Research article

Journal of CHEMOMETRICS

Received: 21 February 2015, Revised: 07 June 2015, Accepted: 17 July 2015, Published online in Wiley Online Library: 26 August 2015

(wileyonlinelibrary.com) DOI: 10.1002/cem.2741

High-dimensional QSAR prediction of anticancer potency of imidazo[4,5-b]pyridine derivatives using adjusted adaptive LASSO

Zakariya Yahya Algamal[a], Muhammad Hisyam Lee[a]*, Abdo M. Al-Fakih[b] and Madzlan Aziz[b]

In high-dimensional quantitative structure–activity relationship (QSAR) studies, identifying relevant molecular descriptors is a major goal. In this study, a proposed penalized method is used as a tool for molecular descriptors selection. The method, called adjusted adaptive least absolute shrinkage and selection operator (LASSO) (AALASSO), is employed to study the high-dimensional QSAR prediction of the anticancer potency of a series of imidazo[4,5-b]pyridine derivatives. This proposed penalized method can perform consistency selection and deal with grouping effects simultaneously. Compared with other commonly used penalized methods, such as LASSO and adaptive LASSO with different initial weights, the results show that AALASSO obtains the best predictive ability not only by consistency selection but also by encouraging grouping effects in selecting more correlated molecular descriptors. Hence, we conclude that AALASSO is a reliable penalized method in the field of high-dimensional QSAR studies. 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: adaptive LASSO; consistency selection; grouping effects; QSAR; anticancer potency

1. INTRODUCTION

Cancer is a term that refers to uncontrolled cellular division, growth, and spread of abnormal cells. It can occur in all body parts. Cancer cells can attack the neighboring undamaged parts of the body and spread to affect other organs [1]. Cancer is a disease that threatens human lives and causes the second highest rate of death globally [2–4]. Although there is continuous progress in the development of cancer treatment, the challenge to develop successful anticancer agents remains [5]. A new method to treat cancer is to use anticancer drugs to act against existing proteins in the proliferation of cancer cells. Aurora kinase, a family of serine/threonine, is a group of kinases that is responsible for regulating the cell cycle [1]. There are three kinds of Aurora kinases; namely, Aurora A, B, and C, which play distinct roles in mitosis regulation [6]. Aurora A is one type of the isoforms of Aurora kinase enzymes and has a catalytic effect during mitosis [1].

According to previous studies, Aurora A has attracted attention in the oncology field because it is involved in a wide range of cancers, such as colorectal, prostate, ovarian, breast, and glioma [7]. Several Aurora kinase inhibitors have been identified as excellent antitumor inhibitors. Recently, it has been reported that a series of imidazo[4,5-b]pyridine derivatives possess excellent potencies as orally bioavailable Aurora A inhibitors [7,8].

Quantitative structure–activity relationship (QSAR) study has become of great importance in computational chemistry and biochemistry. The principle of QSAR is to model several biological activities over a collection of chemical compounds in terms of their structural properties [9]. Consequently, QSAR is a mathematical model that can be used to predict the biological activity of new compounds. Analysis of multiple linear regression (MLR) is one of the most important tools for constructing the QSAR model. It is used for analyzing the relationship between several predictor variables and the response variable. In the area of QSAR modeling, chemical compounds are often treated as observations, molecular descriptors are treated as predictor variables, and the response variable is represented by biological activities such as IC_{50}. Typically, a good QSAR model should possess high predictability and be easily interpretive [10].

The trend today is toward more observations with an even larger number of variables. In chemometrics, there is an example where molecular descriptor generation tools have the capability of producing thousands of molecular descriptors. For instance, commercial software, known as DRAGON 6, can calculate 4885 molecular descriptors [10,11]. The data collected on individual compounds are molecular descriptors, so that a single compound has dimensions in thousands, while there are less than hundreds of compounds available for study. A problem of high dimensionality in QSAR modeling where the number of molecular descriptors, p, exceeds the number of compounds, n, is one of the new challenges facing researchers because conventional

* *Correspondence to: Muhammad Hisyam Lee, Department of Mathematical Sciences, Universiti Teknologi Malaysia, 81310 Skudai, Johor, Malaysia*
E-mail: mhl@utm.my

a *Z. Y. Algamal, M. H. Lee*
Department of Mathematical Sciences, Universiti Teknologi Malaysia, 81310 Skudai, Johor, Malaysia

b *A. M. Al-Fakih, M. Aziz*
Department of Chemistry, Universiti Teknologi Malaysia, 81310 Skudai, Johor, Malaysia

statistical methods, such as MLR, are neither applicable nor suitable [12]. The statistical issues associated with modeling high-dimensional QSAR using MLR include model overfitting and multicollinearity [13].

Molecular descriptor elimination for the highly correlated descriptors is often used by numerous researchers in chemometrics [1,14] in order to avoid multicollinearity, and, at the same time, to reduce dimensionality. However, it may not provide a satisfactory solution if the molecular descriptors dropped from the QSAR model have significant explanatory power relative to the biological activities; that is, omitting molecular descriptors to reduce multicollinearity may damage the predictive power of the QSAR model [15].

Dimensionality reduction and variable selection methods are an attractive way in high-dimensional QSAR studies. The aim of these methods is to select an optimal subset of those molecular descriptors that contain relevant information, and thereby to improve QSAR modeling. This should be observed in terms of predictive performance (by decreasing the effect of multicollinearity) and in interpretability (to prevent overfitting). Principal component analysis and partial least squares have gained attention in this area as dimension reduction methods [16]. They are used to alleviate the effect of multicollinearity and to prevent overfitting by reducing the dimension size. However, these methods lack the ability to interpret the results [17,18]. Traditional variable selection methods and classical model selection criteria, such as backward elimination, forward selection, stepwise selection, Akaike information criterion, and Bayesian information criterion, fail and, computationally, become more expensive in such high-dimensional problems [19,20].

Recently, an attractive framework for penalized regression methods has been adapted and gained popularity among statisticians as a key for performing variable selection and model estimation in high-dimensional data simultaneously. These methods impose a penalty term to be added to the residual sum of squares (RSS). The advantage behind the penalty term is to control the complexity of the model and provide criterion for variable selection, by shrinking the size of the coefficients toward zero. Some penalties simply alleviate the effect of multicollinearity, such as ℓ_2-norm penalty, which is used in ridge regression (RR) [21], while others try to prevent overfitting by reducing the dimension size, such as ℓ_1-norm penalty, which is used in least absolute shrinkage and selection operator (LASSO) [22]. The amount of the penalty term is the trade-off between the variance and bias of the selected model. A small amount leads to select more variables with little bias, but with high variance. Conversely, a large amount leads to select few variables with higher bias, but with less variance. Therefore, a good choice for the amount of the penalty term will improve the prediction accuracy and make an easily interpretable model.

Least absolute shrinkage and selection operator has shown success in many situations; however, it has three shortcomings. Firstly, it cannot select more variables than the number of observations. Secondly, when there is a group of correlated descriptors (grouping effects), LASSO randomly tries to select one or a few of the correlated descriptors [23–25]. Lastly, LASSO does not enjoy the oracle properties, which refer to the probability of selecting the right set of descriptors (with nonzero coefficients) converged to one, and that the estimators of the nonzero coefficients are asymptotically normal with the same means and covariances as if the zero coefficients are known in advance [26,27]. The main objective of this paper is to propose adjusting the adaptive LASSO (ALASSO) by improving the consistency selection and grouping effects. The proposed method has been carried out to establish a reliable QSAR between the IC_{50} values of imidazo[4,5-b]pyridine derivatives and the selected molecular descriptors.

