QSAR Classification Modeling for Bioactivity of Molecular Structure via SPL-Logsum

Liang-Yong Xia, Qing-Yong Wang

Abstract—Quantitative structure-activity relationship (QSAR) modelling is effective 'bridge' to search the reliable relationship related bioactivity to molecular structure. A QSAR classification model contains a larger number of redundant, noisy and irrelevant descriptors. To address this problem, various of methods have been proposed for descriptor selection. Generally, they can be grouped into three categories: filters, wrappers, and embedded methods. Regularization method is an important embedded technology, which can be used for continuous shrinkage and automatic descriptors selection. In recent years, the interest of researchers in the application of regularization techniques is increasing in descriptors selection, such as, logistic regression (LR) with $L_1$ penalty. In this paper, we proposed a novel descriptor selection method based on self-paced learning (SPL) with Logsum penalized LR for predicting the bioactivity of molecular structure. SPL inspired by the learning process of humans and animals that gradually learns from easy samples/smaller losses to hard samples/bigger losses) samples into training and Logsum regularization has capacity to select few meaningful and significant molecular descriptors, respectively. Experimental results on simulation and three public QSAR datasets show that our proposed SPL-Logsum method outperforms other commonly used sparse methods in terms of classification performance and model interpretation.

Index Terms—QSAR; bioactivity; descriptor selection; SPL; Logsum

I. INTRODUCTION

Quantitative structure-activity relationship (QSAR) model is effective 'bridge' to search the reliable relationship between chemical structure and biological activities in the field of drug design and discovery [1]. The chemical structures are represented by a larger number of different descriptors. In general, only a few descriptors associated with bioactivity are favor of the QSAR model. Therefore, descriptor selection plays an important role in the study of QSAR that can eliminates redundant, noisy, and irrelevant descriptors [2].

In the recent years, various of methods have been proposed for descriptor selection. Generally, they can be grouped into three categories: filters, wrappers, and embedded methods [3]. Filter methods select descriptors according to some statistic criterion such as T-test. These selected descriptors will be a part of classification that used to classify the compounds [4] [5]. The drawback of filter method ignores the relationship between descriptors.

Wrapper methods utilize a subset of descriptors and train a model using them [6]. For example, forward selection adds the most important descriptors until the model is not statistically significant [7]. Backward elimination starts with the candidate descriptor and remove descriptors without some statistical descriptors [8]. Particle swarm optimization has a series of initial random particles and then selects descriptors by updating the velocity and positions [9]. Genetic algorithm initializes random particles and then uses the code, selection, exchange and mutation operations to select the subset of descriptors [10] [11]. However, these methods are usually computationally very expensive.

Another method for descriptor selection is embedded method that combines filter methods and wrapper methods [12]. Regularization method [13] is an important embedded technology, which can be used for continuous shrinkage and automatic descriptors selection. Various of regularization types have been proposed, such as, Logsum [14], $L_1$ [15], $L_{EN}$ [16], $L_{1/2}$ [17], SCAD [18], MCP [19], HLR($L_{1/2} + L_2$) [20] and so on. Recently, the regularization has been used in QSRR [21], QSPR [22] and QSTR [23] in the field of chemometrics. However, some individuals have focused their interest and attention on QSAR research. Besides, the interest of applying regularization technology to LR is increasing in the QSAR classification study. The LR model is considered to be an effective discriminant method because it provides the prediction probability of class members. In order to select the small subset of descriptors that can be for QSAR model of interest, various of regularized LR model have been development. For instance, Shevade SK et al [24] proposed sparse LR model with $L_1$ to extract the key variables for classification problem. However, the limitation of $L_1$ is resulting in a biased estimation for large coefficients. Therefore, its model selection is inconsistent and lack of oracle properties. In order to alleviate these, Xu et al [17] proposed $L_{1/2}$ penalty, which has demonstrated many attractive properties, such as unbiasedness, sparsity and oracle properties. Then, Liang et al. [25] used LR based on $L_{1/2}$ penalty to select a small subset of variables.

In this paper, we proposed a novel descriptor selection using SPL via sparse LR with Logsum penalty (SPL-Logsum) in QSAR classification. SPL is inspired by the learning process of humans that gradually incorporates the training samples into learning from easy ones to hard ones [26] [27] [28]. Different from the curriculum learning [29] that learns the data in a predefined order based on prior knowledge, SPL learns the training data in an order from easy to hard dynamically determined by the feedback of the learner itself. Meanwhile, the Logsum regularization proposed by Candès et al [14] produces better solutions with more sparsity. The flow diagram shows the process of our proposed SPL-Logsum for QSAR model in Fig. 1.

Our work has three main contributions:
In this paper, we only consider a general binary classification problem and get a predictor vector \( \mathbf{X} \) and a response variable \( y \), which consists of chemical structure and corresponding biological activities, respectively. Suppose we have \( n \) samples, \( D = (X_1, y_1), (X_2, y_2), \ldots, (X_n, y_n) \), where \( X_i = (x_{i1}, x_{i2}, \ldots, x_{ip}) \) is \( i \)th input pattern with dimensionality \( p \), which means the \( X_i \) has \( p \) descriptors and \( x_{ij} \) denotes the value of descriptor \( j \) for the \( i \)th sample. And \( y_i \) is a corresponding variable that takes a value of 0 or 1. Define a classifier \( f(x) = e^x/(1 + e^x) \) and the LR is given as follows:

\[
P(y_i = 1|X_i) = f(X_i^\prime \beta) = \frac{\exp(X_i^\prime \beta)}{1 + \exp(X_i^\prime \beta)},
\]

where \( \beta = (\beta_0, \beta_1, \ldots, \beta_p) \) are the coefficients that can be estimated, note that \( \beta_0 \) is the intercept. Additionally, the log-likelihood can be expressed as follows:

\[
l(\beta) = -\sum_{i=1}^{n} \{y_i \log[f(x_i^\prime \beta)] + (1 - y_i) \log[1 - f(x_i^\prime \beta)]\}, \quad (2)
\]

the \( \beta \) can be got by minimizing the equation (2). However, in high dimensional QSAR classification problem with \( n > p \), direct to solve the equation (2) can result in over-fitting. Therefore, in order to solve the problem, add the regularization terms to the equation (2).

