Predicting the biological activities of triazole derivatives as SGLT2 inhibitors using multilayer perceptron neural network, support vector machine, and projection pursuit regression models

Jintao Yuan ^{a,c}, Shuling Yu ^{b,*}, Shufang Gao ^b, Ying Gan ^b, Yi Zhang ^b, Ting Zhang ^c,
Yali Wang ^b, Liu Yang ^{b,d}, Jiahua Shi ^{b,*}, Wu Yao ^{a,*}

^a School of Public Health, Zhengzhou University, Zhengzhou, 450001, China

^b Key Laboratory of Natural Medicine and Immune-Engineering of Henan Province, Henan University, Kaifeng, 475004, China

^c Key Laboratory of Environmental Medicine Engineering, Ministry of Education, School of Public Health, Southeast University, Nanjing, 210009, China

^d Pharmaceutical College, Henan University, Kaifeng, 475004, China

* Corresponding author: Wu Yao

School of Public Health, Zhengzhou University, 100 Science Avenue, Zhengzhou, 450001, China. E-mail address: yaowu@zzu.edu.cn (W. Yao).

* Corresponding author: Shuling Yu and Jiahua Shi

Key Laboratory of Natural Medicine and Immune-Engineering of Henan Province, Henan University, Kaifeng, 475004, China. Fax: +86 371 22822134; Tel: +86 371 22822134. E-mail address: yushuling@henu.edu.cn (S.L. Yu), sjiahua@henu.edu.cn (J.H. Shi).

ABSTRACT

Quantitative structure-activity relationship (QSAR) studies were performed in this work to predict the pIC_{50} of non-glycoside sodium-dependent glucose cotransporter-2 (SGLT2) inhibitors (46 triazole derivatives). Four descriptors were selected from the pool of DRAGON descriptors using the enhanced replacement method. Three nonlinear regression methods—multilayer perceptron neural network (MLP NN), support vector machine (SVM), and projection pursuit regression (PPR)—were then used to build the QSAR models. The performance of the obtained models was assessed through the cross-validation and external validation of the test set. PPR produced a better model than MLP NN and SVM with coefficients of determination of 0.962 and 0.871 and root mean square errors of 0.162 and 0.471 for the training and test sets, respectively. This study developed simple and efficient approaches to predict the pIC_{50} of triazole derivatives and provided some insights into their structural features for the development of more potential SGLT2 inhibitors.

Keywords: Quantitative structure-activity relationship; Sodium-dependent glucose cotransporters-2; Multilayer perceptron neural network; Support vector machine; Projection pursuit regression

1. Introduction

Sodium glucose cotransporter 2 (SGLT2) is primarily expressed in the S1 segment of renal proximal tubules and accounts for 90% of renal glucose reabsorption. The inhibition of SGLT2 prevents renal glucose reuptake from the glomerular filtrate and subsequently lowers the blood glucose level. Thus, SGLT2 inhibition is a novel and promising treatment for diabetes. Most of the currently reported SGLT2 inhibitors are glucoside derivatives and are based on phlorizin (i.e., a naturally-occurring O-aryl glycoside) [1]. Some quantitative structure-activity relationship (QSAR) models have been developed for glycoside molecules [2–5]. However, some SGLT2 inhibitors (dapagliflozin, canagliflozin, and empagliflozin are glucosides derived from the prototype phlorizin [6,7]) in clinical development inhibit only 30–50% of glucose reabsorption [8]. The triazole ring can be considered a bioisostere of the amide group because these moieties have a similar H-bond acceptor capacity, a similar distance between substituents ($3.8 \sim 3.9 \text{ \AA}$ in amides and $5.0 \sim 5.1 \text{ \AA}$ in triazoles), and a similar dipolar character (amide 4.0 Debye; triazole 5.0 Debye) [9]. Thus, investigating non-glycoside triazole molecules as SGLT2 inhibitor can provide useful information on synthesizing novel potent SGLT2 inhibitors. However, only a few works focus on non-glycoside molecules [10–12]. In the literature, only one study reports comparative molecular field analysis (CoMFA) and comparative molecular similarity indices analysis (CoMSIA) research on 46 triazole derivatives based on Align Database-Rigid body alignment [12]. CoMFA and CoMSIA are sensitive to molecular alignment, and the conformation obtained by the Align Database-Rigid

body alignment may be far from ‘real’ bioactive conformation. Therefore, an alternative QSAR study on 46 triazole derivatives is conducted in the present study.

