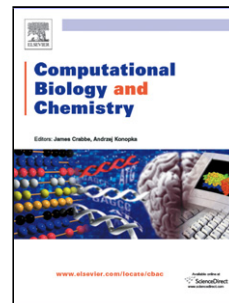


Accepted Manuscript

Title: Predicting human intestinal absorption of diverse chemicals using ensemble learning based QSAR modeling approaches

Author: Nikita Basant Shikha Gupta Kunwar P. Singh



PII: S1476-9271(16)30042-1
DOI: <http://dx.doi.org/doi:10.1016/j.compbiolchem.2016.01.005>
Reference: CBAC 6499

To appear in: *Computational Biology and Chemistry*

Received date: 24-9-2014
Revised date: 18-1-2016
Accepted date: 21-1-2016

Please cite this article as: Basant, Nikita, Gupta, Shikha, Singh, Kunwar P., Predicting human intestinal absorption of diverse chemicals using ensemble learning based QSAR modeling approaches. *Computational Biology and Chemistry* <http://dx.doi.org/10.1016/j.compbiolchem.2016.01.005>

This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

Predicting human intestinal absorption of diverse chemicals using ensemble learning based QSAR modeling approaches

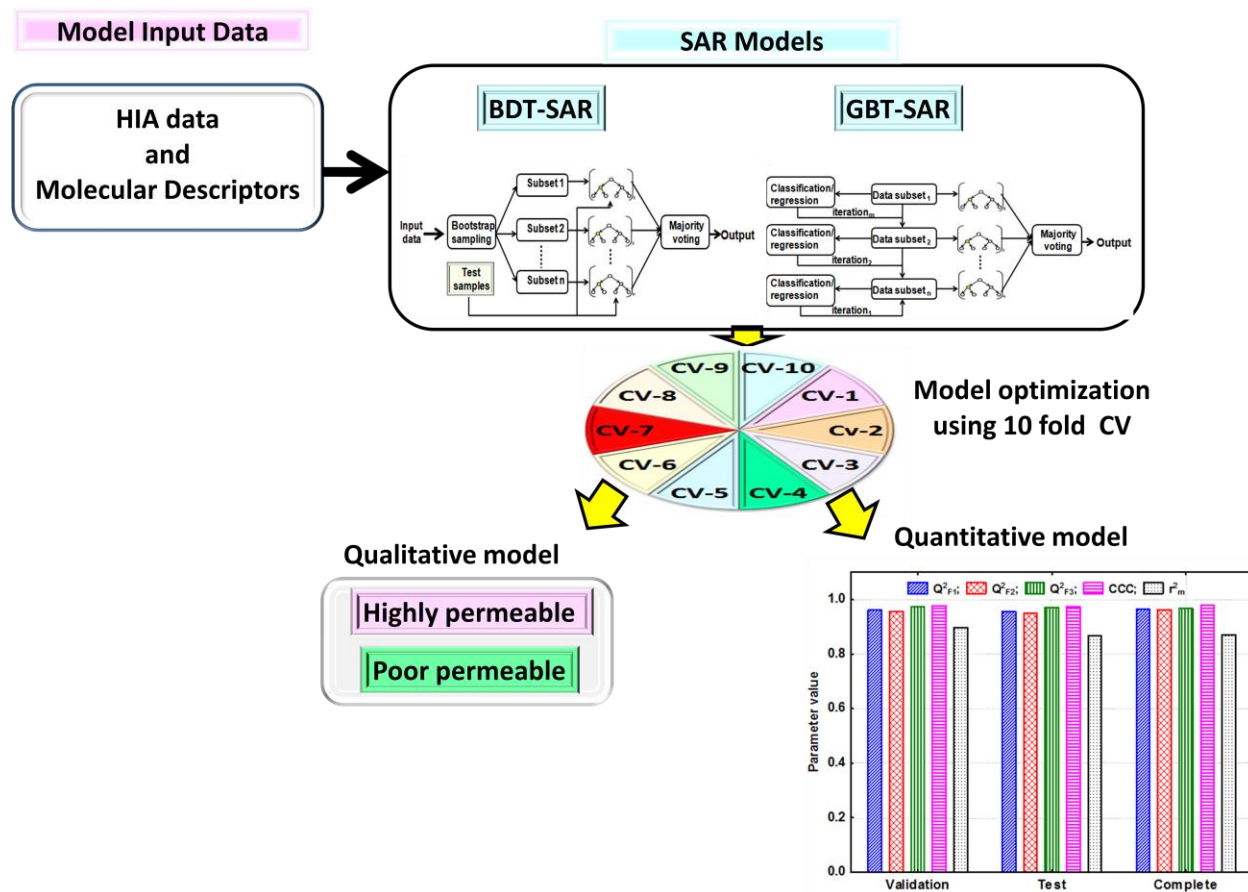
Nikita Basant¹, Shikha Gupta², Kunwar P. Singh^{2*} kpsingh_52@yahoo.com
kunwarpsingh@gmail.com

¹ETRC, Gomtinagar, Lucknow-226010, India

²Environmental Chemistry Division, CSIR-Indian Institute of Toxicology Research, Post Box 80, Mahatma Gandhi Marg, Lucknow-226 001, India

*Corresponding author: Tel: 0091-522-2476091; Fax: 0091-522-2628227

Graphical Abstract



Highlights

- Qualitative/quantitative QSARs developed for predicting HIA of chemicals.
- Structural diversity and nonlinearity in data tested using TSI and BDS statistics.
- QSARs validated through OECD recommended stringent parameters.
- Proposed QSARs precisely predicted HIA of diverse chemicals.
- Proposed QSARs can be useful tools in screening new drug molecules.

Abstract

Human intestinal absorption (HIA) of the drugs administered through the oral route constitutes an important criterion for the candidate molecules. The computational approach for predicting the HIA of molecules may potentiate the screening of new drugs. In this study, ensemble learning (EL) based qualitative and quantitative structure-activity relationship (SAR) models (gradient boosted tree, GBT and bagged decision tree, BDT) have been established for the binary classification and HIA prediction of the chemicals, using the selected molecular descriptors. The structural diversity of the chemicals and the nonlinear structure in the considered data were tested by the similarity index and Brock-Dechert-Scheinkman statistics. The external predictive power of the developed SAR models was evaluated through the internal and external validation procedures recommended in the literature. All the statistical criteria parameters derived for the performance of the constructed SAR models were above their respective thresholds suggesting for their robustness for future applications. In complete data, the qualitative SAR models rendered classification accuracy of >99%, while the quantitative SAR models yielded correlation (R^2) of >0.91 between the measured and predicted HIA values. The performances of the EL-based SAR models were also compared with the linear models (linear discriminant analysis, LDA and multiple linear regression, MLR). The GBT and BDT SAR models performed better than the LDA and MLR methods. A comparison of our models with the previously reported QSARs for HIA prediction suggested for their better performance. The results suggest for the appropriateness of the developed SAR models to reliably predict the HIA of structurally diverse chemicals and can serve as useful tools for the initial screening of the molecules in the drug development process.

