, n)$ can be obtained.

2.3. Weighted sampling particle swarm optimization with partial least square

Figure 1 shows the flowchart of WS-PSO-PLS. The detailed steps are summarized as follows:

Step 1. Define the parameters (all the parameters are listed in Table I), set the generation number $G = 1$, obtain the coefficients by establishing the PLS model with all the descriptors, and normalize the aforementioned coefficients.

Step 2. Initialize the population (i.e., $\overrightarrow{x_i},\ i = 1, \ldots, popsize$): $\alpha * popsize$ individuals are generated based on weighted sampling (sampling part), and the remaining individuals are generated randomly (random part). Next, initialize $\overrightarrow{p_i}$: $\overrightarrow{p_i} = \overrightarrow{x_i},\ i = 1, \ldots, popsize$.

Step 3. Evaluate the population: for each individual $\overrightarrow{x_i}$, we choose the bits, one of which is equal to "1." Then, the corresponding descriptors constitute a subset. Subsequently, the PLS regression is applied to the subset to calculate the fitness Q^2. Clearly, the larger the value of Q^2, the better the individual. Afterward, initialize $\overrightarrow{p_g}$: $\overrightarrow{p_g}$ is equal to the individual with the maximum Q^2.

Step 4. Update each individual according to Equations 1–3, perform the mutation strategy, and evaluate the population. Then, find *subsize* individuals with the worst quality in the population, update these individuals according to Equation 4, and evaluate these *subsize* updated individuals.

Step 5. Update $\overrightarrow{p_i}$ and $\overrightarrow{p_g}$ according to the following rules:

$$\mathit{If}\ Q^2\left(\overrightarrow{x_i}^{G+1}\right) > Q^2\left(\overrightarrow{p_i}^{G}\right)\ \mathit{then}\ \overrightarrow{p_i}^{G+1} = \overrightarrow{x_i}^{G+1}; \tag{6}$$

$$\mathrm{If}\ Q^2\left(\overrightarrow{x_i}^{G+1}\right) > Q^2\left(\overrightarrow{p_g}^{G}\right)\ \mathit{then}\ \overrightarrow{p_g}^{G+1} = \overrightarrow{x_i}^{G+1}; \tag{7}$$

Step 6. If the stopping criterion is satisfied, then stop and output $\overrightarrow{p_g}$, otherwise let $G = G + 1$ and go to Step 4.

3. DATASETS AND SOFTWARE

3.1. Quantitative structure-activity/property relationship datasets

In this study, three QSAR/QSPR datasets are used for demonstrating the effectiveness of our method. The first is artemisinin dataset, which consists of 211 artemisinin analogues [42]. Due to the fact that this dataset has many enantiomeric pairs of activities, the element in each enantiomeric pair with smaller logarithm of the relative activity is used as the output variable (referred as log RA), and the other element is removed [43]. Therefore, the artemisinin dataset used in this paper has 178 compounds. As pointed out in [43], several structural diverse compounds have the same log RA (i.e., −4.0), which makes the

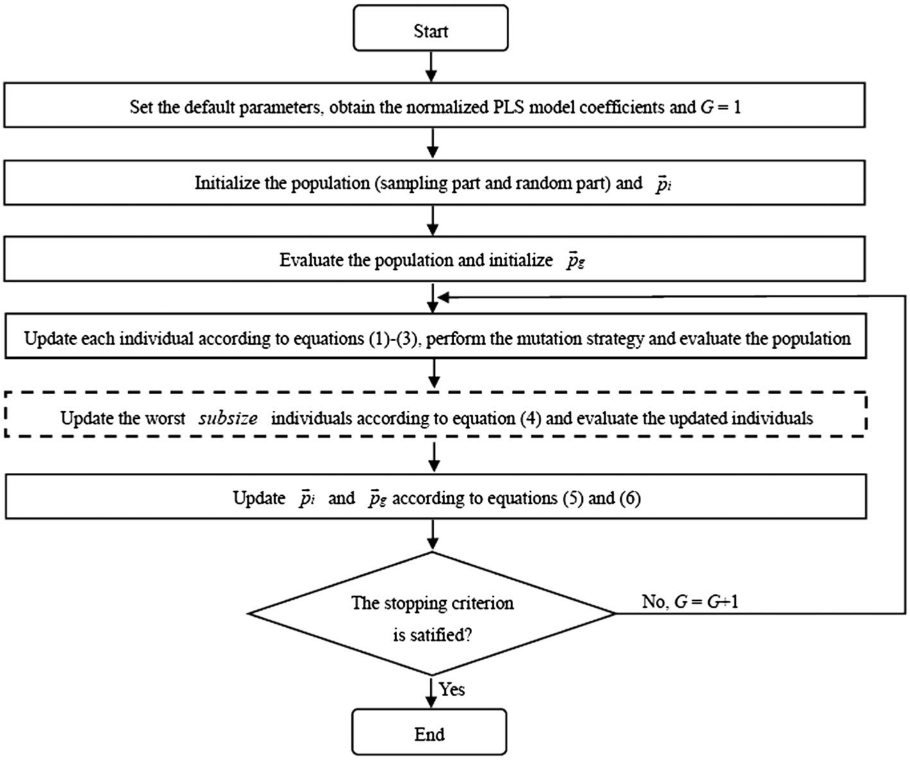
Start
Set the default parameters, obtain the normalized PLS model coefficients and $G = 1$
Initialize the population (sampling part and random part) and $\bar{p}_i$
Evaluate the population and initialize $\bar{p}_g$
Update each individual according to equations (1)-(3), perform the mutation strategy and evaluate the population
Update the worst *subsize* individuals according to equation (4) and evaluate the updated individuals
Update $\bar{p}_i$ and $\bar{p}_g$ according to equations (5) and (6)
The stopping criterion is satified?
No, $G = G$+1
Yes
End

Figure 1. The flowchart of weighted sampling particle swarm optimization with partial least square.