\[
\beta = \arg\min\{l(\beta) + \lambda p(\beta)\}, \quad (3)
\]

where \( l(\beta) \) is loss function, \( p(\beta) \) is penalty function, \( \lambda > 0 \) is a tuning parameter. Note that \( p(\beta) = \sum |\beta|^q \). When \( q \) is equal to 1, the \( L_1 \) has been proposed. Moreover, there are various versions of \( L_1 \), such as \( L_{EN} \), SCAD, MCP, group.
lasso, and so on. We add the $L_1$ regularization to the equation (2). The formula is expressed as follows:

$$\beta = \text{argmin}\{l(\beta) + \lambda \sum_{j=1}^{p} |\beta_j|\}$$  \hspace{1cm} (4)

$L_1$ regularization has capacity to select the descriptors. However, the drawback of $L_1$ regularization can result in a biased estimation for large coefficients and be not sparsity. To address these problems, Xu et al [17] proposed $L_{1/2}$ regulation, which can be taken as a representative of $L_q$ ($0 < q < 1$) penalty. We can rewrite the equation (4) as follows:

$$\beta = \text{argmin}\{l(\beta) + \lambda \sum_{j=1}^{p} |\beta_j|^\frac{1}{2}\}$$  \hspace{1cm} (5)

Theoretically, the $L_0$ regularization produces better solutions with more sparsity, but it is an NP problem. Therefore, Candes et al [14], proposed the Logsum penalty, which approximates the $L_0$ regularization much better. We give the formula based on Logsum regularization as follows:

$$\beta = \text{argmin}\{l(\beta) + \lambda \sum_{j=1}^{p} \log(|\beta_j| + \varepsilon)\}$$  \hspace{1cm} (6)

where $\varepsilon > 0$ should be set arbitrarily small, to closely make the Logsum penalty resemble the $L_0$-norm. The Logsum regularization is non-convex in Fig. 2. The Equation (6) has a local minima [30]. A regularization can should satisfy three properties for the coefficient estimators: unbiasedness, sparsity and continuity in Fig. 3.

![Fig. 2. $L_1$ is convex, $L_{1/2}$ and Logsum are non-convex. The Logsum approximates to $L_0$.](image)

![Fig. 3. Three properties for the coefficient estimators: unbiasedness, sparsity and continuity for (a) $L_1$, (b) $L_{1/2}$ and (c) Logsum](image)

2) A coordinate descent algorithm for the Logsum penalized LR: In this paper, we used the coordinate descent algorithm to solve equation (6). The algorithm is a “one-at-a-time” and solves $\beta_j$ and other $\beta_{j \neq k}$ (represent the parameters remained after $j^{th}$ element is removed) is fixed. Suppose we have $n$ samples, $D = (X_1, y_1), (X_2, y_2),\ldots, (X_n, y_n)$, where $X_i = (x_{i1}, x_{i2},\ldots, x_{ip})$ is $i^{th}$ input pattern with dimensionality $p$, which means the $X_i$ has $p$ descriptors and $x_{ij}$ denotes the value of descriptor $j$ for the $i^{th}$ sample. And $y_i$ is a corresponding variable that takes a value of 0 or 1. According to Friedman et al. [31], Liang et al. [25] and Xia et al. [30], the univariate Logsum thresholding operator can be written as:

$$\beta_j = D(w_j, \lambda, \varepsilon) = \begin{cases} 
\text{sign}(w_j) \frac{e^{w_j} - 1} {\varepsilon} & \text{if } c_2 > 0 \\
0 & \text{if } c_2 \leq 0
\end{cases}$$  \hspace{1cm} (7)

where $\lambda > 0, 0 < \varepsilon < \sqrt{X}, c_1 = \omega_j - \varepsilon, c_2 = c_1^2 - 4(\lambda - w_j \varepsilon)$ and $w_j = \sum_{i=1}^{n} x_{ij}(y_i - \tilde{y}_j^{(i)}), \tilde{y}_j^{(i)} = \sum_{k \neq j} x_{ik}\beta_k$.

Inspired by Liang et al. [25], the equation (6) is linearized by one-term Taylor series expansion:

$$L(\beta, \lambda) = \frac{1}{2n} \sum_{i=1}^{n} (Z_i - X_i \tilde{\beta})^T W_i (Z_i - X_i \tilde{\beta}) + \lambda \sum_{j=1}^{n} \log(|\beta_j| + \varepsilon)$$  \hspace{1cm} (8)

where $\varepsilon > 0, Z_i = X_i \tilde{\beta} + \frac{Y_i - f(X_i, \tilde{\beta})}{f(X_i, \tilde{\beta})(1 - f(X_i, \tilde{\beta}))}$, $W_i = f(X_i, \tilde{\beta})(1 - f(X_i, \tilde{\beta}))$ and $f(X_i, \tilde{\beta}) = \exp(X_i, \tilde{\beta})/(1 + \exp(X_i, \tilde{\beta}))$.