2. MATERIALS AND METHODS

2.1. Dataset

The chemical structures and biological activities (IC_{50}) of 65 imidazo[4,5-b]pyridine derivatives, which are used as anticancer compounds, were selected from the literature [1,7,28,29] and shown in Figure S1 and Table S1 (Supporting information). The logarithmic scale of the IC_{50} values, $pIC_{50} = -\log(IC_{50})$, was used in QSAR modeling as a response variable. All the compounds were randomly split into two subsets, a training set of 45 compounds (70%), which was used to select the tuning parameters, and thereby to do variable selection and a test set of 20 compounds (30%), which was employed to evaluate the prediction ability of the QSAR model.

2.2. Molecular descriptors

The molecular structures of the 65 compounds were drawn using Chem3D software. The molecular structures were optimized using the molecular mechanics (MM2) method and then by a molecular orbital package (MOPAC) module in Chem3D software. Dragon software (version 6.0) was used to generate 4885 molecular descriptors including all 29 blocks based on the optimized molecular structures [11]. To include consistent and useful molecular descriptors, preprocessing steps were carried out as follows: First, those that have constant value for all compounds were excluded from the QSAR study. Then, molecular descriptors in which 60% of their values were zeros were removed. Last, those that have zero values for all compounds were discarded. In total, 2540 descriptors remained for evaluating the QSAR model.

2.3. High-dimensional QSAR model

In QSAR studies, the MLR has been commonly used to link the biological activities as a response variable to the molecular descriptors as predictor variables for data analysis. The resulting ordinary least squares (OLS) method has a closed form, which is easy to compute. However, OLS fails when the number of molecular descriptors, *p*, is greater than the number of compounds, *n*, because the design data matrix $\mathbf{X}$ has more columns than rows and has multicollinearity between molecular descriptors; therefore, $\mathbf{X}^T\mathbf{X}$ is singular [12,30]. Variable selection using penalized methods plays a vital role in statistical modeling with high-dimensional QSAR data. It aims to select only a subset of important descriptors from a large number of molecular descriptors, and thereby to improve the performance of QSAR models in terms of obtaining higher prediction accuracy of the model and easy interpretation. Penalized regression methods provide an estimate QSAR model that has lower prediction error than MLR using OLS, in situations where OLS can be applied.

The reduction in the prediction error using penalized regression, which is measured by mean-squared error, is achieved through a variance–bias trade-off: As the complexity of an MLR increases by including more molecular descriptors in the model, the variance increases and the bias simultaneously decreases. Including more molecular descriptors allows the QSAR model

to adapt to more complicated relationships in the data. However, a model with too many molecular descriptors may overfit the QSAR model. Such overfitting leads to a QSAR model that may not describe future compounds well. Depending on the type of penalty term used, penalized regression can alleviate the problems of multicollinearity and can also produce sparse QSAR models (with small numbers of molecular descriptors) that are consequently easier to interpret scientifically.

The RR proposed by Hoerl and Kennard [21] is one of the most used penalized methods as a remedy for the multicollinearity problem in statistics. RR shrinks the molecular descriptor coefficients toward zero by adding a ℓ_2-norm penalty to the RSS, but never equals zero. Hence, the variances of the molecular descriptor estimators are reduced, which leads to better properties in both estimation and prediction. However, the RR suffers from some limitations. Particularly in high-dimensional QSAR, it does not have the capability to perform variable selection and, therefore, does not give an easily interpretable model. LASSO, introduced by Tibshirani [22], is another frequently used penalized method. LASSO imposes the ℓ_1-norm penalty to the RSS. Because of the ℓ_1-norm property, LASSO can perform variable selection by assigning some molecular descriptor coefficients to zero. Despite the advantage of LASSO, it has some shortcomings. First, it cannot select more molecular descriptors than the number of compounds. Second, when there is a group of correlated descriptors, LASSO arbitrarily selects one or a few correlated descriptors [23–25]. Last, LASSO does not enjoy the oracle properties, which refer to the probability of selecting the right set of molecular descriptors (with nonzero coefficients) converged to one and that the estimators of the nonzero coefficients are asymptotically normal with the same means and covariances as if the zero coefficients are known in advance [26,27].

The classical MLR model for QSAR studies is given by

$$\mathbf{y} = \mathbf{X}\beta + \boldsymbol{\varepsilon} \tag{1}$$

where $\mathbf{y} = (y_1, \ldots, y_n)^T$ is a vector of size $n \times 1$ of the biological activities, $\mathbf{X} = (\mathbf{x_1}, \ldots, \mathbf{x_p})$ is a $n \times p$ design matrix of p molecular descriptors, $\beta = (\beta_1, \ldots, \beta_p)^T$ is a $p \times 1$ vector of unknown molecular descriptor coefficients, and $\boldsymbol{\varepsilon}$ is a vector of size $n \times 1$ of independent and identically distributed random variables with mean 0 and variance σ^2. The usual estimation procedure for the β is OLS by minimization of the RSS with respect to β:

$$\hat{\beta}_{OLS} = \arg\min_{\beta} (\mathbf{y} - \mathbf{X}\beta)^T (\mathbf{y} - \mathbf{X}\beta) \tag{2}$$

Then the OLS estimator is obtained by solving Equation (2) and is defined as

$$\hat{\beta}_{OLS} = \left(\mathbf{X^T X}\right)^{-1} \mathbf{X^T y} \tag{3}$$

Penalization methods for the MLR model, which is called penalized ordinary least squares, are based on penalized RSS, and the estimation of the vector β is obtained by minimizing penalized RSS:

$$\hat{\beta}_{\text{POLS}} = \arg\min_{\beta} \left\{ (\mathbf{y} - \mathbf{X}\beta)^T (\mathbf{y} - \mathbf{X}\beta) + \lambda J(\beta) \right\} \tag{4}$$

The penalty term $\lambda J(\beta)$ depends on the positive tuning parameter λ, which controls the trade-off between fitting the data to the model and the effect of the penalty. In other words, it controls the amount of shrinkage. For the $\lambda = 0$, we obtain the OLS solution. In contrast, for large values of λ, the influence of the penalty term on the coefficient estimates increases. Choosing the tuning parameter is an important part of the model fitting. If focusing on prediction, the tuning parameter should find the right balance between bias and variance to minimize prediction error. Without loss of generality, it is assumed that the molecular descriptors are standardized, $\sum_{i=1}^{n} x_{ij} = 0$ and $(n^{-1}) \sum_{i=1}^{n} x^2_{ij} = 1, \quad \forall j \in \{1, 2, \ldots, p\}$, and the $\mathbf{y}$ is centered, $\sum_{i=1}^{n} y_i = 0$. As a result, the intercept β_0 is not penalized.

Assuming the penalized ordinary least squares in Equation (4), the LASSO estimator of β is defined by

$$\hat{\beta}_{\text{LASSO}} = \arg\min_{\beta} \left\{ (\mathbf{y} - \mathbf{X}\beta)^T (\mathbf{y} - \mathbf{X}\beta) + \lambda \sum_{j=1}^{p} |\beta_j| \right\} \tag{5}$$

where $\lambda \geq 0$ is a tuning parameter. LASSO continuously shrinks the molecular descriptor coefficients toward 0 as λ increases, and some molecular descriptor coefficients are shrunk exactly to 0. Shrinkage often improves prediction accuracy and helps to select irrelevant molecular descriptors. LASSO can be efficiently solved by several methods, such as the least angle regression algorithm [31] and the coordinate descent algorithm [32].