Table I. The parameter values used in different methods

	Parameters					
Methods	Population size: 50	Number of generations: 200	Learning rate *a*: 0.5	Mutation probability *c*: 0.05	Proportion of weighted sampling: $\alpha = 0.5$	Learning rate *b*: 0.8
PSO-PLS	√	√	√		—	
PSO-PLS-1	√	√	√	√	—	
WS-PSO-PLS-1	√	√	√	√	√	
WS-PSO-PLS-2	√	√	√	√	√	√
WS-PSO-PLS	√	√	√	√	√	√

PSO-PLS, particle swarm optimization with partial least square; WS-PSO-PLS, weighted sampling PSO-PLS.
√The parameter have been used in corresponding method.

model development more difficult. For each compound, two-dimensional (2D) descriptors are calculated using ChemoPy software package [44], which is developed by our group. Note that before further descriptor selection by WS-PSO-PLS, two descriptor preselection steps are performed to eliminate some uninformative descriptors: (1) remove the descriptors, the variance of which is near zero or zero and (2) if the correlation of two descriptors is larger than 0.95, then remove one of them. Finally, 89 molecular descriptors are obtained for representing compounds in the artemisinin dataset, and these molecular descriptors are used as inputs for QSAR/QAPR model development. These molecular descriptors include 18 constitutional descriptors, 32 topological structural descriptors, 27 electrotopological state (E-state) descriptors, 5 molecular property descriptors, 4 kappa descriptors, and 3 connectivity descriptors.

The second is benzodiazepine receptors (BZR) dataset. In the BZR dataset, benzodiazepines are a class of psychoactive drugs, which are used to treat anxiety, insomnia, and a range of other circumstances conditions. At the same time, benzodiazepines exhibit sedative, hypnotic, anti-anxiety, anticonvulsant, and muscle relaxant properties, and act via the BZR, which have been extensively researched in QSAR/QSPR [45–47]. The BZR dataset used in our study is presented in [48]. It contains 163 compounds and 75 2.5D descriptors consisting of S_sCH3, S_dssC, CHI-0, and so on.

The third is selwood dataset [49], which has become a benchmark to evaluate the performance of different methods and has been well-studied in QSAR/QSPR [50]. It consists of 29 compounds, 53 descriptors, and a set of corresponding antifilarial antimycin activities expressed as -log(IC50). The molecular descriptors in the selwood dataset include partial atomic charges for atoms 1–10 (ATCH1-ATCH10), dipole vector (DIPV_X, DIPV_Y, and DIPV_Z), dipole moment (DIPMOM), and so on.

The previous three datasets, together with calculated molecular descriptors, could be found in the Supporting Information.

3.2. Software

All the computations are performed with an in-house code in MATLAB (Version 2010a, The MathWorks, Inc., Natick, MA, USA) on a general-purpose computer with Inter® Core® i3 2.4GHz CPU and 2GB of RAM. The MATLAB source code can be obtained from the authors upon request.

4. EXPERIMENTAL RESULTS AND DISCUSSION

The proposed WS-PSO-PLS includes two main features: (1) the mutation strategy and (2) introducing the PLS model coefficients into PSO, including the first generation and the subsequent evolution process. In order to demonstrate the effectiveness of WS-PSO-PLS, five different methods are employed for comparison on the three datasets. These methods include the following: (1) PLS; (2) PSO-PLS[36]; (3) PSO-PLS-1, which is formed by combining the original PSO-PLS with our proposed mutation strategy; (4) WS-PSO-PLS-1, the difference between it and PSO-PLS-1 is that WS-PSO-PLS-1 only incorporates the PLS model information into the population at the first generation; and (5) WS-PSO-PLS-2, the difference between it and PSO-PLS-1 is that WS-PSO-PLS-2 incorporates the PLS model information into the population throughout the search process except the first generation.

By comparing the performance of the involved methods, three performance metrics have been chosen: Q^2, the root mean square error from fivefold cross-validation (denoted as *RMSECV*) and the number of the selected descriptors (denoted as *NSD*). Here, the reason that we choose fivefold cross validation is as follows: Firstly, the fivefold cross-validation and the standard leave-one-out cross-validation are similar owing to that they both belong to the class of k-fold cross-validation. Secondly, in our previous trial and error experiments, we found that fivefold cross-validation has higher computational efficiency than that of the standard leave-one-out cross-validation. Thirdly, maybe you think that the fivefold cross-validation is not as stable as the standard leave-one-out cross-validation. Note that all the involved methods have been implemented 100 times to obtain the statistical results. Thus, the experimental results are credible and steady in our paper. All the data were firstly auto-scaled to have zero mean and unit variance before modeling. Note that in PLS, all the descriptors were directly used for the model development. The maximum number of latent components was set to 20 and the optimal number of latent components was determined by the fivefold cross-validation. Owing to the randomness of PSO, the results may be different in different experiments. Thus, all the methods were implemented 100 times to obtain the statistical results. Moreover, Wilcoxon's rank sum test at a 0.05 significance level was used to check the statistical significance between two methods, in which the Wilcoxon rank sum test is a nonparametric approach to establishing significant difference between two sample groups using magnitude-based ranks [51]. The parameter settings for all the involved methods have been listed in Table I and the experimental results have been presented in Table II.

The first observation from Table II is that when using all the descriptors in PLS, the mean Q^2 is 0.60, 0.40, and 0.24 and the mean *RMSECV* is 0.99, 0.85, and 0.65 for the artemisinin, BZR, and selwood datasets, respectively. In contrast, the average number of the selected descriptors in PSO-PLS is drastically decreased. However, under this condition, the mean Q^2 is 0.7476, 0.5396, and 0.8685 and the mean *RMSECV* is 0.7875, 0.7451, and 0.2668 for the artemisinin, BZR, and selwood datasets, respectively. The aforementioned results suggest that PSO-PLS with less number of descriptors is significantly better than PLS, which verifies the necessity to perform descriptor selection before the QSAR/QSPR model development.