Redefine the partial residual for fitting $\beta_j$ as $Z_i^{(j)} = \sum_{i=1}^{n} W_i (Z_i - \sum_{k \neq j} x_{ik}\beta_k)$ and $\sum_{i=1}^{n} x_{ij}(Z_i - Z_i^{(j)})$.

A pseudocode of coordinate descent algorithm for Logsum penalized LR model is shown in Algorithm 1.

![Algorithm 1: A coordinate descent algorithm for Logsum penalized LR model](image)

B. SPL-Logsum

Inspired by cognitive mechanism of humans and animals, Koller et al. [26], proposed a new learning called SPL (SPL) that learns from easy to hard samples. During the process of optimization, more samples are entered into training set from easy to hard by increasing gradually the penalty of the SPL regularizer. Suppose given a dataset $D = (X_i, y_i)_{i=1}^{n}$ with $n$ samples, $X_i$ and $y_i$ are the $i^{th}$ sample and its label, respectively. The value of $y_i$ is 0 or 1 in classification model. Let $(X_i, \beta)$ denote the learn model and $\beta$ is model parameter that should be estimated. $l(y_i, (X_i, \beta))$ is loss function of the $i^{th}$ sample. The SPL model combines
a weighted loss term and a general self-paced regularizer imposed on sample weight, given as:

\[ \min_{\beta,v \in [0,1]} E(\beta, v, \gamma, \lambda) = \sum_{i=1}^{n} (v_i l(y_i, f(X_i, \beta)) + \lambda p(\beta) + g(v_i, \gamma)) \]

where \( g(v_i, \gamma) = -\gamma v_i \) and \( p(\beta) = \log(|\beta| + \varepsilon) \). Therefore, the equation (9) can be rewritten as:

\[ \min_{\beta,v \in [0,1]} E(\beta, v, \gamma, \lambda) = \sum_{i=1}^{n} (v_i l(y_i, f(X_i, \beta)) + \lambda \log(|\beta| + \varepsilon) - \gamma v_i) \]

where \( \lambda \) is regularization parameter, \( \gamma \) is age parameter for controlling the learning pace, and \( \varepsilon > 0 \). According to Kumar et al [26], with the fixed \( \beta \), the optimal \( v^*_i \) is easily calculated by:

\[ v^*_i = \begin{cases} 1 & \text{if } l(y_i, f(X_i, \beta)) < \gamma \\ 0 & \text{otherwise} \end{cases} \]

Here, we give a explanation for the equation (11). In SPL iteration, when estimating the \( v \) with a fixed \( \beta \), a sample, which is taken as an easy sample (high-confidence sample with smaller loss value), can be selected \((v^* = 1)\) in training if the loss is smaller than \( \gamma \). Otherwise, unselected \((v^* = 0)\). When estimating the \( \beta \) with a fixed \( v \), the classifier only is used easy samples as training set. As the model "age" of \( \gamma \) is increasing, it means that more, probably hard samples with larger losses will be considered to train a more "mature" model. The process of our proposed SPL-Logsum method is shown in Fig. 4. Besides, A pseudocode of our proposed SPL-Logsum is shown in Algorithm 2.

**Algorithm 2: Pseudocode for SPL-Logsum**

**Input:** Input \( X,y,a \) stepsize \( \mu \) and the "age" \( \gamma \)

**Output:** Model parameter \( \beta \)

Initialize \( v^* \) and the value of loss

while not converged do

  update \( \beta^* \) used by Algorithm I

  update \( v^* \) according to E.q. (11)

end

return \( \beta^* \)

(3) Sensitivity, (4) Specificity and (5) P-value, which whether the selected descriptors are significant. In order to evaluate the performance of descriptors selection in simulation data, the sensitivity and specificity are defined as follows [20] [32]:

True Negative(TN):= \(|\bar{\beta} \cdot \bar{\beta}|_0\), False Negative(FN):= \(|\beta \cdot \bar{\beta}|_0\)

True Positive(TP):= \(|\beta \cdot \bar{\beta}|_0\), False Positive(FP):= \(|\bar{\beta} \cdot \bar{\beta}|_0\)

Sensitivity:= \(\frac{TP}{TP+FN}\), Specificity:= \(\frac{TN}{TN+FP}\)

where \(\cdot|_0\) is the element-wise product. \(|\cdot|_0\) calculates the number of non-zero element in a vector, are the on the vectors. The logical "not" operators of \( \beta \) and \( \bar{\beta} \) is \( \beta \) and \( \bar{\beta} \), respectively.

III. DATASETS

A. PubChem AID:651580

The dataset is provided by the Anderson cancer center at the university of Texas. There are 1,759 samples, of which 982 is active, and the 777 is inactive. The 1,875 descriptors can be calculated using QSARINS software. After pretreatment, we used 1,614 samples, of which 914 is active and 700 is inactive. Each sample contains 1,642 descriptors [33].

B. PubChem AID:743297

We could get this dataset from website [34]. By using QSARINS software and preprocessing, we utilized 200 samples, which consist of 58 active and 142 inactive, with 1588 descriptors for model as input.

C. PubChem AID:743263

The QSAR dataset is from Georgetown University. The number of active and inactive is 252 and 868, respectively. After preprocessing, 983 samples could be used for QSAR modelling. Each sample contains 1,612 descriptors. Note that know more about details from website [35].