DRAGON descriptors are widely used to develop numerous QSAR models, ranging from predicting physico-chemical properties to pharmacological and toxicological activities [13–16]. DRAGON descriptors coupled with a multilayer perceptron neural network (MLP NN), support vector machine (SVM), and projection pursuit regression (PPR) have gained much attention and extensive application [17–19]. In this study, MLP NN, SVM, and PPR methods are used to develop models based on selected DRAGON descriptors. The success of QSAR models depends on the molecular descriptors selected to characterize the chemical structure, aside from the appropriate statistical methods. The enhanced replacement method (ERM) as a screening tool for descriptor variables is simple, has low computational cost, and outperforms or equals the performance of the genetic algorithm [20]. Thus, ERM is employed here to select DRAGON descriptors as the independent variables of the MLP NN, SVM, and PPR models. The MLP NN, SVM, and PPR methods tend to achieve accurate predictions, but where obtained models are too complex to be easily understood. To increase interpretability and visualization of “black box” models, sensitivity analysis was employed [21, 22]. Specially, few investigation on PPR combined with sensitivity analysis have been reported.

In this investigation, the ERM method is utilized to reduce the number of descriptors and select the best descriptor subset. Three nonlinear QSAR models are built to predict the pIC_{50} of triazole derivatives with the same descriptors, using MLP

NN, SVM and PPR methods. The presented QSAR models are validated under strict conditions, and the performances of the different models are also compared. The relative importance of each descriptor in the developed models was studied by using sensitivity analysis. The descriptors used in the QSAR models are used to identify the molecular characteristics related to the pIC_{50} of triazole derivatives.

2. Materials and Methods

2.1. Data set

The molecular structures and activities of 46 triazole derivatives were derived from Refs. [10–12]. The IC_{50} values were converted into the corresponding pIC_{50} ($-\log IC_{50}$). The pIC_{50} values are presented in Table 1, and the 2D structures and SMILES strings of the 46 triazole derivatives are shown in Figure 1 and Table S1 (in the Supplementary data), respectively. The modeled pIC_{50} ranged from 5.676 to 8.523. The compounds were divided into training ($n = 39$) and test ($n = 7$) sets, and the respective distribution of the sets was identical to that in Ref. [12] to facilitate the comparison of results. The training set was used to build models, and the independent test set was used to evaluate the predictive ability of the models.

(Figure 1 here)

(Table 1 here)

2.2. DRAGON descriptors

The three-dimensional structures of triazole derivatives were constructed using Sybyl-x 2.1. The energy of each molecule was minimized with the use of the gradient descent method, Tripos force field, and Gasteiger–Huckel charges. The molecular descriptors were generated using the software DRAGON [23], and 4885 different types of molecular descriptors were calculated to describe the structural diversity of the chemicals. The initial pool of descriptors was then reduced by applying the DRAGON built-in variable exclusion procedure. Three types of descriptors were

excluded: constant (i.e., relative standard deviation < 0.001), near-constant (i.e., all the values are equal, except for one), and highly correlated descriptors. For each pair of highly correlated descriptors ($R > 0.9$), the descriptor with the largest mean correlation coefficient among the rest of the descriptors was removed. Finally, the remained descriptors were selected by ERM.

2.3. Variable selection

We used ERM to select the best subset of DRAGON descriptors. The ERM can determine the optimal subset of d ($d \ll D$) descriptors from a large group of D descriptors with a minimum standard deviation (S) [24]:

$$S = \frac{1}{(N-d-1)} \sum_{i=1}^N res_i^2 \quad (1)$$

where N is the number of molecules in the training set, and res_i represents the residual of molecule i . The residual is the difference between the experimental and predicted properties. $S(\mathbf{d}_n)$ is a distribution in a space of $D!/([d!](D-d)!)$ points. The quality of the results achieved with ERM is close to that of an exactly full search of molecular descriptors. More details can be found in Ref. [24,25]. The selection of the ERM variable was implemented in MATLAB.

2.4. MLP NN

The MLP NN used here has a three-layered architecture. The three layers (topology of NN) are the input, hidden, and output layers. The three-layer topology with a single hidden layer is sufficient to solve similar or more complex problems

[26]. The number of the input and output layers is determined by the number of selected descriptors and predicted variables, respectively. The required number of hidden layers is generally determined through testing. The number of neurons (2–10) was tested to determine the optimal number of neurons in the hidden layer. The activation functions of both the hidden and output nodes are as follows: identity, logistic sigmoid, hyperbolic tangent, and negative exponential. The suitable activation functions are determined for the hidden and output nodes through testing. The training process is iterative, and each iteration is called an epoch. During this process, the connection weights are constantly adjusted using the propagation error. Further discussions on the MLP NN can be found in Ref. [27, 28].