Keywords: Human intestinal absorption; ensemble learning; structure-activity relationship; diverse chemicals; qualitative; quantitative models.

1. Introduction

The preclinical ADME (Absorption, Distribution, Metabolism, and Elimination) screening strategies are considered an important tool for the early identification of molecules as a potential candidate in the drug development process (Guerra et al., 2010). Human intestinal absorption (HIA) is one the most important ADME properties as it constitutes one of the key steps during the drugs transporting to their targets. Although, numerous *in vitro* assay methods have been developed for the HIA, the throughput capacity is very low as compared to the high-throughput (HT) activity assay and combinatorial synthesis, thus become a “bottle-neck” in the drug discovery process. Moreover, these methods are still resource-intensive, time-consuming and inadequate to deal with the large number of molecules as the future drug candidates being developed worldwide. Therefore, it is desirable to develop robust and accurate computational methods for predicting the intestinal absorption of the drug molecules. Using *in silico* methods to evaluate the ADME properties has become a practicable alternative choice so far, which could break through the bottleneck in the HT drug discovery process (Talevi et al., 2011). Several qualitative and quantitative structure-activity relationships (SARs) based on the linear and nonlinear modeling methods, such as the multiple linear regression (MLR), partial least squares (PLS), artificial neural networks (ANN), support vector machines (SVMs), decision tree induction (DTI) and decision tree (DT) have been used for predicting the HIA of molecules (Hou et al., 2007a; 2007b; 2007c; Yan et al., 2008; Guerra et al., 2010; Shen et al., 2010; Obrezanova and Segall, 2010; Suenderhauf et al., 2011; Talevi et al., 2011; Ghafouriana et al., 2012). Although, some of these studies reported prediction results in acceptable ranges, most of these were based on a small dataset, and for a limited data, the statistical models often fail due to an over-fitting problem, resulting in a limitation on their use. Many QSAR studies have

limitations of their low predictive power when applied to new chemicals not included in the model building phase. It may be due to the lack of external validation of the model, selection of inappropriate modeling method or irrelevant descriptors. Moreover, the experimental biological activity data generally exhibit nonlinear structure and linear methods fail to capture the nonlinear dependence (Singh et al., 2013a). Therefore, robust QSAR models based on the appropriate modeling approach, relevant set of descriptors, and validated through stringent criteria are required for predicting the HIA of the molecules.

Ensemble learning (EL) methods have been considered as unbiased tools for modeling the complex relationships between the independent (predictor) and the dependent variables (Yang et al., 2010). These methods overcome the problems with weak predictors, alleviate the small sample size problem and reduce the over-fitting in the training phase (Dietterich, 2000). Bagged decision tree (BDT) and gradient boosted tree (GBT) implementing the bagging and the boosting techniques, respectively, improve the accuracy of a predictive function (Yang et al., 2010). These techniques are inherently non-parametric statistical methods and make no assumption regarding the underlying distribution of the values of the predictor variables and can handle numerical data that are highly skewed or multi-modal in nature (Singh et al., 2014; Singh and Gupta, 2014). To our knowledge, EL methods have not yet been applied to the HIA prediction modeling.

This work aimed to develop reliable qualitative and quantitative SAR models for predicting the HIA of chemicals using simple descriptors derived from the molecular structure. Accordingly, EL based GBT and BDT models were constructed using the HIA data to predict the intestinal absorption classes and the end-point values of the chemicals. The predictive and generalization abilities of the constructed SARs were evaluated using various stringent criteria parameters and the external predictive power of these models was established using the test data.

The performances of the EL-based SAR models were also compared with the linear qualitative (linear discriminant analysis, LDA) and quantitative (multiple linear regression, MLR) models.

2. Materials and methods

2.1 Data sets

The dataset used here consisted of the human intestinal absorption (%HIA) data for 578 compounds (Shen et al., 2010). The original data has been compiled by Hou et al. (2007b) from multiple literature sources (Palm et al., 1997; Wessel et al., 1998; Dollery, 1999; Zhao et al., 2001; Deretey et al., 2002; Hou et al., 2007c). According to the information available in the literature, the experimental HIA values of the chemicals are based on (i) the percentage of cumulative drug and its metabolites in urine following oral administration, (ii) indirect measurements, such as bioavailability, the excretion in urine and feces following oral administration, and (iii) the ratio of cumulative urinary excretion of drug-related material following oral and intravenous administration. The data has earlier been used in other modeling studies (Hou et al., 2007a,b; Shen et al., 2010). The compounds in the dataset belong to a wide variety of pharmacological and chemical classes. One of the molecules in the dataset (molecule-618 in supplementary material of Shen et al., 2010) was not included in the present study due to non-availability of structure. Finally, 577 compounds were retained for SAR modeling here. The HIA (%) values of the selected compounds varied between 0 and 100 (Table S1, Supplementary material). A histogram of the experimental HIA values is plotted in Fig. 1. It is evident that the HIA values show a right skewed distribution pattern, which suggests that the majority of the molecules have high intestinal absorption.

2.2 *Molecular descriptors and data processing*

In this study, 104 2D molecular descriptors (constitutional, and topological) were calculated for each compound using the MOSES Descriptor Community Edition (<http://www.molecular-networks.com/services/mosesdescriptors>). For calculating the descriptors, SMILES (simplified molecular input line entry system) of the compound were converted into the sdf files and were used in the MOSES program. The SMILES of the compounds were obtained from the ChemSpider (www.chemspider.com). The chemical structures available in ChemSpider corresponding to the SMILES of the considered molecules were compared with those in the Pubchem (<http://pubchem.ncbi.nlm.nih.gov/compound/>). The compounds for which the chemical structures were found different, the SMILES of such molecules were taken from the Pubchem for descriptor calculation. For the selection of the descriptors, those with low variation (<1) were excluded from the pool. Model fitting approaches were then adopted for the selection of the relevant descriptors for SARs development. Prior to the model fitting, the data were split into the training (70 %), validation (15 %), and test (15 %) subsets using the random distribution approach. The random sampling approach allows an estimate of model prediction accuracy and provides an estimate of its variability. Previous studies have shown that this is a reliable approach to evaluate the model validity (Singh et al., 2013b). Such test sets are commonly accepted as the gold standard to assess the real predictivity of the QSARs (Benigni et al., 2007). The qualitative and quantitative SAR models were constructed with the training data using all the descriptors left (68) in the pool. The optimal values of the model parameters were determined using the respective scoring functions (misclassification rate and root mean squared error) to rank the contribution of features in the current data (training and validation). The lowest ranked features were then removed (Singh et al., 2013a) in the successive steps. The optimal features

were retained and the corresponding prediction accuracy was computed by means of a 10-fold cross validation (CV). Finally, sets of four descriptors for the qualitative and quantitative SAR models were considered in this study (Table 1 and Table S1). The basic statistics of the selected descriptors are given in Table 2.