IV. RESULTS

A. Analyses of simulated data

In this section, we evaluate the performance of our proposed SPL-logsum method in the simulation study. Four methods are compared to our proposed method, including LR with \( L_1 \), \( L_1/2 \), and Logsum penalties, respectively. In addition,
Some factors will be considered for constructing simulation, including the value of loss, model "age" $\gamma$, weight $v$, the confidence of sample, sample size $n$, correlation coefficient $\rho$ and noise control parameter $\sigma$.

1) Loss, "age" $\gamma$ and weight $v$: A sample, which is taken as an easy sample with small loss value, can be selected ($v^* = 1$) in training if the loss is smaller than $\gamma$. Otherwise, unselected ($v^* = 0$). As the model "age" of $\gamma$ is increasing, more probably hard samples with larger losses will be considered to train a more "mature" model. Therefore, we constructed a simple simulation about loss, $\gamma$ and $v$. First of all, a group of the value of loss is given. Then the model "age" is pre-set. At last, the selected samples are based on equation (11). Table I shows that the selection of samples is increasing with "age" $\gamma$.

2) High-confidence samples, medium-confidence samples, low-confidence samples: In this paper, we divided samples into three parts, including high-confidence samples, medium-confidence samples and low-confidence samples. The high-confidence samples are favored by the model, followed by medium-confidence samples and low-confidence samples. The low-confidence samples are probably noise or outline that can reduce the performance of QSAR model. In order to illustrate the process of selected samples for SPL, we constructed a simple simulation. First of all, the simulated data was generated from the LR with using normal distribution to produce $X^{n \times p}$, where $n = 100$, $p = 1000$. Then, a set of coefficients $\beta$ is given. Finally, we can calculate the value of $y$. Fig. 5 shows that a set of samples can be divided into three parts. Therefore, in SPL iteration, we could count up the number of selected samples, where is from high-confidence samples, medium-confidence samples and low-confidence samples. Fig. 6 indicates that at the beginning of the SPL, the model inclines to select high-confidence samples. Afterwards, when the model age gets larger, it tends to incorporate medium-confidence and low-confidence samples to train a mature model.

### TABLE I

| Samples | A | B | C | D | E | F | G | H | I | J | K | L | M | N |
|---------|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Loss    | 0.05 | 0.12 | 0.12 | 0.15 | 0.4 | 0.2 | 0.18 | 0.35 | 0.15 | 0.16 | 0.2 | 0.5 | 0.3 |

When the "age" $\gamma = 0.15$

| V       | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |

[SPL selects: A B C D]

When the "age" $\gamma = 0.2$

| V       | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 0 | 1 | 1 | 0 | 0 | 0 | 0 |

[SPL selects: A B C D E G H J K]

When the "age" $\gamma = 0.25$

| V       | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 1 | 0 | 1 | 1 | 0 | 0 | 0 |

[SPL selects: A B C D E G H J K L]

When the "age" $\gamma = 0.3$

| V       | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 1 | 0 | 1 | 1 | 0 | 1 | 1 |

[SPL selects: A B C D E G H J K L M N]

### Fig. 5
A set of samples can be divided into three parts. The color blue, gray and yellow blocks represent high-confidence samples, medium-confidence samples and low-confidence samples, respectively.

### Fig. 6
The results got by SPL. The number of samples is selected while the model age is increasing ($\gamma = 0.2$, $\mu = 0.05$)

3) Sample size $n$, correlation coefficient $\rho$ and noise control parameter $\sigma$: In this section, we constructed a simulation with sample size $n = 200,300$, correlation coefficient $\rho = 0.2,0.6$ and noise control parameter $\sigma = 0.3,0.9$. The process of construction is expressed as:

Step I: use normal distribution to produce $X$. Here, the number of row($X$) is sample ($n$) and the number of column($X$) is variable ($p = 1000$), respectively.

$$\log\left( \frac{y_i}{1 - y_i} \right) = \beta_0 + \sum_{j=1}^{p} x_{ij}\beta_j + \sigma \epsilon$$  (12)

where $y = (y_1, ..., y_n)^T$ is the vector of $n$ response variables, $X = \{X_1, X_2, ..., X_n\}$ is the generated matrix with $X_i = (x_{i1}, ..., x_{ip})$, $\epsilon = (\epsilon_1, ..., \epsilon_n)^T$ is the random error, $\sigma$ controls the signal to noise.

Step II: add different correlation parameter $\rho$ to simulation data.

$$x_{ij} = \rho \times x_{1j} + (1 - \rho)x_{ij}, i \sim (1, ..., n), j \sim (2, 3, 4, 5)$$  (13)

Step III: validate variable selection, the coefficients(10) are set in advance from 1 to 10.

$$\beta = \begin{bmatrix} 1 & -1 & -1.5 & -3 & 2 & ..., 2 & 0 & 0 & 0, ..., 0 \end{bmatrix}$$  (14)

Where $\beta$ is the coefficient.

Step IV: we can get $y$ from equation (12), (13), (14).