From Table II, it can be seen that PSO-PLS-1 performs better than PSO-PLS in terms of all the performance metrics on the three datasets. For example, with respect to the artemisinin dataset, the mean Q^2 is 0.7545 versus 0.7476, the mean *RMSECV* is 0.7716 versus 0.7875, and the mean *NSD* is 35.05 versus 39.04. As pointed out previously, PSO-PLS-1 is a combination of PSO-PLS with the mutation strategy. Therefore, the mutation strategy can be adopted to enhance the performance of PSO-PLS based on the experimental results.

The PLS model information has been incorporated into WS-PSO-PLS-1 only at the first generation, and PSO-PLS-1 does not use such model information. Compared with PSO-PLS-1, the mean Q^2 of WS-PSO-PLS-1 increase by 0.74%, 1.75%, and 0.77%, the mean *RMSECV* of WS-PSO-PLS-1 decrease by 0.39%, 0.46%, and 2.86%, and the mean *NSD* of WS-PSO-PLS-1 decrease by 1.71%, 5.90%, and 4.54% for the artemisinin, BZR, and selwood datasets, respectively, as shown in Table II. Thus, we can conclude that the incorporation of the PLS model information at the first generation does improve the performance of PSO-PLS-1. In order to further compare WS-PSO-PLS-1 with PSO-PLS-1, Figure 2(A) shows the mean Q^2 of all individuals at the first generation for these two methods on the artemisinin dataset (see Figure S1(A) and Figure S2(A) for the BZR and selwood datasets, respectively). From Figure 2(A), WS-PSO-PLS-1 outperforms PSO-PLS-1 for the first half of the population, while the performance of WS-PSO-PLS-1 and PSO-PLS-1 is nearly the same for the remaining half of the population at the first generation. Note that the first 50% individuals in the population of WS-PSO-PLS-1 are generated by weighted sampling while the remaining 50% individuals are generated randomly. The aforementioned phenomenon verifies that the model information can be exploited to enhance the quality of the population even in the initialization. Moreover, the superiority of WS-PSO-PLS-1 can also be demonstrated by the initial Q^2 in each run for the artemisinin dataset (Figure 2(B)), as well as for the BZR dataset (Figure S1(B)) and the selwood dataset (Figure S2(B)). It is not difficult to understand because the initial individuals are of higher quality by weighted sampling, which has also been verified by our previous study in [30].

As shown in Table II, WS-PSO-PLS-2 is better than PSO-PLS-1 in terms of all the performance metrics on the three datasets. As mentioned previously, the difference between them is that WS-PSO-PLS-2 incorporates the PLS model information into the population after the first generation. As a result, we can conclude that WS-PSO-PLS-2 benefits from the PLS model information during the evolution.

Figure 3, Figure S3, and Figure S4 provide the boxplots of *RMSECV* for the involved methods on the artemisinin, BZR, and selwood datasets, respectively. As shown in these figures, WS-PSO-PLS achieves the best overall performance among the involved methods, which further validates that the mutation strategy and the introduction of the PLS model coefficients into the whole evolution process are both effective for the performance improvement of PSO-PLS.

Table II. Experimental results on the three datasets

Dataset	Methods	Mean Q^2 ± Standard deviation	Mean *RMSECV* ± Standard Deviation	Mean *NSD* ± Standard deviation	Mean *NLV* ± Standard deviation
Artemisinin	PLS	0.60	0.99	89	15
	PSO-PLS[a]	0.7476 ± 0.0093	0.7875 ± 0.0145	39.04 ± 4.40	13 ± 2.25
	PSO-PLS-1	0.7545 ± 0.0107	0.7716 ± 0.0168	35.05 ± 4.19	12 ± 2.65
	WS-PSO-PLS-1	0.7601 ± 0.0090	0.7686 ± 0.0130	34.45 ± 4.17	12 ± 1.57
	WS-PSO-PLS-2	0.7622 ± 0.0075	0.7646 ± 0.0119	34.34 ± 3.98	12 ± 1.72
	WS-PSO-PLS[b]	0.7744 ± 0.0067	0.7447 ± 0.0110	32.24 ± 3.55	10 ± 0.34
BZR	PLS	0.40	0.85	75	8
	PSO-PLS[a]	0.5396 ± 0.0092	0.7451 ± 0.0075	32.35 ± 4.02	7 ± 1.86
	PSO-PLS-1	0.5489 ± 0.0131	0.7330 ± 0.0099	29.14 ± 3.68	7 ± 1.94
	WS-PSO-PLS-1	0.5585 ± 0.0096	0.7296 ± 0.0079	27.42 ± 3.14	7 ± 0.72
	WS-PSO-PLS-2	0.5588 ± 0.0095	0.7289 ± 0.0084	28.34 ± 2.93	6 ± 1.67
	WS-PSO-PLS[b]	0.5632 ± 0.0073	0.7258 ± 0.0061	26.96 ± 2.74	5 ± 0.41
Selwood	PLS	0.24	0.65	53	6
	PSO-PLS[a]	0.8685 ± 0.0337	0.2668 ± 0.0331	20.43 ± 2.79	5 ± 1.78
	PSO-PLS-1	0.8917 ± 0.0334	0.2416 ± 0.0340	18.74 ± 3.07	5 ± 1.65
	WS-PSO-PLS-1	0.8986 ± 0.0228	0.2347 ± 0.0259	17.89 ± 3.01	5 ± 1.41
	WS-PSO-PLS-2	0.9153 ± 0.0122	0.2125 ± 0.0152	17.34 ± 2.34	5 ± 1.89
	WS-PSO-PLS[b]	0.9200 ± 0.0130	0.2090 ± 0.0170	16.44 ± 2.79	4 ± 0.36

PLS, partial least square; PSO-PLS, particle swarm optimization with partial least square; WS-PSO-PLS, weighted sampling PSO-PLS. ***NLV*** is the number of latent variables in each PLS model. The results of Wilcoxon's rank sum test at a 0.05 significance level between *a* and *b* in terms of mean Q^2 are: $p = 4.46 \times 10^{-33}$, $p = 1.11 \times 10^{-30}$, and $p = 9.90 \times 10^{-29}$ on artemisinin, benzodiazepine receptors (BZR), and selwood datasets, respectively.