2.5. SVM

SVM is one of the most recent developments in machine learning modeling. This method is based on the theory of risk minimization and is able to find an optimal separation hyperplane in a multi-dimensional space to perform classification or regression tasks. The performance of SVMs in classification and regression depends on several factors, including the kernel function type and its corresponding parameters, the capacity parameter (C) and insensitive loss function (ϵ). First, the kernel function should be identified to determine the sample distribution in the mapping space. The radial basis function (RBF) is commonly used in many studies because of its good general performance, and only few parameters need to be adjusted [29]. Here, RBF is also employed. The γ of the kernel function greatly affects the number of support

vectors (SVs), which has a close relation with the performance of the SVM and training time. Too many support vectors can result in overfitting and increase the training time. In addition, γ controls the amplitude of the RBF function and therefore controls the generalization ability of SVM. The parameter C controls the trade-off between maximizing the margin and minimizing the training error. The ε -insensitive parameter prevents the entire training set from meeting the boundary conditions and therefore allows the possibility of sparsity in the dual formulation's solution. The optimal value of ε depends on the type of noise present in the data, which is usually unknown. The optimization of these parameters is based on the grid search technique, which is in turn based on 10-fold cross-validation [30].

2.6. PPR

PPR is a nonlinear multivariate regression procedure, which was introduced by Friedman and Stuetzle [31]. It combines both ideas of projection and pursuit. PPR is a powerful tool for exploring the projections of high-dimensional data into low-dimensional space and therefore for overcoming the problem of dimensionality. To data, it has been applied successfully in some fields, such as water supply management [32], pesticide [19], drug discovery [33], analytical chemistry [34], economics [35]. Here, only a brief description is given. Let X be a $(n \times k)$ data matrix, where k is the number of explanatory variables (descriptors) and n is the number of objects under investigation. The projection process can be defined as $z_m = X\alpha_m$, where α_m is the m th projection vector (length k) and z_m is the vector of the scores

after the projection of X on α_m (length n). The two basic factors of PP are: a PP index and a PP algorithm. A PP index, $I(\alpha)$, is a measure of how interesting a projection by matrix α is, where $I(\alpha) = I(\alpha/X)$ implicitly depends on the data $X = (X_1, \dots, X_n)$. The larger the index value is, the more interesting the projection is. A PP algorithm is a numerical optimization algorithm which maximizes the index over all possible α . For a regression problem, PPR aims to approximate the regression pursuit function $f(x)$ through a finite sum of ridge functions with suitable choices of β_i and g_i :

$$f^{(p)}(x) = \sum_{i=1}^p g_i(\beta_i^T x) \quad (2)$$

where the values of β_i are $m \times n$ orthonormal matrices and p is the number of ridge functions. In this algorithm, g_i are found by smoothing operation that entails a backfitting. Specially, given $g^{(0)} = 0$, for $i \geq 1$, it iteratively estimates β_i by maximum of an index and g_i by a low dimensional nonparametric regression estimate based on the projected data (z_j, r_j) , where $r_j = Y_j - g^{(i-1)}(X_j)$ are the residuals at the i th step and $z_i = \beta_i^T X_j, j = 1, \dots, n$. The procedure is repeated forward (and perhaps a backward fitting is allowed to adjust for the previous fitted pair) until the residual sum of squares $\sum r_j^2$ is less than a predetermined values. The basic theory of PPR can be found in Ref. [28, 34, 36, 38,]. PPR is executed in the R environment.

2.7 Model evaluation

Diagnostic statistical tools, namely, goodness of fit and prediction, were used to evaluate the performance of the models. The goodness of fit was examined by the

coefficient of determination (r^2) and root mean square error (RMSE). The goodness of the prediction of models was evaluated to predict the seven compounds in the test set. The presented models were evaluated for robustness using leave-many-out cross validation (LMO CV). The mean of ten readings was given as $q^2_{\text{LMO-5}}$. The mean RMSE of the readings of CV was RMSECVm.

Additionally, several parameters have been proposed in the literature [39] for the external validation of QSAR models which were used to further test the predictive ability of the developed QSAR models.