In this simulation study, first of all, 100 groups of data are constructed with $n$, $\rho$ and $\sigma$. Then, divide the dataset into...
training set and testing set (training set:testing set = 7:3). And then the coefficients are pre-set in advance. Finally, the LR with different penalties are used to select variables and builded model, including our proposed method. Note that the results should be averaged.

| Factors | Methods | Testing dataset | \( \beta \) |
|---------|---------|-----------------|------------|
| \( n \) | \( \rho \) | \( \sigma \) | \[ AUC \] | Sensitivity | Specificity | Accuracy | Sensitivity | Specificity |
| \( L_1 \) | 0.7056 | 0.7000 | 0.7667 | 0.7333 | 0.7000 | 0.7000 | 0.7000 | 0.7000 | 0.7000 |
| \( L_{1/2} \) | 0.7133 | 0.7667 | 0.7333 | 0.7500 | 0.7000 | 0.7000 | 0.7000 | 0.7000 | 0.7000 |
| \( 0.3 \) | Logsum | 0.7578 | 0.8333 | 0.8000 | 0.8000 | 0.8000 | 0.8000 | 0.8000 | 0.8000 |
| SPL-Logsum | 0.8711 | 0.8333 | 0.9000 | 0.8667 | 0.8000 | 0.8000 | 0.8000 | 0.8000 | 0.8000 |
| \( 0.2 \) | Logsum | 0.6296 | 0.5862 | 0.6774 | 0.6333 | 0.7000 | 0.7000 | 0.7000 | 0.7000 |
| SPL-Logsum | 0.7397 | 0.6552 | 0.7742 | 0.7167 | 0.5000 | 0.5000 | 0.5000 | 0.5000 |
| \( 0.9 \) | Logsum | 0.7553 | 0.6897 | 0.7742 | 0.7333 | 0.7000 | 0.7000 | 0.7000 | 0.7000 |
| | SPL-Logsum | 0.7998 | 0.7241 | 0.8710 | 0.8667 | 0.7000 | 0.7000 | 0.7000 | 0.7000 |

B. Analyses of real data

Three public QSAR datasets are got from website, including AID 651580, AID 743297 and AID 74362. We utilized random sampling to divide datasets into training datasets and testing dataset (70% for training set and 30% for testing set). A brief description of these datasets is shown in Table III.

| Dataset | No.of Samples(class/class2) | No.of descriptors |
|---------|-----------------------------|-------------------|
| AID:651580 | 1614(914 ACTIVE/700 INACTIVE) | 1642 |
| AID:743297 | 200(58 ACTIVE/ 142 INACTIVE) | 1588 |
| AID:743263 | 983(229 ACTIVE/ 754 INACTIVE) | 1612 |

Table IV shows that the experiment results are got by \( L_1 \), \( L_{1/2} \), Logsum and SPL-Logsum for testing data. Our proposed SPL-Logsum is better than other methods in terms of AUC, sensitivity, specificity and accuracy. For example, from the view of dataset AID:743297, the AUC obtained by our proposed SPL-Logsum is 0.7816 higher than 0.7368, 0.7045 and 0.7765 of \( L_1 \), \( L_{1/2} \) and Logsum. Moreover, the accuracy of SPL-Logsum is about increased by 17%, 9%, 4% of \( L_1 \), \( L_{1/2} \) and Logsum for dataset AID:651580. Furthermore, our proposed SPL-Logsum method is more sparsity than other methods. For example, for dataset AID:743297 in Fig. VII the number of selected descriptors of SPL-logsum is 10, ranks the first; then next is Logsum with 23; followed

According to existing literature [37], we have learned that the predictive ability of a QSAR model can only be estimated using testing set of compounds. Therefore, we poured more interest and attention into testing dataset, of which can prove the generalization ability of the model. Table II exhibits that the experiment results are got by \( L_1 \), \( L_{1/2} \) and Logsum, including our proposed SPL-Logsum methods. The performance of \( \beta \) got by our proposed SPL-logsum method are better than those of \( L_1 \), \( L_{1/2} \) and Logsum. For example, when \( n = 200, \rho = 0.2 \) and \( \sigma = 0.3 \), the sensitivity of \( L_1 \), Logsum and SPL-Logsum is 0.8, 0.8 and 0.8 higher than 0.7 of \( L_{1/2} \), the specificity obtained by our proposed SPL-Logsum is the highest among four methods with 0.9980. Logsum , \( L_{1/2} \), \( L_1 \) rank in the second, the third and the fourth place with 0.9939, 0.9889 and 0.9495. Besides, we analyzed the performance of testing set. For example, when \( n = 300 \) and \( \rho = 0.6 \), the values of AUC, sensitivity, specificity and accuracy of SPL-logsum are decreased to 0.9222 to 0.9310, 0.8750, 0.9000 and 0.8889, in which \( \sigma \) is from 0.3 to 0.9. When the sample size \( n \) increases, the performances of all the four methods are improved. For example, when \( \rho = 0.2 \) and \( \sigma = 0.3 \), for testing set, the AUC and accuracy of SPL-Logsum is increased by 11% and 10% with \( n = 200 \) and 300. When the \( \rho \) decreases, the performance of four methods are decreased with \( \sigma = 0.9 \). For example, when \( n = 300 \), the results of SPL-logsum are decreased from 0.9310, 0.8750, 0.9000 and 0.8889 to 0.9030, 0.9070, 0.8511 and 0.8778 from \( \rho = 0.6 \) to \( \rho = 0.2 \). In a word, our proposed SPL-logsum approach is superior to \( L_1 \), \( L_{1/2} \) and Logsum in the simulation dataset.
Table V: The results obtained by four methods.

| Dataset | Methods | AUC | Sensitivity | Specificity | Accuracy |
|---------|---------|-----|-------------|-------------|----------|
| L1      | 0.7063  | 0.6000 | 0.6800 | 0.6333 |
| L1/2    | 0.6697  | 0.7353 | 0.6923 | 0.7167 |
| Logsum  | 0.8251  | 0.7429 | 0.8000 | 0.7667 |
| SPL-Logsum | 0.8583 | 0.8000 | 0.8000 | 0.8000 |