2.4. Adaptive LASSO

According to the language of Fan and Li [27], a good penalty term estimator must satisfy three properties: unbiasedness, sparsity, and continuity. Unbiasedness means the resulting estimator has no over-penalization for large parameters to avoid unnecessary modeling biases. Sparsity is another property that an estimator enjoys. In other words, the resulting estimator automatically sets insignificant parameters to zero. Lastly, continuity is the third property, meaning that the resulting estimator is continuous in data in order to avoid instability in model prediction.

One of the main reasons for LASSO not to be consistent, that is, lacking the oracle property [27,33,34], is that it equally penalizes all the regression coefficients, which over-penalizes the irrelevant predictor variables leading it to be a biased estimator. To alleviate this drawback, Zou [26] proposed the ALASSO in which adaptive weights are used for penalizing different coefficients in the ℓ_1-norm penalty. The basic idea behind ALASSO is that by assigning a higher weight to the small coefficients and lower weight to the large coefficients, it is possible to reduce the selection bias; therefore, it can consistently select the model. Furthermore, the ALASSO solution is continuous from its definition, which enables it to enjoy oracle properties. The ALASSO estimator of β is defined by

$$\hat{\beta}_{ALASSO} = \arg\min_{\beta} \left\{ (\mathbf{y} - \mathbf{X}\beta)^T (\mathbf{y} - \mathbf{X}\beta) + \lambda \sum_{j=1}^{p} \mathbf{w}_j |\beta_j| \right\} \tag{6}$$

where $\mathbf{w} = (w_1, \ldots, w_p)^T$ is $p \times 1$ data-driven weight vector. It depends on $\sqrt{n}$ -consistent initial values of $\hat{\beta}$ and $\mathbf{w}_j = \left[abs\left(\hat{\beta}_j\right) \right]^{-\gamma}$, where γ is a positive constant and is usually set to equal 1.

2.5. Adjusted adaptive LASSO

Adaptive LASSO originally used OLS estimates as initial weight [26]. This is no longer valid in high-dimensional data. Several researchers used LASSO estimates as an alternative initial weight [19]. However, using the LASSO estimator in ALASSO when $p > n$ may not be preferable for three reasons. First, the LASSO estimator is inconsistent in itself. In other words, this initial weight is biased in selection variables. Second, it does not take into account the weights for all the variables in any implantation, which means that some variables will be selected and the others will be set to zero. Last, when there is a group of correlated variables, LASSO fails to select the grouped variables together.

To overcome these limitations, the ratio of the standard error of the RR estimator to the RR estimator was proposed as an initial weight in ALASSO. According to the nature of the ℓ_2-norm, the ridge penalty tries to force the estimated coefficients of highly correlated predictor variables to be close to each other. However, this property loses the capability of estimating coefficients of highly correlated predictor variables with different magnitudes, especially with different signs [24]. The advantage of using the standard error of the ridge estimator $s_{\hat{\beta}_{Ridge}}$ is to adjust ALASSO when using RR estimates as an initial value. Cule and De Iorio [35] proposed a procedure to calculate the $s_{\hat{\beta}_{Ridge}}$ depending on the principal component analysis.

Let $\hat{\beta}_{\text{Ridge}} = \left(\hat{\beta}_{1(\text{Ridge})}, \ldots, \hat{\beta}_{p(\text{Ridge})}\right)^T$ be the vector of RR estimate, $\mathbf{s}_{\hat{\beta}_{Ridge}} = (s_{1(\hat{\beta}_{Ridge})}, \ldots, s_{p(\hat{\beta}_{Ridge})})T$ be the vector of the standard error of the RR, and $\mathbf{w}_{Ratio} = (w_{1(ratio)}, \ldots, w_{p(ratio)})^T$ be the ratio weight vector where $w_j = \left[s_{j(\hat{\beta}_{Ridge})}/abs\left(\hat{\beta}_{j(Ridge)}\right)\right]^{-\gamma}$, $j = 1, 2, \ldots, p$. For simplicity, we set $\gamma = 1$. Then a coordinate descent method can be used to solve the adjusted ALASSO (AALASSO). The computation details are given in Algorithm 1.

Algorithm 1. The coordinate descent optimization method for the AALASSO

Step 1: Find $\mathbf{w}_{Ratio}$.

Step 2: Define $\mathbf{x}_j^{**} = \mathbf{x}_j/\mathbf{w}_{Ratio}$, $j = 1, 2, \ldots, p$.

Step 3: Solve the LASSO for all λ values,

$$\hat{\beta}^{**}_{\text{AALASSO}} = \arg\min_{\beta}\left\{(\mathbf{y} - \mathbf{X}\beta)^T(\mathbf{y} - \mathbf{X}\beta) + \lambda \sum_{j=1}^{p} \mathbf{w}_{j(Ratio)}|\beta_j|\right\}.$$

Step 4: Output $\hat{\beta}^*_{j(\text{AALASSO})} = \hat{\beta}_j{**}/\mathbf{w}_{Ratio}$.

2.6. Evaluation criteria

The four methods were evaluated and validated to test the predictive ability of high-dimensional QSAR study of anticancer potency of imidazo[4,5-b]pyridine derivatives. Depending on the training data, two statistical criteria were used: the mean-squared error of the training set (MSE_{train}) and the leave-one-out internal validation (Q^2_{int}) defined by

$$MSE_{train} = \frac{\sum_{i=1}^{n_{train}}\left(y_{i,train} - \hat{y}_{i,train}\right)^2}{n_{train}} \tag{7}$$

and

$$Q^2_{\text{int}} = 1 - \left[\frac{\sum_{i=1}^{n_{train}}\left(y_{i,train} - \hat{y}_{i,train}\right)^2}{\sum_{i=1}^{n_{train}}\left(y_{i,train} - \overline{y}\right)^2}\right] \tag{8}$$

respectively.

Furthermore, the test dataset was used to validate the four methods by computing three criteria: the mean-squared error of the testing set (MSE_{test}), the external validation (Q^2_{ext}), and Pearson correlation between the true pIC_{50} values and the predicted pIC_{50}. The higher the value of the Pearson correlation, the closer the fit of the predicted pIC_{50}. The two former criteria were defined by

$$MSE_{test} = \frac{\sum_{i=1}^{n_{test}}\left(y_{i,test} - \hat{y}_{i,test}\right)^2}{n_{test}} \tag{9}$$

and

$$Q^2_{ext} = 1 - \left[\frac{\sum_{i=1}^{n_{test}}\left(y_{i,test} - \hat{y}_{i,test}\right)^2}{\sum_{i=1}^{n_{test}}\left(y_{i,test} - \overline{y}_{train}\right)^2}\right] \tag{10}$$

respectively, where n_{train} and n_{test} represent the training and testing sample sizes, the $y_{i,train}$, $y_{i,test}$, $\hat{y}_{i,train}$, and $\hat{y}_{i,test}$ stand for the pIC_{50} values of the training set, testing set, and their corresponding predicted pIC_{50} values, while $\overline{y}$ and $\overline{y}_{train}$ represent the mean of the all pIC_{50} values and the mean of the training pIC_{50} values, respectively. For both the internal and external validation criteria, a validation value of greater than 0.5 indicates a good predictive model [36].