$$Q_{F1}^2 = 1 - \frac{\sum_{i=1}^{n_{EXT}} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n_{EXT}} (y_i - \bar{y}_{TR})^2} \quad (3a)$$

$$Q_{F2}^2 = 1 - \frac{\sum_{i=1}^{n_{EXT}} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n_{EXT}} (y_i - \bar{y}_{EXT})^2} \quad (3b)$$

$$Q_{F3}^2 = 1 - \frac{\left[\sum_{i=1}^{n_{EXT}} (y_i - \hat{y}_i)^2 \right] / n_{EXT}}{\left[\sum_{i=1}^{n_{TR}} (y_i - \bar{y}_{TR})^2 \right] / n_{TR}} \quad (3c)$$

$$CCC = \frac{2 \sum_{i=1}^{n_{EXT}} (y_i - \bar{y})(\hat{y}_i - \bar{\hat{y}})}{\sum_{i=1}^{n_{EXT}} (y_i - \bar{y})^2 + \sum_{i=1}^{n_{EXT}} (\hat{y}_i - \bar{\hat{y}})^2 + n_{EXT} (\bar{y} - \bar{\hat{y}})^2} \quad (3d)$$

where y_i is the experimental data values, $\hat{y}_i$ is the predictive value, $\bar{y}_{TR}$ is the mean value in the training set, the symbol TR and EXT are used for the training and test sets, respectively, $\bar{y}$ is the average of the experimental data values, $\bar{\hat{y}}$ is the average of the predicted data values, and n is the number of compounds.

3. Results and Discussion

3.1. Results of ERM

After performing the DRAGON built-in variable exclusion procedure, the pool of descriptors was reduced to 864. A variety of subset sizes were investigated to determine the optimum number of descriptors. If the addition of a descriptor no longer improved the statistical power of a model from its previous level, then the optimum subset size was considered to have been achieved. The influences of the number of descriptors on the r^2 for the training set are shown in Fig. 2. When four descriptors were increased to five descriptors, the increase in r^2 was less than 0.02, and the decrease in RMSE was less than 0.03. To avoid the “overparametrization” of the model, a smaller number of descriptors should be used to represent the chemicals. Thus, the four descriptors constituted the best selection. These selected descriptors and their correlation matrix are summarized in Table 2. The four descriptor variables evidently had a low correlation with one another. Therefore, the four descriptors were used to build the MLP NN, SVM and PPR nonlinear models.

(Figure 2 here)

(Table 2 here)

3.2. MLP NN model

A three-layer MLP NN model was developed using the selected descriptors as input. A 4–8–1 network architecture was obtained after testing different groups of neurons. In the best network, the hidden and output functions were both exponential

functions, with $r^2 = 0.930$, $RMSE = 0.219$, $q^2_{LMO-5} = 0.890$, and $RMSECV_m = 0.275$ for the training set. The results for the test set were satisfactory, with $r^2 = 0.825$ and $RMSE = 0.451$. The statistical parameters recommended by Tropsha et al. were used to check the proposed MLP NN model and are summarized in Table 3. Q^2_{F1} , Q^2_{F2} , and Q^2_{F3} were all higher than 0.6 (threshold value), and CCC was higher than 0.85 (threshold value). The obtained parameters obviously satisfied the generally accepted condition. The predictive pIC_{50} values from the MLP NN model for the entire dataset are listed in Table 1, and the experimental versus predicted values are shown in Fig. 3A. Clearly, all the points were rather uniformly distributed around a diagonal line, thereby suggesting the satisfactory predictive capability of the model.

(Figure 3 here)

(Table 3 here)

3.3. SVM model

SVM was also applied to build a non-linear model with the four descriptors. In this investigation, a searching program named “grid regression” was first adopted. It could search for the best parameters C , γ , and ε , while preventing the over-fitting of the training set. A grid regression was conducted in the training step. To find the optimized parameter combination, the parameter C was found with an increment of 1 within 1–50, γ with an increment of 0.05 within 0.05–0.3, and ε with an increment of 0.001 within 0.01–0.1. Finally, C , γ , and ε were fixed to 7, 0.2, and 0.024, respectively. In this model, the r^2 for the training set was 0.922, and the RMSE was

0.234. The $q^2_{\text{LMO-5}}$ for the LMO CV was 0.867, and the RMSECV_m was 0.314. For the test set, r^2 was increased to 0.846, and RMSE was 0.461, which indicated the good generalization ability of the SVM model. This model satisfied the conditions recommended by Tropsha (Table 3). Satisfactory results were obtained with this model. The results are shown in Table 1 and Fig. 3B.