Since, both the nonlinearities in experimental data and structural diversity of the considered molecules are considered important factors in the selection of the appropriate modeling approach and to develop global predictive QSAR models (Zhao et al., 2006), these were evaluated using the Brock-Dechert-Scheinkman (BDS) statistics (Brock et al., 1996) and Tanimoto similarity index (TSI), respectively. The BDS test is a two-tail nonparametric method for testing the serial independence and nonlinear structure in a data based on the correlation integral. If the computed BDS statistics exceed the critical value at the conventional level, the null hypothesis of linearity is rejected, which reveals the presence of the nonlinear dependence in the data. In this study, the exceeding BDS statistics ($p < 0.01$) suggest for severe nonlinearity in data structure and hence, a nonlinear model is required for developing appropriate QSAR models. TSI is an appropriate distance metric for topology-based chemical similarity studies, which calculates the Tanimoto similarity between the fingerprint of a chemical and a consensus fingerprint, which is 1024 bit fingerprint (Toxmatch, Ideaconult Ltd.). The fingerprint generation is based on the fingerprint implementation of the open source Chemoinformatics library (Steinbeck et al., 2003). The distribution of the TSI values of the considered chemicals is plotted in Fig. 2. The TSI values of the compounds varied between 0.0002 and 0.297 with a mean of 0.017. A good cutoff for the biologically similar molecules is 0.7 or 0.8 (Singh et al., 2013a). The distribution of the TSI values (Fig. 2) suggests that the compounds considered in this work represent sufficiently high structural diversity and warrant model stability with the

suitability of the external test set for assessing the predictive performance of the developed QSAR models.

Criteria for the categorization of the compounds as good intestinal absorption (>30%-100%) and poor absorption ($\leq 30\%$) was used (Shen et al., 2010). In this study, a total 499 compounds belong to the category of good intestinal absorption (class 1) and the rest of the 78 as a poor absorption category (class 2). Further, in view of the highly imbalanced nature of the dataset (499 compounds in class 1, 78 compounds in class 2) in the qualitative SAR modeling, we attempted to construct the qualitative SAR models using a balanced dataset in the training phase (78 compounds in each of the two classes) and keeping the remaining 421 compounds (class 1) in the test set for prediction.

2.3 *Qualitative and quantitative SAR modeling*

2.3.1 *Linear SAR modeling*

Here, linear classification and regression models were established using the linear discriminant analysis (LDA) and multiple linear regression (MLR) methods, respectively. LDA, a classification method constructs a discriminant function (DF) for each group (Singh et al., 2004) as; $f(G_i) = k_i + \sum_{j=1}^n w_{ij}p_{ij}$, where i is the number of groups (G), k_i is the constant inherent to each group, n is the number of parameters used to classify a set of data into a given group, w_j is the weight coefficient assigned by LDA to a given selected parameter (p_j). Here, two groups of chemicals (good and poor intestinal absorption) were considered and bilinear DFs were constructed to discriminate between these groups of chemicals.

The MLR quantifies the relationship between several independent or predictor variables and a dependent variable (Bordbar et al., 2013). A set of coefficients defines the single linear

combination of independent variables (molecular descriptors) that best describes the dependent variable. A MLR model can be represented as: $y = \beta_0 + \beta_1x_1 + \beta_2x_2 + \beta_3x_3 + \dots + \beta_kx_k$, where, k is the number of independent (x) variables, $\beta_1 \dots \beta_k$, the regression coefficients and y is the dependent variable. Regression coefficients represent the independent contributions of each input variable.

2.3.2 *EL-based SAR modeling*

Here, we constructed the EL-based qualitative and quantitative SAR models (GBT, BDT) for predicting the HIA of diverse drugs and drug like compounds. Accordingly, the qualitative SAR model to predict the category of the chemicals (good and poor intestinal absorption) and quantitative SAR model to predict the intestinal absorption (%HIA) of the diverse chemicals were established using the set of selected simple molecular descriptors.

GBT, a non-parametric statistical methodology, combines the strength of the decision tree (DT) and boosting. The central idea behind boosting is the combination of many weak models into a powerful ensemble with a greatly improved performance (Friedman, 2002). Initially, it selects a certain tree population and the first tree is fitted to the data. The residuals from the first tree are then fed into the second tree and repeating the process through a chain of successive trees. The final predicted value is formed by adding the weighted contribution of each tree (Singh et al., 2014). The size of the tree is an important consideration for the model accuracy. Here, the optimal size of the tree was decided using the criteria of minimal cross-validation error. Boosting improves the accuracy of a predictive function by applying it repeatedly in a series and combining the output of each function with weighting, so that the total error of prediction is minimized (Friedman, 2002). The GBT uses the Huber M - regression loss

function which makes it highly resistant to the outliers. The number and depth of the trees are the method's parameters which can be adjusted for a data set at hand. It controls the maximum allowed level of interaction between the variables in the model.

BDT, an ensemble of DTs, combines the individual predictions. It generates an ensemble of individual trees by bootstrap sampling of the training dataset. Multiple samples from the training set are generated by sampling with replacement from the training data. Separate models are produced and used to predict the entire data from the aforesaid sub-sets. Then various estimated models are aggregated by using the mean for regression problems or majority voting for classification problems (Pino-Mejias et al., 2008). The BDTs gaining strength from the bagging technique use the out of bag data rows for model validation. The stochastic element in BDT algorithm makes it highly resistant to over-fitting. The number of trees in a random forest and depth of individual trees are the method's parameters which need to be adjusted for an optimal model selection.