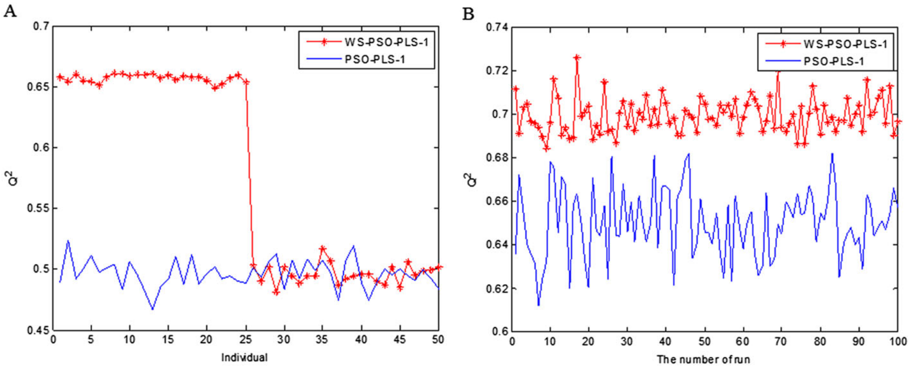

Figure 2. (A) The mean Q^2 of the population at the first generation for the artemisinin dataset. WS-PSO-PLS: the red "*-", and PSO-PLS: the blue "-". (B) The initial Q^2 of PSO-PLS and WS-PSO-PLS for the artemisinin dataset. WS-PSO-PLS: the red "*-", and PSO-PLS: the blue "-". WS-PSO-PLS, weighted sampling PSO-PLS ; PSO-PLS, particle swarm optimization with partial least square.

Next, we will verify the effectiveness of WS-PSO-PLS from another point of view. At the end of the evolution, by making use of all the descriptors in the artemisinin dataset, we plot the bar chart of the normalized coefficients of the PLS model in Figure 4 (A). Figure 4(A) shows that some descriptors have relatively larger normalized coefficients, such as *nring* and *naccr*, the normalized coefficients of which are 0.0892 and 0.0446, respectively. However, some descriptors have relatively smaller normalized coefficients. As pointed out previously, the value of the normalized coefficient represents the importance of a descriptor in QSAR/QSPR model

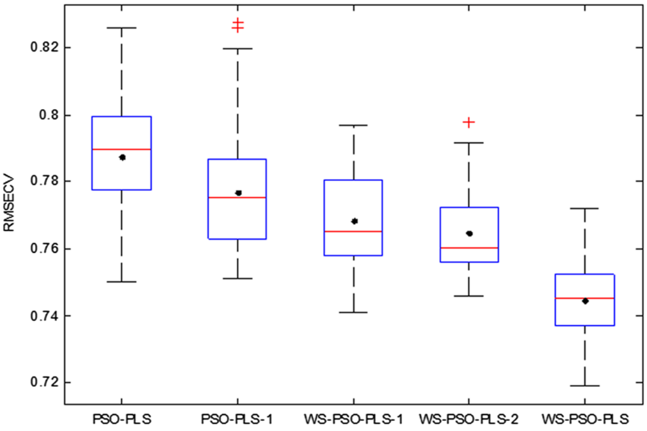

Figure 3. Boxplot of *RMSECV* for different methods on the artemisinin dataset. In each box, the horizontal line inside the box represents the median, the edges of the box are the 25th and 75th percentile, the whiskers extending to the most extreme data points are the maximum and minimum, the dot inside the box is the mean, and red "+" represents outliers. *RMSECV*, root mean square error from fivefold cross-validation

development. According to our careful observation, the descriptors with relatively larger normalized coefficients, marked by the black arrow in Figure 4(A), can be frequently chosen according to weighted sampling strategy. Taking the descriptor of *nring* as an example, it has the largest normalized coefficient. It is interesting to note that *nring* represents the number of rings in the compounds. Because the rings are basic elements of artemisinin analogues, it is undoubted that the use of *nring* with a relatively higher probability is very helpful for model development. Note, however, that PSO-PLS does not use any prior information of the model, and therefore equally treats each descriptor in QSAR/QSPR model development. With respect to the BZR dataset, we find out that the 10*th* descriptor has the largest normalized coefficient, which means that the corresponding descriptor (i.e., *S_sssN*) is very important in model development (Figure S5(A)). It is interesting to note that when the iteration terminates, this descriptor is consistently included by $\overrightarrow{p_g}$ in all the runs, which has the best Q^2 in the population. In addition to the previous two datasets, the similar result can also be observed in the selwood dataset (Figure S6(A)). Overall, PSO-PLS selects descriptors evenly regardless of their normalized coefficients, while WS-PSO-PLS has a reasonable trend to select descriptors with large normalized coefficients (see Figure 4(C), Figure S5(C), and Figure S6(C) for details).

Based on the aforementioned experiments, one can conclude that WS-PSO-PLS with less number of the selected descriptors has the capability to achieve better performance than PSO-PLS. As we all know that the interpretability is important in the PLS model development, which means the fewer descriptors with larger Q^2, the better. Although the number of the selected descriptors is not as few as expected, but WS-PSO-PLS still proves to be effective. We think this problem can be solved by optimizing the number of descriptors and Q^2 at the same time in future work through making use of other methods, such as weighted sum, multi-objective optimization, and so on.