3.4. PPR model

PPR was also applied to build a non-linear model with the same four descriptors as input variables. In investigating PPR, the parameters “nterms,” “max. terms,” “optlevel,” and “span,” must be determined. The parameter “nterms” controls the number of variables to be entered in the model, and “max. terms” is the maximum number of terms to be tried in building the model. The parameter “optlevel” indicates the levels of optimization that differ in terms of the degree at which the models are refitted during this process. At level 0, the existing ridge terms are not refitted. At level 1, the projection directions are not refitted, but the ridge functions and the regression coefficients are both refitted. At levels 2 and 3, all the terms are refitted and are equivalent for one response. Level 3 is more careful in re-balancing the contributions from each regressor at each step, so that the convergence to a saddle point of the sum of squares criterion is less likely. The parameter “span” defines the fraction of the observations in the span of running the lines smoothly. In this investigation, “nterms,” “max. terms,” “optlevel,” and “span” were determined to be 2, 3, 2, and 0, respectively. The model yielded the following values for the training set:

$r^2 = 0.962$, $RMSE = 0.162$, $q^2_{LMO-5} = 0.821$, and $RMSECVm = 0.351$. The proposed PPR model yielded an r^2 of 0.871 and an RMSE of 0.471 for the test set. Q^2_{F1} , Q^2_{F2} , Q^2_{F3} , and CCC were all higher than the threshold value (Table 3). Therefore, the PPR model was used to predict the pIC_{50} values, and the predicted results are shown in Table 1 and Fig. 3C. The figure shows that the plots exhibit a relatively uniform scattering around the diagonal line, but no obvious outliers can be found.

3.4. Comparison of the results obtained by different approaches

To check the qualities of the three different methods (MLP NN, SVM, and PPR), we compared the three models presented above with those obtained in literature [12]. Table 4 shows the statistical parameters of the obtained models from the same set of compounds. The values of the r^2 of the SVM and PPR models in this study for the training and test sets were all better than those reported for the CoMSIA model. The value of r^2 of the MLP NN model for the training set was better than that reported for the CoMSIA model. The values of the r^2 of the PPR model for the training and test sets were very close to those reported for the CoMFA model. The comparative results of these three methods indicated that PPR was the most accurate approach.

(Table 4 here)

3.4. Sensitivity analysis

The “black-box” problem of the nonlinear method, in which the exact relationship between the input and output data is difficult to determine, was resolved

using sensitivity analysis [22]. Here, the sequential zeroing of weights as sensitivity analysis method was performed to rank the order of importance of the descriptors in the proposed models [40]. The differences (Rmdiffi) between the RMSE of the prediction of the complete model and the actual RMSE were obtained when the *i*th variable was excluded from the trained model (RMSE_i). The differences of these RMSE (Rmdiffi) were calculated as follows: $Rmdiffi = RMSE_i - RMSE$. The most important variable was the variable that resulted in the highest value of Rmdiffi. The values of Rmdiffi for the MLP NN, SVM, and PPR models were calculated and are plotted in Figs. 4A, C, and E, respectively. The relative contributions of the four descriptors in MLP NN, SVM, and PPR are shown in Figs. 4B, D, and F, respectively. The importance of the four descriptors in MLP NN, SVM, and PPR evidently followed the same order: $SpMax_AEA(dm) > GATS6i > CATS2D_02_AL > L2p$. As the pie charts show, the minimum relative contribution of a descriptor was 17%, which suggested that each descriptor was important for the MPL NN, SVM, and PPR models to predict pIC_{50} .

(Figure 4 here)

3.5. Descriptor analysis and interpretation

The developed QSAR models should not only have a reliable prediction capability but also provide insight into the factors that are likely to influence the pIC_{50} of triazole derivatives by interpreting the meaning of the selected descriptors.

$SpMax_AEA(dm)$ is the most important descriptor in the three models based on

the above sensitivity analysis. SpMax_AEA(dm) is an edge adjacency index, which means that it leads the eigenvalue from the augmented edge adjacency (AEA) matrix weighted by the dipole moment. SpMax is the largest eigenvalue of the matrix spectrum. This descriptor is weighted by the dipole moment. The dipole moment is related to the polarity of a molecule [41]. The emergence of the dipole moment is due to the asymmetry in the 3D structure and charge distribution of the molecule, which is consistent with the findings in the literature [12]. The molecular volume, shape, and hydrophobicity of the substituent group affect the pIC₅₀ of triazole derivatives [12].

The Geary autocorrelation–lag 6/weighted by the ionization potential (GATS6i) descriptor is a 2D autocorrelation molecular descriptor that deals with the distribution of atomic ionization potentials along a molecular topological structure. It is calculated from an H included molecular graph and may be calculated as follows:

$$GATS6i = \frac{(1/2\Delta) \sum_{i=1}^n \sum_{j=1}^n \delta_{ij} (i_i - i_j)^2}{(1/n-1) \sum_{i=1}^n (i_i - \bar{i})^2} \quad (4)$$

where i_i is the atomic ionization potential, $\bar{i}$ is the average value of the atomic ionization potentials in the molecule, n is the number of atoms, and δ_{ij} is the Kronecker delta (i.e., equal to 1 if $\delta_{ij} = 6$; otherwise, 0, with δ_{ij} being the topological distance between the two considered atoms). Δ is the sum of the Kronecker deltas, that is, the number of atom pairs at a distance equal to 6. [42, 43]

Its value is directly proportional to the number of atoms and inversely proportional to the number of atom pairs at a distance equal to 6. As the number of types of hetero atoms in a molecule increases, the deviation from the average atomic

ionization potential also increases, and the descriptor value in turn decreases.