2.4 *Model parameter optimization and validation*

The HIA data were split into the training, validation and test sets and the qualitative (binary) and quantitative SAR models based on the linear (LDA, MLR) and EL approach (GBT, BDT) were established with the training and validation data, while keeping the test data for the external validation of the constructed SARs. The optimal parameters of the models and the number of relevant descriptors were determined following the k-fold CV procedure. In CV, the data D are divided into k non-overlapping subsets, D_1, D_2, \dots, D_k . At each iteration i ($i = 1$ to k), the model is trained with $D - D_i$ and tested on D_i . This approach has an advantage that each test set is independent of the others (Singh et al., 2013a). The optimal models were selected on the basis of the misclassification rate (MR) and the root mean squared error (RMSE) in the training and

validation data.

Randomization tests were performed to avoid any possible inclusion of the fortuitous correlations in the proposed SAR models. New models were recalculated for randomly shuffled target values with original independent descriptors matrix. The models, thus obtained from the training set with randomized target values should have significantly lower predicted accuracy values than the proposed ones because the relationship between the structure and the property has been broken (Xu et al., 2011).

External validation of the constructed SAR models was performed by using the test set composed of the data not used in the model building phase. Various stringent criteria for validation of the QSAR models recommended in the literature were considered. The concordance correlation coefficient (CCC) in the external validation measures the precision and accuracy of the predictive model (Lin, 1992). Since, it does not involve the training set information, it is considered a true external validation measure independent of the sampled chemical space. Other criteria, such as Q_{F1}^2 , Q_{F2}^2 , and Q_{F3}^2 have also been proposed in the literature for external validation of the predictive models (Shi et al., 2001; Schuurmann et al., 2008; Consonni et al., 2009). Further, Roy et al. (2013) proposed additional criterion based on the correlations between the observed and model predicted values of the target property with (r^2) and without (r_o^2) intercept for the least-square lines. The performance of the proposed quantitative SAR models in predicting the intestinal absorption of the chemicals was also assessed by calculating the R^2 and the RMSE values in the training, validation, test and complete data arrays. Tests based on the Cooper's statistics (Cooper et al., 1979) were used to evaluate the predictive performance of the qualitative SAR models for binary discrimination of the compounds. Accordingly, the values of the sensitivity, specificity, accuracy and Matthew's

correlation coefficient (MCC) were enumerated for the prediction results.

Modelling was performed using DTREG (<http://www.dtreg.com>). All other computations were done in Excel 97. The TSI values were calculated using Toxmatch (https://eurl-ecvam.jrc.ec.europa.eu/laboratoriesresearch/predictive_toxicology/qsar_tools/toxmatch).

2.5 *Applicability domain*

For a predictive SAR model aimed for screening new molecules, it is very important to define its applicability domain (AD), which is a theoretical region in the space defined by the descriptors used in the model building and the modeled response. In the present study, the AD of the constructed SARs was determined by the range of the individual descriptors in the training and external sets, as well as by the leverage approach. In the leverage approach, Williams plot was drawn using the standardized residuals and the leverages. The leverage value, h_i for each i^{th} compound is calculated (Puzyn et al., 2011) from the descriptors ($i \times j$) matrix (\mathbf{X}) as, $h_i = \mathbf{x}_i^T (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{x}_i$, where \mathbf{x}_i is a raw vector of the molecular descriptors for a particular i^{th} compound. The obtained Williams plot was used to detect both the response outliers and the structurally influential chemicals in the model. The warning leverage h^* is, generally, fixed at $3(p+1)/n$, where p is the number of variables in the model, and n is the number of training compounds. The value of h_i greater than the critical h^* value indicates that the structure of the compound differs from those used for the calibration. The leverage can be used as a quantitative measure of the model AD suitable for evaluating the degree of extrapolation.

3. Results and discussion

Basic statistics of the selected molecular descriptors for SARs modeling (Table 2) suggest that the standard deviation (SD) and the coefficient of variation (CoV) indicated high variability within the selected descriptors. The CoV values of the descriptors vary between 42.40% (NAtoms) and 157.47% (HDon_N). The qualitative and quantitative SAR models were established using the training (403 compounds) and validation (87 compounds) data, while keeping the test (87 compounds) subsets away from the model building phase, which were then used for the external validation purpose. The prediction reliability can be best checked by means of an external test set with new moieties not included in model building (Benigni et al., 2007).

3.1 Qualitative SAR model:

The qualitative SAR models were constructed using four molecular descriptors (Table 1) for discriminating the compounds based on their intestinal absorption (good and poor). Accordingly, EL- based SAR models (GBT, BDT) were established and optimized using the 10-fold CV, which finally have the total number of trees in series, maximum depth of any tree, and the number of average group splits of 501, 7, 245.9 and 200, 14, 34.9, respectively, yielding the mean MR of 8.16% (GBT) and 8.16% (BDT), which are close to the test (2.30 %) set value. The results show no obvious over-fitting of the data. In Y-randomization, the average MR values of the scrambled SAR models were 15.16% (GBT), and 13.93% (BDT), respectively, which are significantly higher than those of the original models in training (0.25 %). This suggests that both the original qualitative models are relevant and unlikely to arise as a result of chance of correlation.

The contributions of the four selected descriptors in two qualitative models (GBT, BDT) ranged between 25.38% -100 % and 31.37% - 100 %, respectively, with the highest contribution

of TPSA (100%) followed by the XLogP in both the cases. Yan et al. (2008) reported that TPSA and XLogP were the most significant descriptors in the intestinal absorption prediction. The individual contributions of all the four descriptors in GBT and BDT models are plotted in Fig. 3. The higher the relative influence of a feature, the more important it is in the prediction process (Singh et al., 2014).

The optimal qualitative SAR models were then applied to the test, and complete datasets. The performance parameters of the qualitative SAR models in the training, validation, test and complete data are shown in Table 3. A nearly perfect (~ 99%) classification of the chemicals into two categories (complete data) by both the modeling approaches here may be due to the fact that the chemicals in two classes differ significantly in their characteristics (descriptors) considered for modeling. A cursory look on the values of these descriptors corresponding to the chemicals in two different categories (Table 4) revealed that the range and mean values of these descriptors for the compounds with poor HIA category were; 0.00-13.0, 2.78 (HDon_O); -7.48-6.78, -0.16 (XLogP); 0.00-530.49, 144.75 (TPSA); 15.0-176.0, 56.77 (NAtoms), whereas, for the compounds in good HIA class were; 0.00-6.0, 0.72 (HDon_O); -3.34-7.82, 2.45 (XLogP); 3.24-201.85, 64.13 (TPSA); 9.0-136.0, 41.85 (NAtoms), respectively. It is evident that in general, the mean values of all the descriptors (except XLogP) are higher for the chemicals in poor HIA class, and the values of all the descriptors for the chemicals in two classes are significantly different.