Finally, to guard against the possibility of having learned such chance models, we have taken the artemisinin dataset and one of the optimal descriptor subsets as an example to validate the reliability of our QSPR/QSPR model. We used Y-randomization to check the robustness and chance correlation of the models. In Y-randomization test, the log RA values were randomly shuffled to change their true order. Thus, although the log RA values (and the statistical distribution) stayed the same, their position against the appropriate compound and its descriptors was now altered, thus destroying any meaningful relation that may have existed between independent variables and response values. By these new data such obtained, we can construct a large number of QSAR/QSPR models (e.g., 500) to get metrics like Q^2 and *RMSECV*. These metrics can be compared with those from the true model to obtain some hints about chance correlation. By constructing 500 models based on the permutated response values, we found that Q^2 and *RMSECV* are located in the range of [−0.2582, 0.0782] and [1.5054, 1.7587], respectively. Compared with the true model (i.e., mean $Q^2 = 0.7744$, mean $RMSECV = 0.7447$), there is a significant difference between the Q^2 of these shuffled models and the real one. The bad prediction statistics of these shuffled models suggest that our previous model indeed

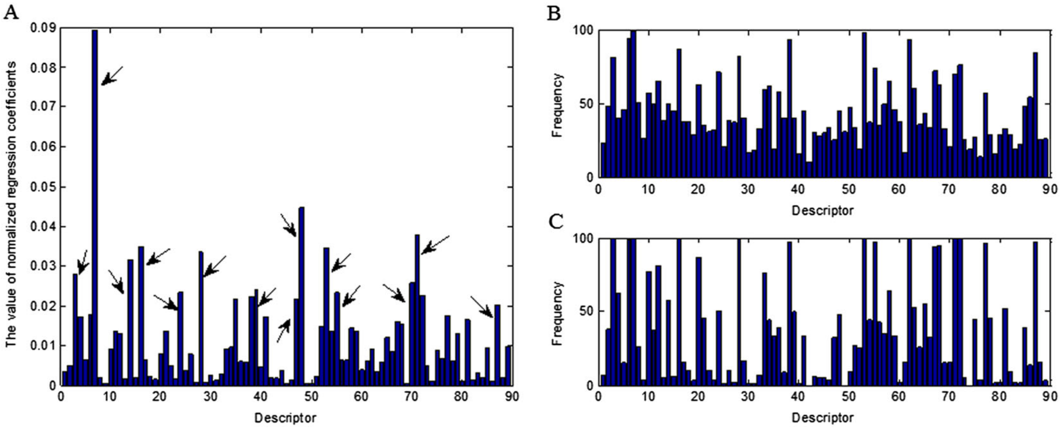

Figure 4. (A) Bar chart of the normalized PLS coefficients for the artemisinin dataset. In WS-PSO-PLS, the large normalized PLS regression coefficients, marked by the black arrow, can easily be selected according to the weighted sampling strategy. (B) The selection frequency of each descriptor for the artemisinin dataset in PSO-PLS. (C) The selection frequency of each descriptor for the artemisinin dataset in WS-PSO-PLS. WS-PSO-PLS, weighted sampling PSO-PLS; PSO-PLS, particle swarm optimization with partial least square.

reflects the true relationship between molecular descriptors and log RA values rather than from chance correlation.

4.2. Effect of the parameters

In this section, the effect of the parameters on the performance of WS-PSO-PLS has been discussed through various experiments.

4.2.1. *Effect of the function evaluations*

To fairly compare the performance between PSO-PLS and WS-PSO-PLS, the number of function evaluations (FES) was set to 10 000 in the previous experiments. Under this condition, the performance of WS-PSO-PLS is better than that of PSO-PLS on all the datasets. One question that naturally arises is whether the performance of PSO-PLS is similar to or even better than that of WS-PSO-PLS with the increase of FES. In order to answer the aforementioned question, we empirically studied on the effect of FES on the performance of WS-PSO-PLS and PSO-PLS. In our experiment, the number of FES was added to 50 000, and other parameters were kept unchanged. We chose the selwood dataset as an example to test the performance of WS-PSO-PLS and PSO-PLS. The evolutionary curves of the mean fitness provided by PSO-PLS and WS-PSO-PLS are shown in Figure 5. As shown in Figure 5, after 10 000 FES, the performance of PSO-PLS and WS-PSO-PLS slightly improves. Note, however, that the performance of WS-PSO-PLS is consistently superior to that of PSO-PLS during the evolution. Therefore, we can conclude that WS-PSO-PLS is able to find better descriptor subset with less number of the FES.

4.2.2. *Effect of the parameter b*

As mentioned previously, the main aim of the parameter *b* in Equation 4 is to control how much the PLS model information will be incorporated into the inferior individuals of the population. If the value of *b* is too small, too much model information will be injected into the poor individuals in the population. More importantly, nearly all the dimensions of the poor individuals in the population will be replaced by the corresponding dimensions of the weighted sampling individuals, which might have a side effect on the search ability of PSO. On the other hand, WS-PSO-PLS is almost the same as PSO-PLS if the value of *b* is too large, because little model information will be incorporated into the evolutionary process. In order to ascertain the effect of *b* on the performance of WS-PSO-PLS, we tested six different values of *b*: 0, 0.2, 0.4, 0.6, 0.8, and 1. In our experiments, the other parameters were kept unchanged, and the selwood dataset was taken as an example.