CATS2D_02_AL belongs to the category of chemically advanced template search-2D (CATS2D) descriptors, which can be considered to be pharmacophoric fingerprints that are based on topological distance constraints [41]. CATS2D_02_AL defines the same pharmacophore points: hydrogen-bond acceptor (A) and lipophilic (L) at topological distances of 2 bonds between the points. This descriptor is probably related to existing nitrogen atoms in the rings (the hydrogen bond-acceptor point) and hydrophobic side chains. This analysis is consistent with the finding in the literature [12] that the best CoMSIA model was developed using steric, electrostatic, hydrophobic, and hydrogen-bond acceptor fields.

L2p belongs to the category of weighted holistic invariant molecular (WHIM) descriptors and corresponds to the second component size directional WHIM indexed/weighted by atomic polarizabilities. The WHIM descriptors contain information on the size, shape, symmetry, and atom distribution of the whole 3D molecular structure. They are calculated from the 3D coordinates of a molecule within different weighting schemes. The descriptor L2p is calculated from the atomic polarizability weighting scheme. The atomic polarizability of carbon is 1.760. However, the atomic polarizabilities of fluorine, chlorine, bromine, nitrogen, sulfur, and oxygen are 0.557, 2.180, 3.050, 1.100, 2.900, and 0.802, respectively. [42] When these hetero atoms are included in the triazole derivatives, the polarizabilities of the molecules change.

4. Conclusion

In this work, MLP NN, SVM, and PPR nonlinear models were developed to predict the pIC_{50} of selected triazole derivatives. The PPR model produced more satisfactory results than the MLP NN and SVM models and exhibited good predictive ability. The sensitivity analysis revealed that the four selected descriptors had roles in the activity of these compounds. The selected descriptors provided insight into the structural features related to the pIC_{50} of triazole derivatives. Thus, the three simple and effective methods not only accurately predicted the pIC_{50} of triazole derivatives but also provided useful information for designing new triazole derivatives as SGLT2 inhibitors.

Acknowledgements

This work was supported by the China Postdoctoral Science Foundation (Grant No. 2014M550387), Postdoctoral Research Sponsorship in Henan Province (Grant No. 2013014), the Joint Fund of National Natural Science Foundation of China and Henan Province (U1404508), Scientific and Technological Key Project Education Department of Henan Province (Grant No. 14B350005), National Science Foundation for Young Scientists of China (Grant No. 81302461), and Scientific Research Foundation of Henan University (Grant No. 2013YBZR039).

Conflict of interest

The authors declare no competing financial interest.

Appendix A. Supplementary data

Supplementary data to this article can be found online at

References

- [1] J.R.L. Ehrenkranz, N.G. Lewis, C.R. Kahn, J. Roth, Phlorizin: a review, *Diabetes Metab. Res. Rev.* 21 (2005) 31–38.
- [2] S. Nakka, L. Guruprasad, Structural insights into the active site of human sodium dependent glucose Co-transporter 2: homology modelling, molecular docking, and 3D-QSAR studies, *Aust. J. Chem.* 65 (2012) 1314–1324.
- [3] V. Suryanarayanan, A. Sudha, S. Rajamanikandan, R. Vanajothi, P. Srinivasan, Atom-based 3D QSAR studies on novel Nb-D-xylosylindole derivatives as SGLT2 inhibitors, *Med. Chem. Res.* 22 (2012) 615–624.
- [4] K. Rebekah, A.J. Prasoon, Y. Mukesh, S. Nishant, N.S. Anuraj, J. Shobha, Optimization of gaussian Kernel function in support vector machine aided QSAR studies of C-Aryl glucoside SGLT2 inhibitors, *Interdiscip. Sci. Comput. Life Sci.* 5 (2013) 45–52.
- [5] V.K. Vyas, H.G. Bhatt, P.K. Patel, J. Jalu, C. Chintha, N. Gupta, M. Ghate, CoMFA and CoMSIA studies on C-aryl glucoside SGLT2 inhibitors as potential antidiabetic agents, *SAR QSAR Environ. Res.* 24 (2013) 519–551.
- [6] S. Nair, J.P.H. Wilding, Sodium glucose cotransporter 2 inhibitors as a new

treatment for diabetes mellitus, *J. Clin. Endocrinol. Metab.* 95 (2010) 34–42.