The MR and the MCC values yielded by the binary discrimination models (GBT, BDT) in the complete data were 0.69% and 0.97, respectively. MR represents the percentage of the misclassified compounds in two classes. MCC value equal to 1 is regarded as a perfect prediction, whereas, 0 is for a completely random prediction. As shown in Table 3, the

performance of the proposed SAR models was excellent. The results further showed that the sensitivity and specificity values for both the qualitative models were more than 99% in complete data. A QSAR model for screening of new molecules in drug discovery should exhibit high sensitivity. A low sensitivity value indicates the low ability of a model to recognize the target property of diverse compounds. On the other hand, the high specificity value indicates the high ability of the model to recognize the false positive compounds (Singh et al., 2013a). An investigation of the results suggests that in total four compounds each by GBT (bephenium, obidoxime, poldine, propantheline) and BDT (bephenium, edrophonium, obidoxime, poldine) were misclassified. This may be due to the fact that these compounds were not appropriately represented by the set of selected descriptors in the qualitative SAR models. Further, the performances of the EL-based (GBT, BDT) qualitative SAR models were also compared with that of the LDA, which in training, validation and test sets yielded classification accuracies of 89.33%, 95.40%, and 94.25%, respectively. It is evident that the EL-based SAR models performed relatively better than the LDA (Table 3). The LDA is a linear classification method which is unable to capture the nonlinearities in data.

Qualitative SAR models (GBT and BDT) were also constructed with the balanced data considering an equal number of compounds ($n=78$) in both the classes in the training set and applied to predict the test set with the remaining 421 compounds (class 1). These models (GBT, BDT) yielded classification accuracies of 99.36% in the training and 99.52% in the test data, respectively. These results are closely comparable with the original models (Table 3).

Several models based on the PLS, DTI, DT, ANN, and SVM approaches have been proposed earlier for the qualitative prediction of HIA of the chemicals (Table 5). These studies were based on different numbers of compounds and descriptors and except a few (based on the

fingerprints descriptors), generally reported low prediction accuracies. Among these, Hou et al. (2007a) and Shen et al. (2010) developed SVM based binary classification models using the similar number of compounds and reported the classification accuracies of 98 % and 99 %, respectively, in the test data. These results were closely comparable with those of the present study. Moreover, the proposed EL based SAR models using a few simple descriptors here yielded higher classification accuracies as compared to the other earlier studies (Table 5). It may be due to the fact that the proposed approaches in this study implement stochastic gradient boosting and bagging algorithms which improve the prediction accuracy of the weak learners (Breiman, 1996; Singh et al., 2014).

Further, in order to make a closer comparison of the modeling results with those of Hou et al. (2007a), we constructed the EL-based qualitative SAR models (GBT, BDT) considering an identical split (479 in training and 98 in test) of compounds. The two models (GBT, BDT) yielded the classification accuracies of 99.37%, 99.58% in the training and 95.92%, 94.90% in the test phases, respectively. These results are closely comparable with those of the Hou's work (Table 5).

3.2 *Quantitative SAR model:*

EL-based quantitative SAR models (GBT, BDT) were constructed for predicting the intestinal absorption (HIA%) of the chemicals using three and four molecular descriptors (Table 1). The optimal values of the model parameters, total number of trees in series, maximum depth of any tree, and the number of average group splits were 540, 13, 1167.8 (GBT) and 247, 29, 179.4 (BDT), respectively, and captured 96.5% and 87.63% of the data variance in the training. In 10-fold CV, the mean RMSE values for the GBT and BDT were 26.56 and 21.69,

respectively. The Y-randomization results (mean R^2) for the constructed SAR models derived through a 10-fold CV procedure were 0.002 for each, which revealed that the original models are relevant and unlikely to arise as a result of chance of correlation.

The contributions of the selected descriptors in the models (GBT, BDT) ranged between 21.41% -100 % and 16.03% -100 %, respectively. In both the models, the contribution of TPSA was the highest (100%) followed by NAtoms and HDon_O. The individual contributions of the descriptors in both the quantitative models are shown in Fig. 4. The optimal quantitative SAR models were applied to the test and complete data arrays. The GBT and BDT models explained 95.75%, 92.04%; 95.12%, 87.28% and 96.33%, 88.21% of the data variance in validation, test, and complete data subsets. The performance parameters of the two models are presented in Table 6. In complete data array, the two models (GBT, BDT) yielded RMSE and R^2 values of 6.16, 0.968 and 11.04, 0.908, respectively. Both the models yielded considerably low prediction error (RMSE) and high correlations (R^2) between the measured and predicted value of the response variable in the training, validation and test data (Figure 5; Table 6). The model is considered acceptable when the value of R^2 in external set exceeds 0.6 (Tropsha et al., 2011).

Further, the values of various coefficients such as CCC, Q_{F1}^2 , Q_{F2}^2 , Q_{F3}^2 and r_m^2 derived for the external validation, test, and complete data are displayed in Fig. 6. It is evident that the respective values of these coefficients for the two models in three datasets ranged between 0.93-0.98; 0.88-0.96; 0.87-0.96; 0.90-0.97 and 0.67-0.90. The value of a metric nearer to one indicates for the better quality of the model in terms of the external validation. According to Chirico and Gramatica (2012), the threshold criteria of 0.7 for Q_{F1}^2 , Q_{F2}^2 , Q_{F3}^2 ; 0.85 for CCC; and 0.65 for r_m^2 were proposed as an indicator for the acceptability of the predictive QSAR model. The validation of the proposed SAR models using the external data yielded criteria parameter values above their

respective thresholds (Fig. 6), which demonstrate high predictive power of the constructed SAR models for external prediction. The performance of the EL-based quantitative SAR models (GBT, BDT) were also compared with that of the MLR, which in training, validation and test sets yielded correlations (R^2) and RMSE values of 0.259, 29.37 (training); 0.447, 20.53 (validation); 0.683, 20.11 (test), and 0.308, 26.96 (complete data), respectively. It is evident that the performance of the EL-based quantitative SAR models was better than that of the MLR model, which is a linear method and unable to capture the nonlinear dependence in data.