Table VI: The 12 top-ranked meaningful and significant descriptors identified by L1, L1/2, Logsum and SPL-Logsum from the AID: 651580 dataset. (The common descriptors are emphasized in bold.)

| Rank | L1 | L1/2 | logsum | SPL-Logsum |
|------|----|------|--------|------------|
| 1    | L3i | MLFER_BH | MATS2e | gmin |
| 2    | L3u | MATS5m | ATSC6v | SHBint8 |
| 3    | L3s | maxHCats | nHibn4 | GATS2v |
| 4    | L3e | SpManS_Bhe | gmin | TBD6e |
| 5    | maxN | maxN | ATSC3e | C1SP1 |
| 6    | SssNH | minsC | ATSC8v | MLFER_A |
| 7    | SpMAD_Dzs | ATSC3e | GATS8e | MATS6m |
| 8    | minN | TDB5t | SHBint9 | minHbnu4 |
| 9    | MATS8m | naaO | minsC | ATSC2i |
| 10   | nHoaNH | ATSC4i | ASP-3 | GATS3m |
| 11   | RDF40s | SssssC | SssssCH | E2p |
| 12   | RDF85s | ATSC4s | MATS1e | nHCH2 |

Table VII: The 7 top-ranked meaningful and significant descriptors identified by L1, L1/2, Logsum and SPL-Logsum from the AID: 743297 dataset. (The common descriptors are emphasized in bold.)

| Rank | L1 | L1/2 | logsum | SPL-Logsum |
|------|----|------|--------|------------|
| 1    | SIC1 | AAF33s | VCH-3 | nAcid |
| 2    | BIC1 | ndN | ATSC5e | ATSC3s |
| 3    | MATS7v | RDF30m | ATSC7v | ATSC5e |
| 4    | ATSC7v | ETA_Shape_P | nHibn4 | SssssCH |
| 5    | MLFER_S | GATS4e | AATS3s | minHcaHu |
| 6    | RDF30m | nO | ATSC7i | minsNH |
| 7    | RDF70p | ATSC4c | GATS2p | RDF100v |

V. Conclusion

In the field of drug design and discovery, only a few descriptors related to bioactivity are selected that can be for QSAR model of interest. Therefore, descriptor selection is an attractive method that reflect the bioactivity in QSAR modeling. In this paper, we proposed a novel descriptor selection using SPL via sparse LR with Logsum penalty in QSAR classification. SPL can identify the easy and hard samples adaptively according to what the model has already learned and gradually add harder samples into training and Logsum regularization can select few meaningful and significant molecular descriptors simultaneously, respectively.

Both experimental results on artificial and three QSAR datasets demonstrate that our proposed SPL-Logsum method is superior to L1, L1/2 and Logsum. Therefore, our proposed method is the effective technique in both descriptor selection and prediction of biological activity.

In this paper, SPL has capacity to learn from easy samples to hard samples. However, we ignore an important aspect in learning: diversity. We plan to incorporate this information into our proposed method in our future work.

Appendix A

Abbreviations:

QSAR: Quantitative structure–activity relationship
QSRR: Quantitative structure–(chromatographic) retention relationships
QS PR: Quantitative structure-property relationship
QSTR: Quantitative structure-toxicity relationship
MCP: Maximum concave penalty
SCAD: Smoothly clipped absolute deviation
L1: LASSO
L1_1: Elastic net
LR: Logistic regression
SPL: Self-paced learning
AUC: Area under the curve

References

[1] A. R. Katritzky, M. Kuanar, S. Slavov, C. D. Hall, M. Karelson, I. Kahn, and D. A. Dobchev, “Quantitative correlation of physical and chemical properties with chemical structure: utility for prediction,” Chemical reviews, vol. 110, no. 10, pp. 5714–5789, 2010.
The 10 top-ranked meaningful and significant descriptors identified by $L_1$, $L_{1/2}$, LOGSUM and SPL-LOGSUM from the AID: 743263 dataset. (The common descriptors are emphasized in bold.)

| Rank | $L_1$ | $L_{1/2}$ | logsum | SPL-logsum |
|------|-------|-----------|--------|------------|
| 1    | maxsF | RDF130u   | LipinskiFailures | AATS55v    |
| 2    | minF  | nBondsD2  | nFRing | AATS65v    |
| 3    | SRW9  | RDF25m    | SCH-3   | ATSC2i     |
| 4    | SBWT  | nhA/NH    | SsaNH   | MATS3m     |
|      | MDEC-13 | minM1n    | minHBin2 | nBondsD2  |
| 6    | ETA_dAlpha_A | SpMint7_Bhi | ATSC3p | nHBin3     |
| 7    | minICats | nhH/CH   | RDF45u  | minHAvn    |
| 8    | RDF95u | RDF45s    | ATSC1v  | maxHBin10  |
| 9    | TDB10m | C3SP2     | MDEO-11 | nFRing     |
| 10   | C3SP2 | RDF75v    | AATS8p  | RDF20v     |