[7] I. Idris, R. Donnelly, Sodium–glucose co-transporter-2 inhibitors: an emerging new class of oral antidiabetic drug, *Diabetes Obes. Metab.* 11 (2009) 79–88.

[8] J.J. Liu, T.W. Lee, R.A. DeFronzo, Why do SGLT2 inhibitors inhibit only 30–50% of renal glucose reabsorption in humans? *Diabetes.* 61 (2012) 2199–2204.

[9] D. Gonzaga, M.R. Senger, F.C. da Silva, V.F. Ferreira, F.P. Silva -Jr, 1-phenyl-1H- and 2-phenyl-2H-1,2,3-triazol derivatives: Design, synthesis and inhibitory effect on alpha-glycosidases, *Eur. J. Med. Chem.* 74 (2014) 461–476.

[10] A.R. Li, J. Zhang, J. Greenberg, T. Lee, J. Liu. Discovery of non-glucoside SGLT2 inhibitors, *Bioorg. Med. Chem.* 21 (2011) 2472–2475.

[11] X. Du, M. Lizarzaburu, S. Turcotte, T. Lee, J. Greenberg, B. Shan, P. Fan, Y. Ling, J.C. Medina, J. Houze. Optimization of triazoles as novel and potent nonphlorizin SGLT2 inhibitors, *Bioorg. Med. Chem. Lett.* 21 (2011) 3774–3779.

[12] H. Zhi, J. Zheng, Y. Chang, Q. Li, G. Liao, Q. Wang, P. Sun, QSAR studies on triazole derivatives as sglt inhibitors via CoMFA and CoMSIA, *J. Molecu. Struct.* 1098 (2015) 199–205.

- [13] E. Ibezim, P.R. Duchowicz, E.V. Ortiz, E.A. Castro, QSAR on aryl-piperazine derivatives with activity on malaria, *Chemom. Intell. Lab. Syst.* 110 (2012) 81–88.
- [14] F. Gharagheizi, M. Mehrpooya, Prediction of some important physical properties of sulfur compounds using quantitative structure–properties relationships, *Mol Divers* 12 (2008) 143–155.
- [15] I. García, Y. Fall, X. García-Mera, F. Prado-Prado, Theoretical study of GSK-3 α : neural networks QSAR studies for the design of new inhibitors using 2D descriptors, *Mol Divers* 15 (2011) 947–955.
- [16] M.D. Ertürk, M.T. Saçan, M. Novicb, N. Minovski, Quantitative structure–activity relationships (QSARs) using the novel marine algal toxicity data of phenols, *J. Mol. Graphics Modell.* 38 (2012) 90–100.
- [17] M.H. Fatemi, P. Izadiyan, Cytotoxicity estimation of ionic liquids based on their effective structural features, *Chemosphere* 84 (2011) 553–563.
- [18] M. Goodarzi, M.P. Freitas, C.H. Wu, P.R. Duchowicz, pKa modeling and prediction of a series of pH indicators through genetic algorithm-least square support vector regression, *Chemom. Intell. Lab. Syst.* 101 (2010) 102–109.

- [19] H. Du, J. Wang, Z. Hu, X. Yao, X. Zhang, Prediction of Fungicidal Activities of Rice Blast Disease Based on Least-Squares Support Vector Machines and Project Pursuit Regression, *J. Agric. Food Chem.* 56 (22)(2008) 10785–10792.
- [20] A.G. Mercader, P.R. Duchowicz, F.M. Fernández, E.A. Castro, Replacement method and enhanced replacement method versus the genetic algorithm approach for the selection of molecular descriptors in QSPR/QSAR theories, *J. Chem. Inf. Model* 50(2010) 1542–1548.
- [21] P. Cortez, M.J. Embrechts, Using sensitivity analysis and visualization techniques to open black box data mining models, *Inform. Sciences* 225 (2013) 1–17.
- [22] F.O. Andersson, M. Åberg, S.P. Jacobsson, Algorithmic approaches for studies of variable influence, contribution and selection in neural networks, *Chemom. Intell. Lab. Syst.*, 51(2000) 61–72.
- [23] Talete srl. Dragon (Software for Molecular Descriptor Calculation) Version 6.0. <http://www.talete.mi.it/>.
- [24] M. Rybka, A.G. Mercader, E.A. Castro, Predictive QSAR study of chalcone derivatives cytotoxicity activity against HT-29 human colon adenocarcinoma cell lines, *Chemom. Intell. Lab. Syst.* 132 (2014) 18–29.