Numerous models of HIA based on different descriptors have been gathered on different reviews. It reveals that the MLR, PLS, ANN, and SVM are the most common methods used in the oral absorption prediction. These models correlate the intestinal absorption with different types of descriptors (Table 7). It may be noted that different studies considered a different number of compounds and descriptors, thus, making it difficult to compare the prediction results with the present work. However, Hou et al. (2007b) and Reynolds et al. (2009) considering almost a similar number of compounds proposed the regression models based on the MLR and nonlinear approaches and reported R^2 values of 0.71 and 0.93, respectively, in the training data. The proposed quantitative SAR models in our study considering a significantly larger dataset and fewer simple descriptors performed better than the earlier ones. Moreover, the fact that a rigorous external validation was made in our models is an additional indication that these bear good predictive abilities.

3.3 *Model interpretation*

Here, qualitative and quantitative SAR models were developed using the constitutional (XLogP, NAtoms, HDon_O, HDon_N) and topological (TPSA) descriptors. All the selected descriptors, except XLogP, exhibited significant ($p < 0.05$) negative correlations with the HIA of

the considered compounds. All of the descriptors above can be used in combination to correlate with the intestinal absorption; however, the correlation decreases significantly when the descriptors are used independently, highlighting that the absorption is a complex process and is reliant and influenced by a number of different descriptors, not just one (Clark, 2011). In both the qualitative and quantitative SAR models, the TPSA has the highest contribution (100 %). This descriptor is a measure of the hydrophilicity and H-bonding potential of the molecule (Nuez and Rodriguez, 2008). Hydrophobicity is another physiologically important parameter in intestinal absorption and descriptors relating to it, such as XLogP, have a positive contribution to the predictions for passively absorbed compounds (Zakeri-Milani et al., 2006). An increase in the hydrophobicity would increase the permeation of the compound into and through the cell membrane in the intestine. A significant correlation of XLogP with HIA indicates that the highly lipophilic compounds are capable of moving across the membrane by diffusion. NAtoms, HDon_O and HDon_N are the constitutional descriptors reflecting the molecular composition of a compound without any information about its molecular geometry. The negative coefficient of these descriptors revealed that intestinal permeability decreases as the number of atoms and H-donors on the compound increases.

3.4 Applicability domain of the proposed models

The AD of the proposed SAR models was determined using both the descriptor's range and leverage approaches. The ranges of the implemented descriptors in both the qualitative and quantitative SAR models are presented in Table 8. The results show that none of the compounds in the qualitative and quantitative SAR model was out of their respective ADs.

From the Williams plot (Fig. 7), it is evident that sixteen compounds in GBT and twenty compounds in BDT SAR models exceeded their respective critical leverage values (0.030,

0.037), whereas four and eight compounds in two models exhibited standardized residual value above 3 i/j (Table 9). A deeper investigation revealed that the high leverage compounds generally have higher molecular mass (MW 342.3 - 1449.3 g/mol), whereas those with high residual values have relatively low molecular mass (MW < 440 g/mol). The absorption processes are strongly affected by the molecular mass and it is important to demonstrate the discriminating capacity of the permeability models for molecules with a molecular mass > 400 which due to high lipophilicity can become poorly permeable (Pham The et al., 2011). Further analysis revealed that the majority of the compounds exhibiting both the high leverage and high residual values in both the models belong to the category of poor absorption compounds ($HIA \leq 30\%$). In general, these compounds are natural products or their derivatives. Among these, compounds with low HIA values were over-estimated, whereas, those with high HIA values were under-estimated by both the constructed quantitative models. Such a pattern may be due to the highly imbalanced nature of the data (large variations in HIA values ranging between 0 and 100). Further, this may be due to the fact that these compounds were not appropriately represented by the set of the selected descriptors in the models.

4. Conclusions

EL-based qualitative and quantitative SAR models were established for the HIA prediction of the diverse molecules. Accordingly, the SAR models based on GBT and BDT methods were developed for the binary classification of the compounds and prediction of their HIA using simple molecular descriptors. The novelty of this work relies on the robustness of the developed models evaluated using several stringent external validation parameters, such as the classification accuracy, sensitivity, specificity, CCC, Q_{F1}^2 , Q_{F2}^2 , Q_{F3}^2 and r_m^2 , which rendered

highly reliable results. Moreover, the EL-based SAR models performed relatively better than the linear classification (LDA) and regression (MLR) models. A comparison of our models with the previously reported QSARs for HIA prediction also stresses their characteristics. The excellent predictive power of the proposed EL-based SARs models may be attributed to their ability for capturing the nonlinear dependence in the data; producing more robust models through combining the predictions from the ensemble of weak learners. Overall, the proposed SAR models may be a useful prediction tool in the initial screening of compounds in the drug development process.

Acknowledgement

The authors thank the Director, CSIR-Indian Institute of Toxicology Research, Lucknow (India) for his keen interest in this work and providing all necessary facilities. We also thank the unanimous reviewers for their critical suggestions leading to improvement of manuscript.

References

- Benigni, R., Netzeva, T.I., Benfenati, E., Bossa, C., Franke, R., Helma, C., Hulzebos, E., Marchant, C., Richard, A., Woo, Y.P., Yang, C., 2007. The expanding role of predictive toxicology: An update on the (Q)SAR models for mutagens and carcinogens. *J. Environ. Sci. Health Part C* 25, 53-97.
- Bordbar, M., Ghasemi, J., Faal, A.Y., Fazaeli, R., 2013. Chemometric Modeling to Predict Aquatic Toxicity of Benzene Derivatives Using Stepwise-Multi Linear Regression and Partial Least Square. *Asian J. Chem.* 25, 331-342.
- Breiman, L., 1996. Bagging Predictors. *Mach. Learn.* 24, 123–140.
- Brock, W.A., Dechert, W., Scheinkman, J.A., LeBaron, B.A., 1996. Test for independence based on the correlation dimension. *Economet Rev.* 15, 197-235.
- Chirico, N., Gramatica, P., 2012. 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, 2044-2058.
- Clark, D.E., 2011. What has polar surface area ever done for drug discovery? *Future Med. Chem.* 3, 469–484.
- Consonni, V., Ballabio, D., Todeschini, R., 2009. Comments on the definition of the Q^2 parameter for QSAR validation. *J. Chem. Inf. Model.* 49, 1669-1678.
- Cooper, J.A., Saracci, R., Cole, P., 1979. Describing the validity of carcinogen screening test. *Br. J. Cancer* 39, 87–89.