3. RESULTS AND DISCUSSION

In the high-dimensional QSAR prediction of anticancer potency of imidazo[4,5-b]pyridine derivatives, AALASSO was performed, and the results were compared with LASSO, $ALASSO_{lasso}$, and $ALASSO_{ridge}$ in terms of the selected molecular descriptors and the accuracy of the prediction.

3.1. Molecular descriptors selection

In this study, all 2540 molecular descriptors were given the chance in QSAR study. In order to select the most informatics descriptors, the training set was used to select the descriptors through finding the optimal value for the tuning parameter of each method. The K-fold cross-validation method was employed with $K = 5$ to find the optimal values of λ. Table I summarizes the tuning parameter values and the number of molecular descriptors selected by each of these four methods in the training set. The names of the selected molecular descriptors and their corresponding coefficient values of the AALASSO, LASSO, $ALASSO_{lasso}$,

Table I. Tuning parameter values and the selected molecular descriptor numbers for the four methods

	λ	No. of descriptors selected
AALASSO	0.073	26
LASSO	0.106	19
$ALASSO_{lasso}$	0.141	15
$ALASSO_{ridge}$	0.186	12

AALASSO, adjusted adaptive least absolute shrinkage and selection operator; LASSO, least absolute shrinkage and selection operator; ALASSO, adaptive least absolute shrinkage and selection operator.

and $ALASSO_{ridge}$ are shown in Figure 1. Among the 26 molecular descriptors selected by AALASSO, 16 descriptors, 12 descriptors, and 7 descriptors were also selected by LASSO, $ALASSO_{lasso}$, and $ALASSO_{ridge}$, respectively. The names and the description of all the selected molecular descriptors by the four penalized methods are given in Table S2.

As can be seen from Table I and Figure 1, AALASSO selected more molecular descriptors than the other three methods. Most of these selected descriptors were correlated. For example, the highest correlation among molecular descriptors was 0.997 between RFD and NNRS, which belong to the ring block. These two correlated descriptors were selected together by AALASSO, while none of them was selected by any of the other methods. Furthermore, AALASSO selected four correlated descriptors, MATS4v, MATS1e, MATS4p, and MATS1s, which belong to the 2D autocorrelations block. However, the MATS1e descriptor, which yielded medium correlation values with the rest, was not selected by LASSO, $ALASSO_{lasso}$, or $ALASSO_{ridge}$, respectively. Besides, $ALASSO_{lasso}$ only selected the RDF135m descriptor, and LASSO selected the RDF135m and RDF095m descriptors, while AALASSO selected four correlated descriptors from the RDF block, which are RDF095m, RDF110m, RDF135m, and RDF040i. Again, the AALASSO method succeeded in selecting more correlated descriptors from the 3D-MoRSE block. It selected Mor03m, Mor11s, Mor16s, and Mor31s, while the LASSO method only selected Mor11s. On the other hand, it can be observed that the $ALASSO_{ridge}$ failed to select highly correlated molecular descriptors, although it selected less molecular descriptors compared with the other methods, which may give easier interpretation. The success of AALASSO in selecting more correlated molecular descriptors than the other methods, especially $ALASSO_{ridge}$, is due to its ability to adjust the adaptive weight $w_j = \left[1/abs\left(\hat{\beta}_{j(Ridge)}\right)\right]$ by $s_{\hat{\beta}_{Ridge}}$.

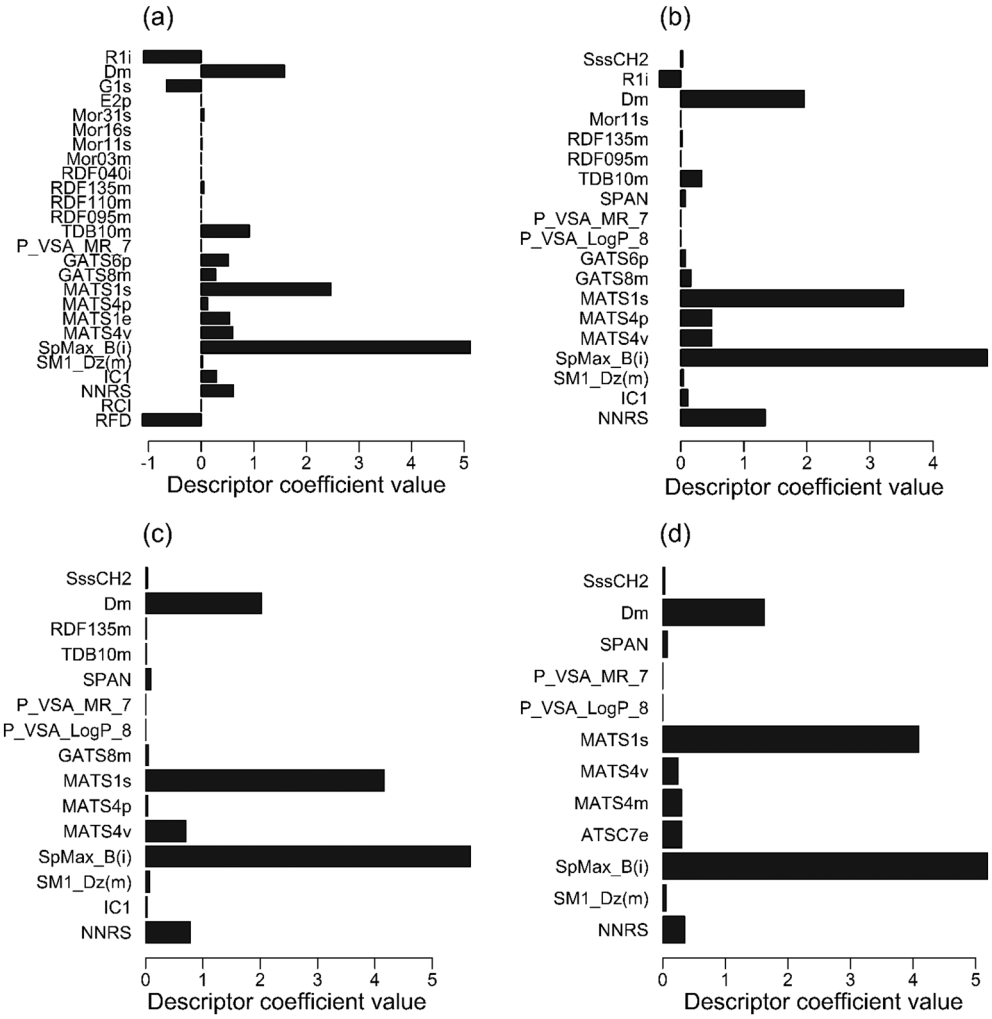

Figure 1. Selected molecular descriptor coefficients (a) AALASSO, (b) LASSO, (c) $ALASSO_{lasso}$, and (d) $ALASSO_{ridge}$.