[25] J.S. Torrecilla, J. García, E. Rojo, F. Rodríguez, Estimation of toxicity of ionic liquids in leukemia rat cell line and acetylcholinesterase enzyme by principal component analysis, neural networks and multiple linear regressions, *J. Hazard. Mater.* 164(2009) 182–194.

[26] J.S. Torrecilla, J. García, E. Rojo, F. Rodríguez, Estimation of toxicity of ionic liquids in Leukemia Rat Cell Line and Acetylcholinesterase enzyme by principal component analysis, neural networks and multiple linear regressions, *J. Hazard. Mater.* 164 (2009) 182–194.

[27] J. Parinet, M. Julien, P. Nun, R.J. Robins, G. Remaud, P. Höhener, Predicting equilibrium vapour pressure isotope effects by using artificial neural networks or multi-linear regression – A quantitative structure property relationship approach, *Chemosphere* 134 (2015) 521–527.

[28] Y. Yuan, R. Zhang, R. Hu, X. Ruan, Prediction of CCR5 receptor binding affinity of substituted 1-(3,3-diphenylpropyl)-piperidiny l amides and ureas based on the heuristic method, support vector machine and projection pursuit regression, *Eur. J. Med. Chem.* 44 (2009) 25–34

[29] K.P. Singh, N. Basant, S. Gupta, Support vector machines in water quality

management, *Anal. Chim. Acta* 703(2011) 152–162.

[30] H. Du, J. Wang, J. Watzl, X. Zhang, Z. Hu, Prediction of inhibition of matrix metalloproteinase inhibitors based on the combination of Projection Pursuit Regression and Grid Search method, *Chemom. Intell. Lab. Syst.* 93 (2008) 160–166.

[31] J.H. Friedman, W. Stuetzle, Projection Pursuit Regression, *J. Am. Stat. Assoc.* 76 (1981) 817–823.

[32] M. Herrera, L. Torgo, J. Izquierdo, R. Pérez-García. Predictive models for forecasting hourly urban water demand, *J. Hydrol.* 387 (2010) 141–150.

[33] H. Du, J. Wang, J. Watzl, X. Zhang, Z. Hu, Prediction of inhibition of matrix metalloproteinase inhibitors based on the combination of Projection Pursuit Regression and Grid Search method, *Chemom. Intell. Lab. Syst.* 93 (2008) 160–166.

[34] Y. Du, Y. Liang, D. Yun, Data Mining for Seeking an Accurate Quantitative Relationship between Molecular Structure and GC Retention Indices of Alkenes by Projection Pursuit, *J. Chem. Inf. Comput. Sci.* 42 (2002) 1283–1292.

[35] R. Gupta, A. Majumdar, Reconsidering the welfare cost of inflation in the US: a nonparametric estimation of the nonlinear long-run money-demand equation using projection pursuit regressions, *Empir. Econ.* 46 (2014) 1221–1240.

[36] Y. Yuan, R. Zhang, R. Hu, X. Ruan, Prediction of CCR5 receptor binding affinity

of substituted 1-(3,3-diphenylpropyl)-piperidinyI amides and ureas based on the heuristic method, support vector machine and projection pursuit regression, *Eur. J. Med. Chem.* 44 (2009) 25–34.

[37] J.A. van Leeuwen, R.J. Jonker, R. Gill, Octane number prediction based on gas chromatographic analysis with non-linear regression techniques, *Chemom. Intell. Lab. Syst.* 25 (1994) 325–340.

[38] N. Chirico, P. Gramatica, Real external predictivity of QSAR models. part 2. new intercomparable thresholds for different validation criteria and the need for scatter plot inspection, *J. Chem. Inf. Model.* 52 (2012) 2044–2058.

[39] M. Ebrahimi, T. Khayamian, Interactions of G-quadruplex DNA binding site with berberine derivatives and construct a structure-based QSAR using docking descriptors, *Med. Chem. Res.* 23 (2014) 1327–1339.

[40] M. Karelson, V.S. Lobanov, Quantum-Chemical Descriptors in QSAR/QSPR Studies, *Chem. Rev.* 96 (1996) 1027–1043.

[41] N. Khatri, A.K. Madan, Models for H3 receptor antagonist activity of sulfonylurea derivatives, *J. Mol. Graphics Modell.* 48 (2014) 87–95.

[42] R. Todeschini, V. Consonni, *Molecular Descriptors for Chemoinformatics*, vol. I/II, 2nd ed., Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim, 2009.

[43] R. Guha, S.C. Schürer, Utilizing high throughput screening data for predictive toxicology models: protocols and application to MLSCN assays, *J. Comput. Aided Mol. Des.* 22 (2008) 367–384.

[illegible]

Figure 2

Figure 3

Figure 4