The experimental results have been given in Figure 6. As shown in Figure 6, $Q^2_{\max}$ of WS-PSO-PLS with $b = 0$, 0.2, 0.4, 0.6, 0.8, and 1.0 is equal to 0.9086, 0.9091, 0.9098, 0.9138, 0.9209, and 0.9012, respectively, which means the algorithm exhibits the best performance when $b = 0.8$. In addition, $Q^2_{\max}$ of PSO-PLS is equal to 0.8653. Therefore, WS-PSO-PLS has an advantage over PSO-PLS with all the values of *b*. It is necessary to note that when $b = 1$, the implementation of WS-PSO-PLS is equivalent to PSO-PLS expect for the initialized process. Due to the fact that WS-PSO-PLS utilizes the PLS model information in the initial population, WS-PSO-PLS performs better than PSO-PLS when $b = 1$.

We have tested the other two datasets and found that similar conclusion can be made. Based on the earlier experiments, $b = 0.8$ is recommended in our method. However, generally speaking, the optimal *b* value should be optimized according to the problems investigated.

4.3. Some further comments

Some key points are worth again highlighting in WS-PSO-PLS. In general, WS-PSO-PLS could be considered as a supervised version of PSO-PLS. With the help of the weighted sampling, PLS model information can be effectively incorporated into PSO to rapidly guide the search process at each generation. Thus, WS-PSO-PLS could progressively focus on that important descriptor subset for modeling. This idea is analogous to that from boosting to some extent. The other is the sampling way for descriptors during the generating of individuals at each generation. Instead of the simple random sampling in PSO-PLS, weighted sampling is used to select the more eligible descriptor subset for each

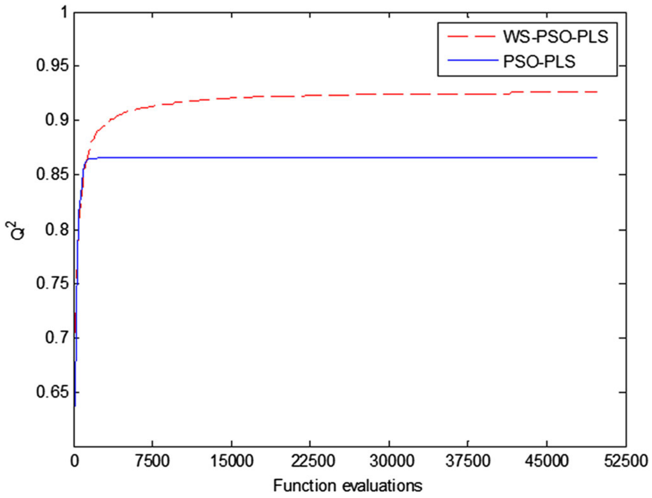

Figure 5. Relationship between the mean Q^2 and the function evaluations in PSO-PLS and WS-PSO-PLS for the selwood dataset. WS-PSO-PLS, weighted sampling PSO-PLS; PSO-PLS, particle swarm optimization with partial least square.

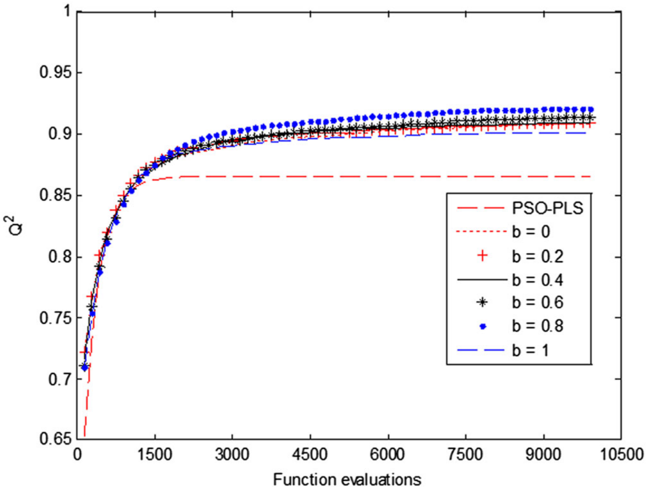

Figure 6. The mean Q^2 of PSO-PLS and WS-PSO-PLS with different *b* for the selwood dataset WS-PSO-PLS, weighted sampling PSO-PLS; PSO-PLS, particle swarm optimization with partial least square..

individual in WS-PSO-PLS. Unlike simple random sampling, which treats all descriptors impartially, weighted sampling takes a new strategy called “the survival of the fittest” and gives these descriptors different weights according to their contributions to regression. The descriptors with high weights are considered to be important for model development and therefore should have a high probability that WS-PSO-PLS selects it. The bigger the weight of the descriptor is, the higher the probability that WS-PSO-PLS selects the descriptor is. Ideally, if one descriptor does not contain any information at all for modeling, we hope that WS-PSO-PLS should not take this descriptor into account. That is, the descriptor should be removed from the descriptor pool. Thus, WS-PSO-PLS could effectively focus on those meaningful and important descriptors by feedback information from the PLS model information or prior information. Herein, it should be noted that we only used the simple PLS model coefficients to select the informative variables. In fact, other variable importance information could also be used in weighting sampling to guide the search process.

5. CONCLUSION

In this paper, we have introduced a novel idea of incorporating the PLS model information into PSO-PLS to improve its performance, and a new method called WS-PSO-PLS has been proposed. In WS-PSO-PLS, some individuals are generated through weighted sampling, which is conducted based on the normalized coefficients obtained by building the PLS model with all the descriptors, and some dimensions of these individuals are used to replace the corresponding dimensions of a few individuals with poor quality in the population at each generation. Moreover, a mutation strategy is performed to avoid getting trapping into a local optimum for WS-PSO-PLS. From the experimental results on three QSAR/QSPR datasets, we can conclude that the performance of WS-PSO-PLS is statistically better than that of the original PLS and PSO-PLS. Therefore, WS-PSO-PLS is a good alternative for descriptor selection in QSAR/QSPR model development. In the near future, we will try to apply our idea to other evolutionary algorithm diagrams in QSAR/QSPR studies for solving the problems with high-dimensional descriptor space. In addition, we also intend to design other methods to exploit the model information for descriptor selection in QSAR/QSPR.