- Deconinck, E., Ates, H., Callebaut, N., Gyseghemb, Y.V., Heyden, Y.V., 2007. Evaluation of chromatographic descriptors for the prediction of gastro-intestinal absorption of drugs. *J. Chromatograph. A* 1138, 190-202.
- Deretey, E., Feher, M., Schmidt, J.M., 2002. Rapid Prediction of Human Intestinal Absorption. *Quant.Struct.-Act.Relat.* 21, 493- 506.
- Dietterich, T.G., 2000. Ensemble methods in machine learning. *Lect. Notes Comput. Sc.* 1857, 1–15.
- Dollery, C.T., 1999. *Therapeutic Drugs*, second ed, Churchill Livingstone, Edinburgh, UK, p 2.
- Friedman, J.H., 2002. Stochastic gradient boosting. *Comput. Stat. Data An.* 38, 367-378.
- Ghafouriana, T., Freitas, A.A., Newby, D., 2012. The impact of training set data distributions for modelling of passive intestinal absorption. *Int. J. Pharm.* 436, 711-720.
- Guerra, A., Campillo, N.E., Páez, J.A., 2010. Neural computational prediction of oral drug absorption based on CODES 2D descriptors. *Eur. J. Med. Chem.* 46, 930-940.
- Hou, T., Wang, J., Li, Y., 2007a. ADME evaluation in drug discovery. 8. The prediction of human intestinal absorption by a support vector machine. *J. Chem. Inf. Model.* 47, 2408-2415.
- Hou, T., Wang, J., Zhang, W., Xu, X., 2007b. ADME Evaluation in Drug Discovery. 7. Prediction of Oral Absorption by Correlation and Classification. *J. Chem. Inf. Model.* 47, 208-218.
- Hou, T., Wang, J., Zhang, W., Xu, X., 2007c. ADME Evaluation in Drug Discovery. 6. If the Oral Bioavailability in Human Can be Effectively Predicted by Simple Molecular Properties? *J. Chem. Inf. Model.* 47, 460-463.

- Iyer, M., Tseng, Y.J., Senese, C.L., Liu, J., Hopfinger, A.J., 2007. Prediction and mechanistic interpretation of human oral drug absorption using MI-QSAR analysis. *Mol. Pharm.* 4, 218–231.
- Lin, L.I., 1992. Assay validation using the concordance correlation coefficient. *Biometrics* 48, 599–604.
- Nuez, A.D.I., Rodríguez, R., 2008. Current methodology for the assessment of ADME-Tox properties on drug candidate molecules. *Biotechnol. Appl.* 25, 97–110.
- Obrezanova, O., Segall, M.D., 2010. Gaussian processes for classification: QSAR modeling of ADMET and target activity. *J. Chem. Inf. Model.* 50, 1053–1061.
- Palm, K., Stenberg, P., Luthman, K., Artursson, P., 1997. Polar Molecular Surface Properties Predict the Intestinal Absorption of Drugs in Humans. *Pharm. Res.* 14, 568–571.
- Pham The, H., González-Álvarez, I., Bermejo, M., Mangas Sanjuan, V., Centelles, I., Garrigues, T.M., Cabrera-Pérez, M.Á., 2011. In silico prediction of caco-2 cell permeability by a classification QSAR approach. *Mol. Inf.* 30, 376–385.
- Pino-Mejias, R., Jimenez-Gamero, M.D., Cubiles-de-la-Vega, M.D., Pascual-Acosta, A., 2008. Reduced bootstrap aggregating of learning algorithms. *Pattern Recogn. Lett.* 29, 265–271.
- Puzyn, T., Rasulev, B., Gajewicz, A., Hu, X., Dasari, T.P., Michalkova, A., Hwang, H.M., Toropov, A., Leszczynska, D., Leszczynska, J., 2011. Using nano-QSAR to predict the cytotoxicity of metal oxide nanoparticles. *Nat. Nanotechnol.* 6, 175–178.
- Reynolds, D.P., Lanevskij, K., Japertas, P., Didziapetris, R., Petrauskas, A., 2009. Ionization-specific analysis of human intestinal absorption. *J. Pharm. Sci.* 98, 4039–4054.

- Roy, K., Chakraborty, P., Mitra, I., Ojha, P. K., Kar, S., Das, R. N., 2013. Some case studies on application of r^2_m metrics for judging quality of quantitative structure-activity relationship predictions: Emphasis on scaling of response data. *J. Comput. Chem.* 34, 1071-1082.
- Schuurmann, G., Ebert, R., Chen, J., Wang, B., Kuhne, R., 2008. External validation and prediction employing the predictive squared correlation coefficient test set activity mean vs training set activity mean. *J. Chem. Inf. Model.* 48, 2140–2145.
- Shen, J., Cheng, F., Xu, Y., Li, W., Tang, Y., 2010. Estimation of ADME properties with substructure pattern recognition. *J. Chem. Inf. Model.* 50, 1034-1041.
- Shi, L.M., Fang, H., Tong, W., Wu, J., Perkins, R., Blair, R.M., Branham, W.S., Dial, S.L., Moland, C.L., Sheehan, D.M., 2001. QSAR models using a large diverse set of estrogens. *J. Chem. Inf. Comput. Sci.* 41, 186–195.
- Singh, K.P., Malik, A., Mohan, D., Sinha, S., 2004. Multivariate statistical techniques for the evaluation of spatial and temporal variations in water quality of Gomti River (India)—a case study. *Water Res.* 38, 3980–3992.
- Singh, K.P., Gupta, S., Rai, P., 2013a. Predicting acute aquatic toxicity of structurally diverse chemicals in fish using artificial intelligence approaches. *Ecotox. Environ. Safe.* 95, 221–233.
- Singh, K.P., Gupta, S., Rai, P., 2013b. Predicting carcinogenicity of diverse chemicals using probabilistic neural network modeling approaches. *Toxicol. Appl. Pharmacol.* 272, 465-475.
- Singh, K.P., Gupta, S., Mohan, D., 2014. Evaluating influences of seasonal variations and anthropogenic activities on alluvial groundwater hydrochemistry using ensemble learning approaches. *J. Hydrol.* 511, 254-266.