[2] M. Shahlaei, “Descriptor selection methods in quantitative structure–activity relationship studies: a review study,” Chemical reviews, vol. 113, no. 10, pp. 8093–8103, 2013.
[3] I. Guyon, S. Gunn, A. Ben-Hur, and G. Dror, “Result analysis of the nips 2003 feature selection challenge,” in Advances in neural information processing systems, 2005, pp. 545–552.
[4] W. Duch, “Filter methods,” in Feature Extraction. Springer, 2006, pp. 105–117.
[5] L. Yu and H. Liu, “Feature selection for high-dimensional data: A fast correlation-based filter solution,” in Proceedings of the 20th international conference on machine learning (ICML-03), 2003, pp. 856–863.
[6] A. G. Karegowda, M. Jayaram, and A. Manjunath, “Feature subset selection problem using wrapper approach in supervised learning,” International journal of Computer applications, vol. 1, no. 7, pp. 13–17, 2010.
[7] D. C. Whitely, M. G. Ford, and D. J. Livingstone, “Unsupervised forward selection: a method for eliminating redundant variables,” Journal of Chemical Information and Computer Sciences, vol. 40, no. 5, pp. 1160–1168, 2000.
[8] K. Z. Mao, “Orthogonal forward selection and backward elimination algorithms for feature subset selection,” IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics), vol. 34, no. 1, pp. 629–634, 2004.
[9] J. Kennedy, “Particle swarm optimization,” in Encyclopedia of machine learning. Springer, 2011, pp. 760–766.
[10] K. Deb, A. Pratap, S. Agarwal, and T. Meyarivan, “A fast and elitist multiobjective genetic algorithm: Nsga-ii,” IEEE transactions on evolutionary computation, vol. 6, no. 2, pp. 182–197, 2002.
[11] B. Hemmatatnejad, M. Akhond, R. Miri, and M. Shamisipour, “Genetic algorithm applied to the selection of factors in principal component-artificial neural networks: application to qsr study of calcium channel antagonist activity of 1, 4-dihydropyridines (nifedipine analogues),” Journal of chemical information and computer sciences, vol. 43, no. 4, pp. 1328–1334, 2003.
[12] T. N. Lal, O. Chapelle, J. Weston, and A. Elisseeff, “Embedded methods,” in Feature extraction. Springer, 2006, pp. 137–165.
[13] B. Schölkopf and A. J. Smola, Learning with kernels: regularization, optimization, and beyond. MIT press, 2002.
[14] E. J. Candes, M. B. Wakin, and S. P. Boyd, “Enhancing sparsity by reweighted 1 minimization,” Journal of Fourier analysis and applications, vol. 14, no. 5-6, pp. 877–905, 2008.
[15] R. Tibshirani, “Regression shrinkage and selection via the lasso,” Journal of the Royal Statistical Society. Series B (Methodological), pp. 267–288, 1996.
[16] H. Zou and T. Hastie, “Regularization and variable selection via the elastic net,” Journal of the Royal Statistical Society: Series B (Statistical Methodology), vol. 67, no. 2, pp. 301–320, 2005.
[17] Z. Xu, H. Zhang, Y. Wang, X. Chang, and Y. Liang, “L 1/2 regularization,” Science China Information Sciences, vol. 53, no. 6, pp. 1159–1169, 2010.
[18] J. Fan and R. Li, “Variable selection via nonconcave penalized likelihood and its oracle properties,” Journal of the American statistical Association, vol. 96, no. 456, pp. 1348–1360, 2001.
[19] C.-H. Zhang et al., “Nearly unbiased variable selection under minimax concave penalty,” The Annals of statistics, vol. 38, no. 2, pp. 894–942, 2010.
[20] H.-H. Huang, X.-Y. Liu, and Y. Liang, “Feature selection and cancer classification via sparse logistic regression with the hybrid 11/2+ 2 regularization,” PloS one, vol. 11, no. 5, p. e0149675, 2016.
[21] E. Daghir-Wojtkowiak, P. Wiczing, S. Bocian, Ł. Kubik, P. Kośliński, B. Buszewski, R. Kaliszewska, and M. J. Markuszewski, “Least absolute shrinkage and selection operator and dimensionality reduction techniques in quantitative structure retention relationship modeling of retention in hydrophilic interaction liquid chromatography,” Journal of Chromatography A, vol. 1403, pp. 54–62, 2015.
[22] M. Goodarzi, T. Chen, and M. P. Freitas, “Qsr predictions of heat of fusion of organic compounds using bayesian regularized artificial neural networks,” Chemometrics and Intelligent Laboratory Systems, vol. 104, no. 2, pp. 260–264, 2010.
[23] R. Aalizadeh, C. Peter, and N. S. Thomaids, “Prediction of acute toxicity of emerging contaminants on the water flea daphnia magna by ant colony optimization–support vector machine qsr models,” Environmental Science: Processes & Impacts, vol. 19, no. 3, pp. 438–448, 2017.
[24] S. K. Shevade and S. S. Keerthi, “A simple and efficient algorithm for gene selection using sparse logistic regression,” Bioinformatics, vol. 19, no. 17, pp. 2246–2253, 2003.
[25] Y. Liang, C. Liu, X.-Z. Luan, K.-S. Leung, T.-M. Chan, Z.-B. Xu, and H. Zhang, “Sparse logistic regression with a 1 1/2 penalty for gene selection in cancer classification,” BMC bioinformatics, vol. 14, no. 1, p. 198, 2013.
[26] M. P. Kumar, B. Packer, and D. Koller, “Self-paced learning for latent variable models,” in Advances in Neural Information Processing Systems, 2010, pp. 1189–1197.
[27] L. Jiang, D. Meng, S.-I. Yu, Z. Lan, S. Shan, and A. Hauptmann, “Self-paced learning with diversity,” in Advances in Neural Information Processing Systems, 2014, pp. 2078–2086.
[28] D. Meng, Q. Zhao, and L. Jiang, “What objective does self-paced learning indeed optimize?” arXiv preprint arXiv:1511.06049, 2015.
[29] Y. Bengio, J. Louradour, R. Collobert, and J. Weston, “Curriculum learning,” in Proceedings of the 26th annual international conference on machine learning. ACM, 2009, pp. 41–48.
[30] L.-Y. Xia, Y.-W. Wang, D.-Y. Meng, X.-J. Yao, H. Chai, and Y. Liang, “Descriptor selection via log-sum regularization for the biological activities of chemical structure,” International journal of molecular sciences, vol. 19, no. 1, p. 30, 2017.
[31] J. Friedman, T. Hastie, and R. Tibshirani, “Regularization paths for generalized linear models via coordinate descent,” Journal of statistical software, vol. 35, no. 1, p. 1, 2010.
[32] W. Zhang, Y.-w. Wan, G. I. Allen, K. Pang, M. L. Anderson, and Z. Liu, “Molecular pathway identification using biological network-regularized logistic models,” BMC genomics, vol. 14, no. 8, p. S7, 2013.
[33] AID:651580, “Pubchem bioassay summary,” https://pubchem.ncbi.nlm.nih.gov/bioassay/651580, accessed April 19, 2018.
[34] AID:743297, “Pubchem bioassay summary,” https://pubchem.ncbi.nlm.nih.gov/bioassay/743297, accessed April 19, 2018.
[35] AID:743263, “Pubchem bioassay summary,” https://pubchem.ncbi.nlm.nih.gov/bioassay/743263, accessed April 19, 2018.
[36] L. Jiang, D. Meng, Q. Zhao, S. Shan, and A. G. Hauptmann, “Self-paced curriculum learning,” in AAAI, vol. 2, no. 5.4, 2015, p. 6.
[37] A. Golbraikh and A. Tropsha, “Beware of q2!” Journal of molecular graphics and modelling, vol. 20, no. 4, pp. 269–276, 2002.
TABLE IX
THE DETAIL INFORMATION OF DESCRIPTORS OBTAINED BY SPL-LOGSUM METHOD.