Table II. The frequencies of the selected molecular descriptors obtained by the four methods over 25 times

Selected descriptors	AALASSO	LASSO	$ALASSO_{lasso}$	$ALASSO_{ridge}$
RFD	23	0	0	0
RCI	23	0	0	0
NNRS	25	24	25	23
IC1	24	23	22	0
SM1_Dz.m.	24	24	25	25
SpMax_B.i.	25	25	25	25
ATSC7e	0	0	0	15
MATS4m	0	0	0	20
MATS4v	25	22	23	22
MATS1e	23	0	0	0
MATS4p	24	21	22	0
MATS1s	25	22	21	22
GATS8m	25	22	23	0
GATS6p	24	21	0	0
P_VSA_LogP_8	0	10	12	14
P_VSA_MR_7	23	13	14	10
SPAN	0	20	22	22
TDB10m	25	21	22	0
RDF095m	24	21	0	0
RDF110m	24	0	0	0
RDF135m	25	25	23	0
RDF040i	23	0	0	0
Mor03m	23	0	0	0
Mor11s	25	23	0	0
Mor16s	24	0	0	0
Mor31s	23	0	0	0
E2p	24	0	0	0
G1s	25	0	0	0
Dm	25	25	25	25
R1i.	25	24	0	0
SssCH2	0	15	14	16

AALASSO, adjusted adaptive least absolute shrinkage and selection operator; LASSO, least absolute shrinkage and selection operator; ALASSO, adaptive least absolute shrinkage and selection operator.

Table III. Training and testing evaluation criteria values for the four methods

Methods	Training set		Testing set		Pearson Correlation
	MSE_{train}	Q^2_{int}	MSE_{test}	Q^2_{ext}	
AALASSO	0.075	0.942	0.150	0.867	0.940
LASSO	0.110	0.915	0.226	0.799	0.909
$ALASSO_{lasso}$	0.139	0.893	0.267	0.763	0.890
$ALASSO_{ridge}$	0.177	0.865	0.297	0.737	0.880

AALASSO, adjusted adaptive least absolute shrinkage and selection operator; LASSO, least absolute shrinkage and selection operator; ALASSO, adaptive least absolute shrinkage and selection operator.

To further evaluate the performance of the AALASSO ability in encouraging the selected group of correlated molecular descriptors, all compounds were divided 25 times into training and

Table IV. The true and predicted pIC_{50} values of the training and testing sets for the four methods

Molecule no.	True pIC_{50}	Predicted pIC_{50}			
		AALASSO	LASSO	$ALASSO_{lasso}$	$ALASSO_{ridge}$
57	8.000	7.480	7.380	7.280	7.250
24	7.796	7.530	7.503	7.496	7.465
56	7.854	7.750	7.766	7.758	7.718
32	6.558	6.730	6.795	6.862	6.927
14	4.796	4.950	5.008	5.136	5.293
29	6.939	6.938	6.926	6.937	6.940
26	7.284	7.225	7.181	7.127	7.133
51	7.921	7.918	7.833	7.821	7.783
21	7.301	7.282	7.291	7.306	7.289
15	4.538	5.056	5.171	5.249	5.372
2	6.244	5.875	5.863	5.826	5.795
17	5.357	5.447	5.426	5.439	5.462
1	5.367	5.133	5.157	5.211	5.299
53	7.523	7.722	7.731	7.720	7.681
52	8.699	8.286	8.251	8.202	8.119
23	6.287	6.698	6.811	6.883	6.930
11	5.620	5.864	5.877	5.872	5.884
7	4.745	5.322	5.503	5.635	5.750
25	7.092	7.114	7.145	7.150	7.175
34	6.588	6.646	6.785	6.874	6.976
13	5.180	4.928	5.033	5.133	5.238
45	7.495	7.427	7.346	7.319	7.349
64	8.523	8.348	8.191	8.065	7.914
4	6.131	5.967	5.929	5.970	6.044
50	8.222	8.246	8.191	8.093	7.990
44	7.092	7.567	7.611	7.568	7.561
46	8.398	7.828	7.677	7.597	7.542
8	5.046	5.444	5.579	5.610	5.638
18	6.060	6.542	6.64	6.687	6.747
19	7.377	7.293	7.296	7.267	7.220
65	8.000	7.921	7.863	7.797	7.700
27	8.523	8.005	7.852	7.763	7.655
36	7.071	7.175	7.219	7.258	7.271
20	7.260	7.248	7.217	7.229	7.235
37	7.102	6.946	6.937	6.980	6.986
48	7.824	7.666	7.614	7.570	7.534
10	5.155	5.240	5.309	5.401	5.508
47	7.678	7.437	7.418	7.387	7.344
5	5.161	5.347	5.313	5.329	5.401
30	6.928	7.077	7.148	7.179	7.167
61	8.000	8.027	8.059	8.017	7.944
39	8.046	7.707	7.627	7.573	7.536
38	7.260	7.584	7.518	7.486	7.442
9	6.602	6.608	6.550	6.473	6.380
31	7.125	7.198	7.224	7.226	7.178
3^{test}	6.310	5.895	6.054	6.080	6.049
6^{test}	4.699	5.691	5.785	5.851	5.887
12^{test}	5.000	5.145	5.127	5.220	5.370
16^{test}	5.337	5.908	6.226	6.330	6.362
22^{test}	7.276	6.991	7.086	7.147	7.222
28^{test}	6.438	6.549	6.757	6.851	6.921
33^{test}	6.750	6.618	6.500	6.508	6.570
35^{test}	6.678	6.846	6.971	7.005	7.057
40^{test}	8.097	7.914	7.825	7.731	7.664

(*Continues*)

Table IV. *(Continued)*

Molecule no.	True pIC_{50}	Predicted pIC_{50}			
		AALASSO	LASSO	$ALASSO_{lasso}$	$ALASSO_{ridge}$
41^{test}	7.398	7.779	7.686	7.620	7.531
42^{test}	6.187	7.094	7.461	7.538	7.594
43^{test}	7.770	7.685	7.684	7.670	7.667
49^{test}	8.301	8.044	7.999	7.941	7.876
54^{test}	8.222	7.916	7.943	7.907	7.846
55^{test}	8.000	8.226	8.307	8.294	8.194
58^{test}	7.260	7.434	7.274	7.146	7.083
59^{test}	7.523	7.401	7.315	7.209	7.198
60^{test}	7.538	7.608	7.456	7.292	7.188
62^{test}	7.824	8.105	8.129	8.114	8.024
63^{test}	8.000	7.867	8.017	8.015	7.903

AALASSO, adjusted adaptive least absolute shrinkage and selection operator; LASSO, least absolute shrinkage and selection operator; ALASSO, adaptive least absolute shrinkage and selection operator.
^{test}The molecule belongs to test set.

testing sets. The performance in terms of the frequency of the selected descriptors for each method is reported in Table II. It is seen that AALASSO gave consistent selection and succeeded in selecting the same correlated descriptors as it selected them originally with a percentage equal to 92%. The correctly selected descriptor percentages of LASSO, $ALASSO_{lasso}$, and $ALASSO_{ridge}$, on the other hand, were 40%, 48%, and 40%, respectively.

3.2. Evaluation of AALASSO

Training and testing dataset were used to measure the predictive accuracy of the AALASSO, and the results were compared with LASSO, $ALASSO_{lasso}$, and $ALASSO_{ridge}$. The results are reported in Table III. It can be seen that the MSE_{train} of the AALASSO was about 31.81%, 46.04%, and 57.62% lower than that of LASSO, $ALASSO_{lasso}$, and $ALASSO_{ridge}$, respectively. Moreover, the prediction performance in the training set using Q^2_{int} of the AALASSO was 0.942, which was much better than 0.915, 0.893, and 0.893 obtained by the LASSO, $ALASSO_{lasso}$, and $ALASSO_{ridge}$, respectively, which indicated the better predictive ability of the AALASSO than the other three methods.