Acknowledgements

This study was supported by the National Natural Science Foundation of China (grant nos. 81402853, 61273314, and 61175064) and the Chinese Postdoctoral Science Foundation (2014T70794, 2014M562142). The studies meet with the approval of the university’s review board.

REFERENCES

1. Nantasenamat C, Isarankura-Na-Ayudhya C, Naenna T, Prachayasittikul V. A practical overview of quantitative structure-activity relationship. *EXCLI J.* 2009; **8**: 1611–2156.
2. Wold S, Martens H, Wold H. The multivariate calibration problem in chemistry solved by the PLS method. In *Conference Proceeding Matrix Pencils*. Springer: Berlin Heidelberg, 1983; **973**: 286–293.
3. Geladi P, Kowalski B. Partial least-regression: a tutorial. *Anal. Chim. Acta* 1986; **185**: 1–17.
4. Rosipal R, Kramer N. *Overview and Recent Advances in Partial Least Squares*. Springer: Berlin Heidelberg, 2006; **3940**: 34–51.
5. Cao DS, Xu QS, Liang YZ, Chen X, Li HD. Automatic feature subset selection for decision tree-based ensemble methods in the prediction of bioactivity. *Chemometr Intell Lab* 2010; **103**: 129–136.
6. Fu GH, Cao DS, Xu QS, Li HD, Liang YZ. Combination of kernel PCA and linear support vector machine for modeling a nonlinear relationship between bioactivity and molecular descriptors. *J. Chemometr.* 2011; **25**: 92–99.
7. Huang X, Cao DS, Xu QS, Shen L, Huang JH, Liang YZ. A novel tree kernel support vector machine classifier for modeling the relationship between bioactivity and molecular descriptors. *Chemometr Intell Lab* 2013; **120**: 71–76.
8. Svetnik V, Liaw A, Tong C, Culberson JC, Sheridan RP, Feuston BP. Random forest: a classification and regression tool for compound classification and QSAR modeling. *J. Chem. Inf. Comput. Sci.* 2003; **43**: 1947–1958.
9. Cao DS, Yang YN, Zhao JC, Yan J, Liu S, Hu QN, Xu QS, Liang YZ. Computer-aided prediction of toxicity with substructure pattern and random forest. *J. Chemometr.* 2012; **26**: 7–15.
10. Svetnik V, Wang T, Tong C, Liaw A, Sheridan RP, Song QH. Boosting: an ensemble learning tool for compound classification and QSAR modeling. *J. Chem. Inf. Model.* 2005; **45**: 786–799.
11. Cao DS, Xu QS, Liang YZ, Zhang LX, Li HD. The boosting: a new idea of building models. *Chemometr Intell Lab* 2010; **100**: 1–11.
12. Lin WQ, Jiang JH, Shen Q, Shen GL, Yu RQ. Optimized block-wise variable combination by particle swarm optimization for partial least squares modeling in quantitative structure-activity relationship studies. *J. Chem. Inf. Model.* 2005; **2**: 486–493.
13. Shahlaei M. Descriptor selection methods in quantitative structure-activity relationship studies: a review study. *Chem. Rev.* 2013; **113**: 8093–8103.
14. Saeys Y, Inza I, Larranaga P. A review of feature selection techniques in bioinformatics. *Bioinformatics* 2007; **23**: 2507–2517.
15. Mehmood T, Liland KH, Snipen L, Sæbø S. A review of variable selection methods in partial least squares regression. *Chemometr Intell Lab* 2012; **118**: 62–69.
16. Tibshirani R. Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. B.* 1996; **58**: 267–288.
17. Efron BB, Hastie T, Johnstone I, Tibshirani R. Least angle regression. *Ann. Stat.* 2004; **32**: 407–499.
18. Yang SP, Song ST, Tang ZM, Song HF. Optimization of antisense drug design against conservative local motif in simulant secondary structures of HER-2 mRNA and QSAR analysis. *Acta Pharmacol. Sin.* 2003; **9**: 897–902.
19. Li HD, Liang YZ, Xu QS, Cao DS. Key wavelengths screening using competitive adaptive reweighted sampling method for multivariate calibration. *Anal. Chim. Acta* 2009; **648**: 77–84.
20. Cao DS, Wang B, Zeng MM, Liang YZ, Xu QS, Zhang LX, Li HD, Hu QN. A new strategy of exploring metabolomics data using Monte Carlo tree. *Analyst* 2011; **136**: 947–954.
21. Granitto PM, Furlanello C, Biasioli F, Gasperi F. Recursive feature elimination with random forest for PTR-MS analysis of agroindustrial products. *Chemometr Intell Lab* 2006; **83**: 83–90.
22. Teofilo RF, Martins JPA, Ferreira MMC. Sorting variables by using informative vectors as a strategy for feature selection in multivariate regression. *J. Chemometr.* 2009; **1**: 32–48.
23. Centner V, Massart DL, DeNoord OE. Jong Sde, Vandeginste BM, Sterna C. Elimination of uninformative variables for multivariate calibration. *Anal. Chem.* 1996; **68**: 3851–3858.
24. Izrailev S, Agrafiotis D. Variable selection for QSAR by artificial ant colony system. *SAR QSAR Environ. Res.* 2002; **13**: 417–423.
25. Niculescu SP. Artificial neural networks and genetic algorithms in QSAR. *J. Mol. Struc-THEOCHEM.* 2003; **622**: 71–83.
26. Dutta D, Guha R, Wild D, Chen T. Ensemble feature selection: consistent descriptor subsets for multiple QSAR models. *J. Chem. Inf. Model.* 2007; **47**: 989–997.
27. Yun YH, Wang WT, Tan ML, Liang YZ, Li HD, Cao DS, Lu HM, Xu QS. A strategy that iteratively retains informative variables for selecting optimal variable subset in multivariate calibration. *Anal. Chim. Acta* 2014; **807**: 36–43.
28. Kalivas JH, Roberts N, Sutter JM. Global optimization by simulated annealing with wavelength selection for ultraviolet–visible spectrophotometry. *Anal. Chem.* 1989; **61**: 2024–2030.
29. Arakawa M, Yamashita Y, Funatsu K. Genetic algorithm-based wavelength selection method for spectral calibration. *J. Chemometr.* 2011; **25**: 10–19.
30. Bangalore AS, Shaffer RE, Small GW, Arnold MA. Genetic algorithm-based method for selecting wavelengths and model size for use with

partial least-squares regression: application to near-infrared spectroscopy. *Anal. Chem.* 1996; **68**: 4200–4212.