| Descriptors | Name                                                                 | P-value       | Class |
|-------------|----------------------------------------------------------------------|---------------|-------|
| gmin        | Minimum E-State                                                     | 0.00050       | 2D    |
| TDB6e       | 3D topological distance based autocorrelation - lag 6 / weighted by Sanderson electronegativities | 0.00368       | 3D    |
| GATS2s      | Geary autocorrelation - lag 2 / weighted by I-state                 | 0.00576       | 2D    |
| C1SP1       | Triply bound carbon bound to one other carbon                       | 0.00849       | 2D    |
| MLFER_A     | Overall or summation solute hydrogen bond acidity                    | 0.01383       | 2D    |
| SHBint8     | Sum of E-State descriptors of strength for potential hydrogen bonds of path length 8 | 0.01620       | 2D    |
| MATS3m      | Moran autocorrelation - lag 8 / weighted by mass                    | 0.02395       | 2D    |
| E2p         | 2nd component accessibility directional WHIM index / weighted by relative polarizabilities | 0.03345       | 3D    |
| minHBint4   | Minimum E-State descriptors of strength for potential Hydrogen Bonds of path length 4 | 0.04056       | 2D    |
| nHiCCH2     | Minimum atom-type H E-State: =CH2                                   | 0.04883       | 2D    |
| GATS3m      | Geary autocorrelation - lag 3 / weighted by mass                    | 0.03929       | 2D    |
| ATSC2i      | Centered Broto-Moreau autocorrelation - lag 2 / weighted by first ionization potential | 0.06968       | 2D    |
| nAcid       | Number of acidic groups. The list of acidic groups is defined by these SMARTS "([O;H1]-[C,S,P]=O)" | 0.00569       | 2D    |
| ATSC3s      | Centered Broto-Moreau autocorrelation - lag 3 / weighted by I-state | 0.00866       | 2D    |
| AATSC5e     | Average centered Broto-Moreau autocorrelation - lag 5 / weighted by Sanderson electronegativities | 0.00210       | 2D    |
| SssCH       | Sum of atom-type E-State: > CH-                                    | 0.01260       | 2D    |
| minHCsatu   | Minimum atom-type H E-State: H\ on C sp3 bonded to unsaturated C    | 0.04492       | 2D    |
| minssNH     | Minimum atom-type E-State: -NH2+                                   | 0.00025       | 2D    |
| RDF100v     | Radial distribution function - 100 / weighted by relative van der Waals volumes | 0.01730       | 3D    |
| AATS5v      | Average Broto-Moreau autocorrelation - lag 5 / weighted by van der Waals volumes | 0.08453       | 2D    |
| AATS6s      | Average Broto-Moreau autocorrelation - lag 6 / weighted by I-state  | 0.03047       | 2D    |
| ATSC2t      | Centered Broto-Moreau autocorrelation - lag 2 / weighted by first ionization potential | 0.00536       | 2D    |
| MATS3m      | Moran autocorrelation - lag 3 / weighted by mass                    | 0.00037       | 2D    |
| nBondsD2    | Total number of double bonds (excluding bonds to aromatic bonds)    | 0.00059       | 2D    |
| nHBint3     | Count of E-State descriptors of strength for potential Hydrogen Bonds of path length 3 | 0.00041       | 2D    |
| minHAvin    | Minimum atom-type H E-State: H on C vinyl bonded to C aromatic      | 0.02756       | 2D    |
| maxHBint10  | Maximum E-State descriptors of strength for potential Hydrogen Bonds of path length 10 | 0.04263       | 2D    |
| nFRing      | Number of fused rings                                               | 0.03005       | 2D    |
| RDF20s      | Radial distribution function - 020 / weighted by relative I-state   | 0.05537       | 3D    |