Depending on the test set, AALASSO reduced the MSE_{test} significantly in comparison with the other three methods. The

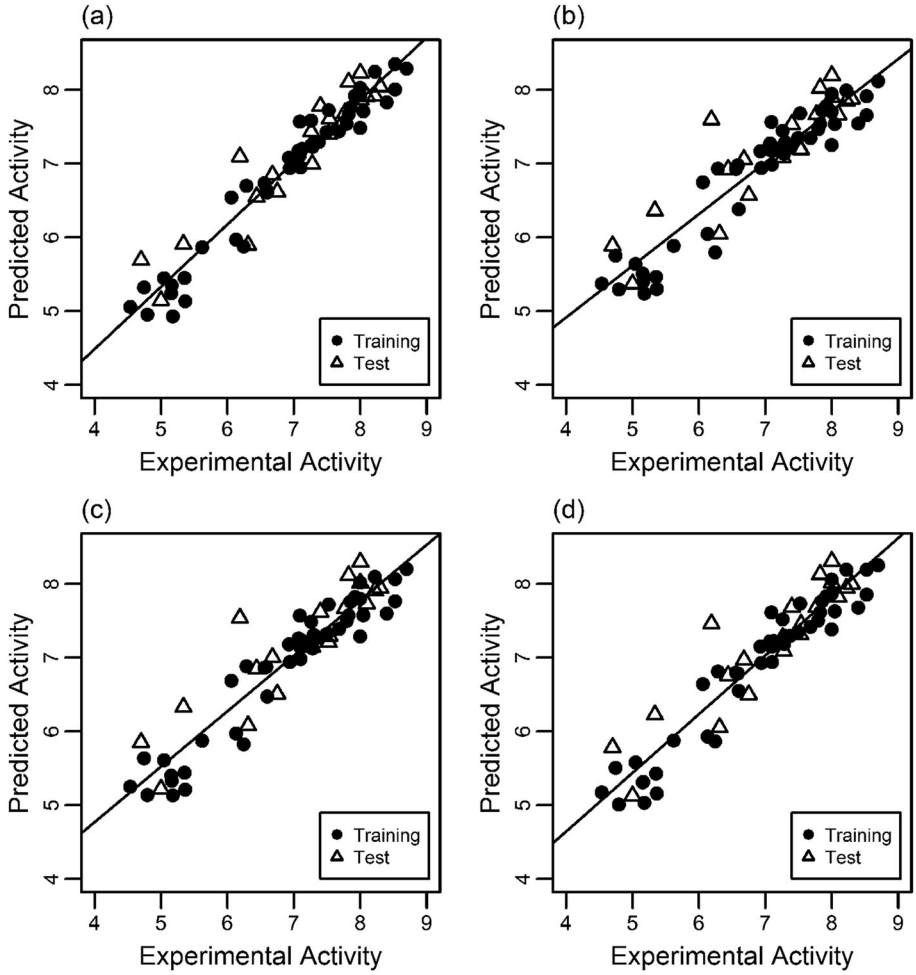

Figure 2. Plot of true versus predicted pIC_{50} values as obtained from the training and testing sets (a) AALASSO, (b) LASSO, (c) $ALASSO_{lasso}$, and (d) $ALASSO_{ridge}$.

reduction of MSE_{test} using AALASSO was 34%, 44%, and 50% compared with LASSO, $ALASSO_{lasso}$, and $ALASSO_{ridge}$, respectively. Furthermore, it is apparent that the Q^2_{ext} value for AALASSO was higher than the other three methods, which indicated that AALASSO has greater predictive ability than LASSO, $ALASSO_{lasso}$, and $ALASSO_{ridge}$. Moreover, the reliability of AALASSO was also assessed from its Pearson correlation value. It ranked the AALASSO above the LASSO, $ALASSO_{lasso}$, and $ALASSO_{ridge}$. On the other hand, $ALASSO_{ridge}$ generally performed slightly worse than the other three methods in terms of validation, although it did select less molecular descriptors.

All the predicted pIC50 results for both the training and testing data are listed in Table IV. Figure 2 displays the correlation between the true values of the pIC_{50} and the corresponding predicted values for both the training and the test sets. Figure 2(a) clearly reveals that the predicted pIC_{50} values were in good agreement in comparison with the true values.

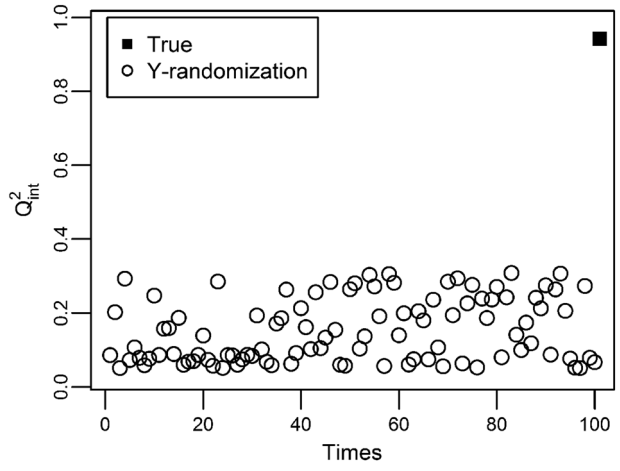

Figure 3. The Q^2_{int} values over 100 times of the Y-randomization test for AALASSO.

3.3. Y-randomization test

Adjusted ALASSO was further validated by applying the Y-randomization test [37]. This was in order to ensure that the predictive power of AALASSO was not based on chance. This test randomly shuffled the biological activity, pIC_{50}, several times and applied AALASSO each time. After that, the Q^2_{int} was calculated for each time. If all the obtained Q^2_{int} values were less than the Q^2_{int} of the constructed AALASSO, then the constructed AALASSO was not due to chance correlation indicating that the AALASSO method could lead to an acceptable method using the training