31. Yun YH, Cao DS, Tan ML, Yan J, Ren DB, Xu QS, Yu L, Liang YZ. A simple idea on applying large regression coefficient to improve the genetic algorithm-PLS for variable selection in multivariate calibration. *Chemometr Intell Lab* 2014; **130**: 76–83.
32. Kubinyi H. Evolutionary variable selection in regression and PLS analyses. *J. Chemom.* 1996; **10**: 119–133.
33. Luke BT. 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.
34. Goodarzi M, Saeys W, Deeb O, etc. Particle swarm optimization and genetic algorithm as feature selection techniques for the QSAR modeling of imidazo[1,5-a]pyrido[3,2-e]pyrazines, inhibitors of phosphodiesterase 10A. *Chem. Biol. Drug Des.* 2013; **82**: 685–696.
35. Cheng ZJ, Zhang YT, Zhang WJ. QSAR studies of imidazopyridine derivatives as Et-PKG inhibitors using the PSO-SVM approach. *Med Chem Res.* 2010; **19**: 1307–1325.
36. Prakasvudhisarn C, Lawtrakul L. Feature set selection in QSAR of 1-[(2-Hydroxyethoxy) methyl]-6-(phenylthio)thymine (HEPT) analogues by using swarm intelligence. *Monatshefte für Chemie-Chemical Monthly* 2008; **139**: 197–211.
37. Shen Q, Jiang JH, Jiao CX, Shen GL, Yu RQ. Modified particle swarm optimization algorithm for variable selection in MLR and PLS modeling: QSAR studies of antagonism of angiotensin II antagonists. *Eur. J. Pharm. Sci.* 2004; **22**: 145–152.
38. Kennedy J, Eberhart R. Particle swarm optimization. In *IEEE International Conference on Neural Networks*. IEEE: Washington, DC, USA, 1995; **4**: 1942–1948.
39. Shi Y, Eberhart R. A modified particle swarm optimizer. In *IEEE International Conference on Evolutionary Computation (CEC' 98)*. IEEE: Anchorage, AK, 1998; 69–73.
40. Lucasius CB, Beckers MLM, Kateman G. Genetic algorithms in wavelength selection: a comparative study. *Anal. Chim. Acta* 1994; **286**: 135–153.
41. Cao DS, Liu S, Fan L, Liang YZ. QSAR analysis of the effects of OATP1B1 transporter by structurally diverse natural products using a particle swarm optimization-combined multiple linear regression approach. *Chemometr Intell Lab* 2014; **130**: 84–90.
42. Avery MA, Gaston MA, Rodrigues CR, Barreiro EJ, Cohen FE, Sabnis YA, Woolfrey JR. Structure activity relationships of the antimalarial agent artemisinin. The development of predictive in vitro potency models using CoMFA and HQSAR methodologies. *J. Med. Chem.* 2002; **45**: 292–303.
43. Guha R, Jurs P. The development of QSAR models to predict and interpret the biological activity of artemisinin analogues. *J. Chem. Inf. Comput. Sci.* 2004; **44**: 1440–1449.
44. Cao DS, Xu QS, Hu QN, Liang YZ. ChemoPy: freely available python package for computational biology and chemoinformatics. *Bioinformatics* 2013; **29**: 1092–1094.
45. Litina DH, Hansch C. Quantitative structure-activity relationships of the benzodiazepines. A review and reevaluation. *Chem. Rev.* 1994; **6**: 1483–1505.
46. Lu AJ, Liu B, Liu HB, Zhou JJ. 3D-QSAR study of benzodiazepines at five recombinant GABAA/benzodiazepine receptor subtypes. *Acta Phys. Chim. Sin.* 2004; **20**: 488–493.
47. Neaz MM, Muddassar M, Pasha FA, Cho SJ. 2D-QSAR of non-benzodiazepines to benzodiazepines receptor (BZR). *Med Chem Res.* 2009; **18**: 98–111.
48. Haefely W, Kyburz E, Gerecke M, Mohler H. Recent advances in the molecular pharmacology of benzodiazepine receptors and in the structure-activity relationships of their agonists and antagonists. *Adv. Drug Res.* 1985; **14**: 165–322.
49. Selwood DL, Livingstone DJ, Comley JC, Dowd BAO', Hudson AT, Jackson P, Jandu KS, Rose VS, Stables JN. Structure-activity relationships of antifilarial antimycin analogues: a multivariate pattern recognition study. *J. Med. Chem.* 1990; **33**: 136–142.
50. Nicolotti O, Gillet VJ, Fleming PJ, Green DVS. Multiobjective optimization in quantitative structure-activity relationships: deriving accurate and interpretable QSARs. *J. Med. Chem.* 2002; **45**: 5069–5080.
51. Winston H. *Wilcoxon Rank Sum Test. Encyclopedia of Systems Biology*. 2013; 2354–2355. DOI: 10.1007/978-1-4419-9863-7_1185.

SUPPORTING INFORMATION

Additional supporting information can be found in the online version of this article at the publisher's website.