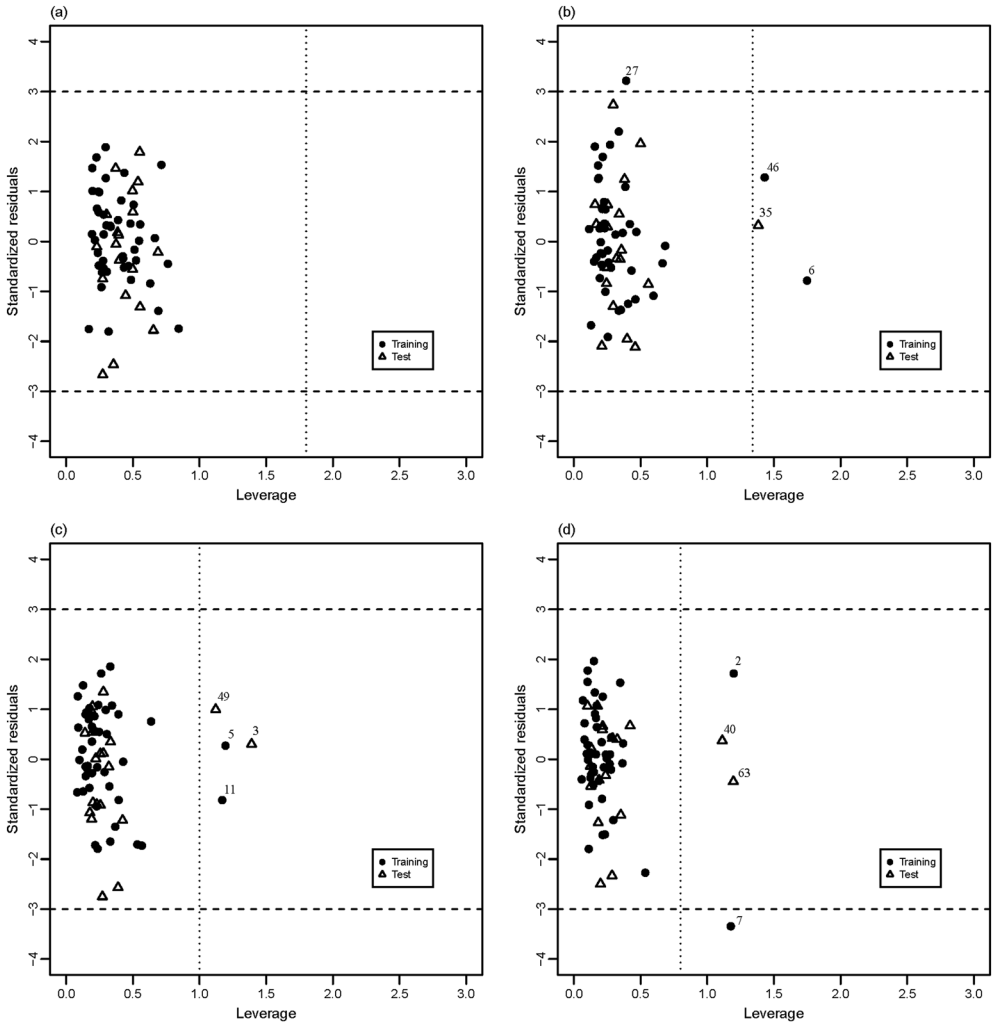

Figure 4. Williams plot of the AD for the training and testing sets (a) AALASSO, (b) LASSO, (c) $ALASSO_{lasso}$, and (d) $ALASSO_{ridge}$.

dataset. Figure 3 shows the results for the Y-randomization test for 100 times of Q^2_{int} values.

It can be clearly seen from Figure 3 that the Q^2_{int} values were in the range of 0.050 to 0.308. In comparison with true Q^2_{int} values of AALASSO ($Q^2_{int} = 0.942$), these values indicated that the high-dimensional QSAR prediction of anticancer potency of imidazo[4,5-b]pyridine derivatives by using proposed AALASSO were not due to chance correlation or structural dependence of the training dataset.

3.4. Applicability domain assessment

To further evaluate the ability of AALASSO in generating a reliable and robust QSAR model, an applicability domain (AD) assessment was used. According to Gramatica [38], AD is defined as AD is a theoretical region in chemical space, defined by the model descriptors and modeled response, and thus by the nature of the chemicals in the training set, as represented in each model by specific molecular descriptors.

Leverage approach can be used as an AD assessment [38,39].

Figure 4 (Williams plot) depicts the results of the leverage values against the standardized residuals for each compound for the AALASSO and the other three methods (the dotted line indicates the leverage threshold, while the dashed line represents the standardized residual limits). The influential compound can be detected when its leverage value is greater than a leverage threshold ($h^* = 3(p+1)/n$) [39], where p is the number of the selected descriptors in the final QSAR model and n represents the number of compounds. It is easily clear from Figure 4(a) that there are no compounds with a higher standardized residual than the limit ±3, which can be considered as biological activity outliers. Also, from Figure 4(b), it is obvious that only compound number 27 in the training set was identified as an outlier in the chemical activity, while compound numbers 46 and 6 of training set and compound number 35 of test set are considered influential compounds.

In the Williams plot (Figure 4(c)), we can observe that there are two compounds from the training set, 5 and 11, and from the testing set, 3 and 49, identified as influential chemical compounds. As demonstrated in Figure 4(d), only compound 7 from the training set was considered as a chemical activity outlier and identified as influential compound simultaneously. On the other hand, compound 2 and compounds 40 and 64, from the training and testing sets, respectively, were identified as influential chemical compounds. Thus, it is clearly demonstrated that all the results confirm that the constructed QSAR model by AALASSO is more reliable and robust compared with LASSO, $ALASSO_{lasso}$, and $ALASSO_{ridge}$.

To summarize, it is obvious from Tables II and III, and Figures 2–4 that AALASSO has superior results in terms of evaluation. In addition, it outperforms the other competitor methods in terms of both consistency selection and selection of a group of correlated molecular descriptors. It selected many correlated molecular descriptors. In comparison, LASSO, $ALASSO_{lasso}$, and $ALASSO_{ridge}$ were only able to pick a few correlated descriptors. Generally, our proposed method, AALASSO, achieved better performance, especially in the selection of molecular descriptors and can be successfully applied to the high-dimensional QSAR studies.

4. CONCLUSION

A proposed penalized method as a tool for molecular descriptors selection, AALASSO, was carried out to study the high-dimensional QSAR of a series of anticancer potency of imidazo[4,5-b]pyridine derivatives. The proposed method was tested on training and testing sets, and the results showed that AALASSO outperforms LASSO, $ALASSO_{lasso}$, and $ALASSO_{ridge}$ in terms of consistency selection and grouping effects. The potential advantage of AALASSO is its ability to consistently select more correlated molecular descriptors. To conclude, the prediction accuracy of the high-dimensional QSAR of the anticancer potency demonstrated the advantage of the AALASSO and could further be applied in other high-dimensional QSAR studies.

REFERENCES

1. Bababdani BM, Mousavi M. Gravitational search algorithm: a new feature selection method for QSAR study of anticancer potency of imidazo[4,5-b]pyridine derivatives. *Chemometr. Intell. Lab. Syst.* 2013; **122**: 1–11.
2. Ali I, Haque A, Saleem K, Hsieh MF. Curcumin-I Knoevenagel's condensates and their Schiff's bases as anticancer agents: synthesis, pharmacological and simulation studies. *Bioorg. Med. Chem.* 2013; **21**: 3808–3820.
3. Kamel MM, Megally A, Nadia Y. Synthesis of novel 1,2,4-triazoles, triazolothiadiazines and triazolothiadiazoles as potential anticancer agents. *Eur. J. Med. Chem.* 2014; **86**: 75–80.
4. Zhang S, Luo Y, He L-Q, Liu Z-J, Jiang A-Q, Yang Y-H, Zhu H-L. Synthesis, biological evaluation, and molecular docking studies of novel 1,3,4-oxadiazole derivatives possessing benzotriazole moiety as FAK inhibitors with anticancer activity. *Bioorg. Med. Chem.* 2013; **21**: 3723–3729.
5. Ma L-Y, Pang L-P, Wang B, Zhang M, Hu B, Xue D-Q, Shao K-P, Zhang B-L, Liu Y, Zhang E, Liu H-M. Design and synthesis of novel 1,2,3-triazole-pyrimidine hybrids as potential anticancer agents. *Eur. J. Med. Chem.* 2014; **86**: 368–380.
6. Bavetsias V, Crumpler S, Sun C, Avery S, Atrash B, Faisal A, Moore AS, Kosmopoulou M, Brown N, Sheldrake PW, Bush K, Henley A, Box G, Valenti M, de Haven Brandon A, Raynaud FI, Workman P, Eccles SA, Bayliss R, Linardopoulos S, Blagg J. Optimization of imidazo[4,5-b]pyridine-based kinase inhibitors: identification of a dual FLT3/Aurora kinase inhibitor as an orally bioavailable preclinical development candidate for the treatment of acute myeloid leukemia. *J. Med. Chem.* 2012; **55**: 8721–8734.
7. Lan P, Chen W-N, Chen W-M. Molecular modeling studies on imidazo[4,5-b]pyridine derivatives as Aurora A kinase inhibitors using 3D-QSAR and docking approaches. *Eur. J. Med. Chem.* 2011; **46**: 77–94.
8. Bavetsias V, Faisal A, Crumpler S, Brown N, Kosmopoulou M, Joshi A, Atrash B, Pérez-Fuertes Y, Schmitt JA, Boxall KJ, Burke R, Sun C, Avery S, Bush K, Henley A, Raynaud FI, Workman P, Bayliss R, Linardopoulos S, Blagg J. Aurora isoform selectivity: design and synthesis of imidazo[4,5-b]pyridine derivatives as highly selective inhibitors of Aurora-A kinase in cells. *J. Med. Chem.* 2013; **56**: 9122–9135.
9. Pourbasheer E, Aalizadeh R, Shokouhi TS, Ganjali MR, Norouzi P, Shadmanesh J. 2D and 3D quantitative structure–activity relationship study of hepatitis C Virus NS5B polymerase inhibitors by comparative molecular field analysis and comparative molecular similarity indices analysis methods. *J. Chem. Inf. Model.* 2014; **54**: 2902–2914.
10. Huang J, Fan X. Reliably assessing prediction reliability for high dimensional QSAR data. *Mol. Divers.* 2013; **17**: 63–73.
11. Talete. http://www.talete.mi.it/index.htm/ [10 Sep 2014].
12. Filzmoser P, Gschwandtner M, Todorov V. Review of sparse methods in regression and classification with application to chemometrics. *J. Chemom.* 2012; **26**: 42–51.
13. Al-Fakih AM, Aziz M, Abdallah HH, Algamal ZY, Lee MH, Maarof H. High dimensional QSAR study of mild steel corrosion inhibition in acidic medium by furan derivatives. *Int. J. Electrochem. Sci.* 2015; **10**: 3568–3583.
14. Xu J, Wang L, Wang L, Shen X, Xu W. QSPR study of Setschenow constants of organic compounds using MLR, ANN, and SVM analyses. *J. Comput. Chem.* 2011; **32**: 3241–3252.
15. Yan X, Su XG. *Linear Regression Analysis: Theory and Computing*, World Scientific: Singapore, 2009; 238–250.
16. Liu P, Long W. Current mathematical methods used in QSAR/QSPR studies. *Int. J. Mol. Sci.* 2009; **10**: 1978–1998.

17. Tran TN, Afanador NL, Buydens LMC, Blanchet L. Interpretation of variable importance in partial least squares with significance multivariate correlation (sMC). *Chemometr. Intell. Lab. Syst.* 2014; **138**: 153–160.
18. Zou H, Hastie T, Tibshirani R. Sparse principal component analysis. *J. Comput. Graph. Stat.* 2006; **15**: 265–286.
19. Bühlmann P, Van De Geer S. *Statistics for High-dimensional Data: Methods, Theory and Applications*, Springer: Heidelberg, Germany, 2011; 25–26.
20. Chen J, Chen Z. Extended BIC for small-n-large-P sparse GLM. *Stat. Sin.* 2012; **22**: 555–574.
21. Hoerl AE, Kennard RW. Ridge regression: biased estimation for nonorthogonal problems. *Technometrics* 1970; **12**: 55–67.
22. Tibshirani R. Regression shrinkage and selection via the LASSO. *J. R. Stat. Soc. Ser. B* 1996; **58**: 267–288.
23. ter Braak CJF. Regression by L1 regularization of smart contrasts and sums (ROSCAS) beats PLS and elastic net in latent variable model. *J. Chemom.* 2009; **23**: 217–228.
24. Wang S, Nan B, Rosset S, Zhu J. Random LASSO. *Ann. Appl. Stat.* 2011; **5**: 468–485.
25. Zou H, Hastie T. Regularization and variable selection via the elastic net. *J. R. Stat. Soc. Ser. B* 2005; **67**: 301–320.
26. Zou H. The adaptive LASSO and its oracle properties. *J. Am. Stat. Assoc.* 2006; **101**: 1418–1429.
27. Fan J, Li R. Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Am. Stat. Assoc.* 2001; **96**: 1348–1360.
28. Bavetsias V, Large JM, Sun C, Bouloc N, Kosmopoulou M, Matteucci M, Wilsher NE, Martins V, Reynisson J, Atrash B, Faisal A, Urban F, Valenti M, de Haven Brandon A, Box G, Raynaud FI, Workman P, Eccles SA, Bayliss R, Blagg J, Linardopoulos S, McDonald E. Imidazo [4,5-b]pyridine derivatives as inhibitors of Aurora kinases: lead optimization studies toward the identification of an orally bioavailable preclinical development candidate. *J. Med. Chem.* 2010; **53**: 5213–5228.
29. Bavetsias V, Sun C, Bouloc N, Reynisson J, Workman P, Linardopoulos S, McDonald E. Hit generation and exploration: Imidazo[4,5-b]pyridine derivatives as inhibitors of Aurora kinases. *Bioorg. Med. Chem. Lett.* 2007; **17**: 6567–6571.
30. Ross Kunz M, She Y. Multivariate calibration maintenance and transfer through robust fused LASSO. *J. Chemom.* 2013; **27**: 233–242.
31. Efron B, Hastie T, Johnstone I, Tibshirani R. Least angle regression. *Ann. Stat.* 2004; **32**: 407–499.
32. Friedman J, Hastie T, Tibshirani R. Regularization paths for generalized linear models via coordinate descent. *J. Stat. Softw.* 2010; **33**: 1–22.
33. Alhamzawi R, Yu K, Benoit DF. Bayesian adaptive LASSO quantile regression. *Stat. Model.* 2012; **12**: 279–297.
34. Benoit DF, Alhamzawi R, Yu K. Bayesian LASSO binary quantile regression. *Comput. Stat.* 2013; **28**: 2861–2873.
35. Cule E, De Iorio M. Ridge regression in prediction problems: automatic choice of the ridge parameter. *Genet. Epidemiol.* 2013; **37**: 704–714.
36. Zhu H, Guo W, Shen Z, Tang Q, Ji W, Jia L. QSAR models for degradation of organic pollutants in ozonation process under acidic condition. *Chemosphere* 2015; **119**: 65–71.
37. Rücker C, Rücker G, Meringer M. y-Randomization and its variants in QSPR/QSAR. *J. Chem. Inf. Model.* 2007; **47**: 2345–2357.
38. Gramatica P. Principles of QSAR models validation: internal and external. *QSAR Comb. Sci.* 2007; **26**: 694–701.
39. Minovski N, Zuperl S, Drgan V, Novic M. Assessment of applicability domain for multivariate counter-propagation artificial neural network predictive models by minimum Euclidean distance space analysis: a case study. *Anal. Chim. Acta* 2013; **759**: 28–42.

SUPPORTING INFORMATION

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