- Singh, K.P., Gupta, S., 2014. Nano-QSAR modeling for predicting biological activity of diverse nanomaterials. *RSC Adv.* 4, 13215-13230.
- Steinbeck, C., Han, Y., Kuhn, S., Horlacher, O., Luttmann, E., Willighagen, E., 2003. The chemistry development kit (CDK): an open-source java library for chemo- and bioinformatics. *J. Chem. Inf. Comput. Sci.* 43, 493–500.
- Suenderhauf, C., Hammann, F., Maunz, A., Helma, C., Huwyler, J., 2011. Combinatorial QSAR modeling of human intestinal absorption. *Mol. Pharm.* 8, 213-224.
- Talevi, A., Goodarzi, M., Ortiz, E.V., Duchowicz, P.R., Bellera, C.L., Pesce, G., Castro, E.A., Bruno-Blanch, L.E., 2011. Prediction of drug intestinal absorption by new linear and non-linear QSPR. *Eur. J. Med. Chem.* 46, 218-228.
- Tropsha, A., Golbraikh, A., Cho, W.J., 2011. Development of kNN QSAR models for 3-arylisoquinoline antitumor agents. *Bull. Korean Chem. Soc.* 32, 2397–2404.
- Wessel, M.D., Jurs, P.C., Tolan, J.W., Muskal, S.M., 1998. Prediction of Human Intestinal Absorption of Drug Compounds from Molecular Structure. *J. Chem. Inf. Comput. Sci.* 38, 726-735.
- Xu, J., Wang, L., Liu, L., Bai, Z., Wang, L., 2011. QSPR study of the absorption maxima of azobenzene dyes. *Bull. Korean Chem. Soc.* 32, 3865-3872.
- Yan, A., Wang, Z., Cai, Z., 2008. Prediction of human intestinal absorption by GA feature selection and support vector machine regression. *Int. J. Mol. Sci.* 9, 1961-1976.
- Yang, P., Yang, Y.H., Zhou, B.B., Zomaya, A.Y., 2010. A review of ensemble methods in bioinformatics. *Curr. Bioinforma.* 5, 296–308.
- Zakeri-Milani, P., Tajerzadeh, H., Islambolchilar, Z., Barzegar, S., Valizadeh, H., 2006. The relation between molecular properties of drugs and their transport across the intestinal membrane. *DARU* 14, 164–171.

- Zhao, Y.H., Le, J., Abraham, M.H., Hersey, A., Eddershaw, P.J., Luscombe, C.N., Boutina, D., Beck, G., Sherborne, B., Cooper, I., Platts, J.A., 2001. Evaluation of Human Intestinal Absorption Data and Subsequent Derivation of a Quantitative Structure-Activity Relationship (QSAR) with the Abraham Descriptors. *J. Pharm. Sci.* 90, 749-784.
- Zhao, C.Y., Zhang, H.X., Zhang, X.Y., Liu, M.C., Hu, Z.D., Fan, B.T., 2006. Application of support vector machine (SVM) for prediction toxic activity of different data sets. *Toxicology* 217, 105-119.

Figure Captions

Figure 1: Histogram of the HIA (%) values of the chemicals in complete dataset.

Figure 2: Histogram of the TSI values of the chemicals in complete dataset.

Figure 3: Plot of the contributions of the selected descriptors in qualitative SAR models.

Figure 4: Plot of the contributions of the selected descriptors in quantitative SAR models.

Figure 5: Plot of the experimental and predicted values of HIA% in training, validation and test data for (a) GBT-SAR, and (b) BDT-SAR models.

Figure 6: Plot of the validation criteria parameters for (a) GBT-SAR, and (b) BDT-SAR models in various datasets.

Figure 7: Williams plot for the quantitative (a) GBT-SAR, and (b) BDT-SAR model.

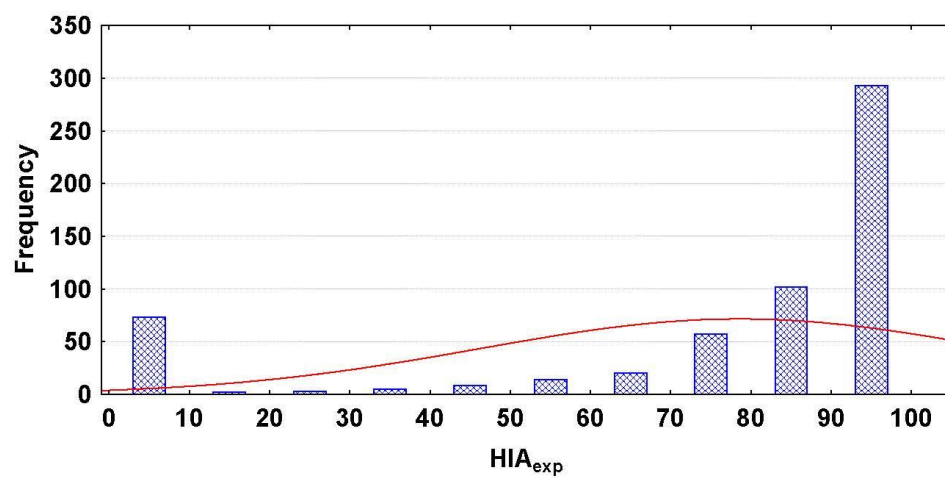


Figure 1

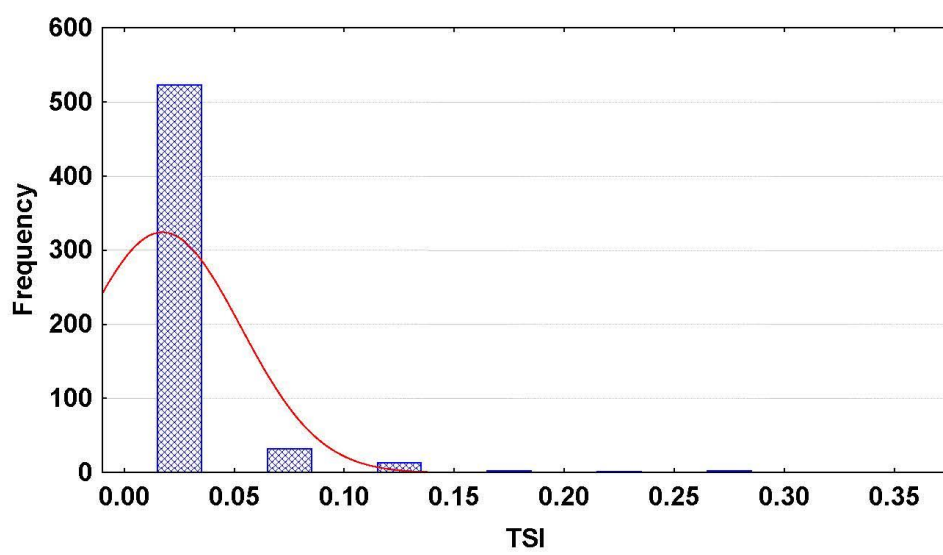


Figure 2

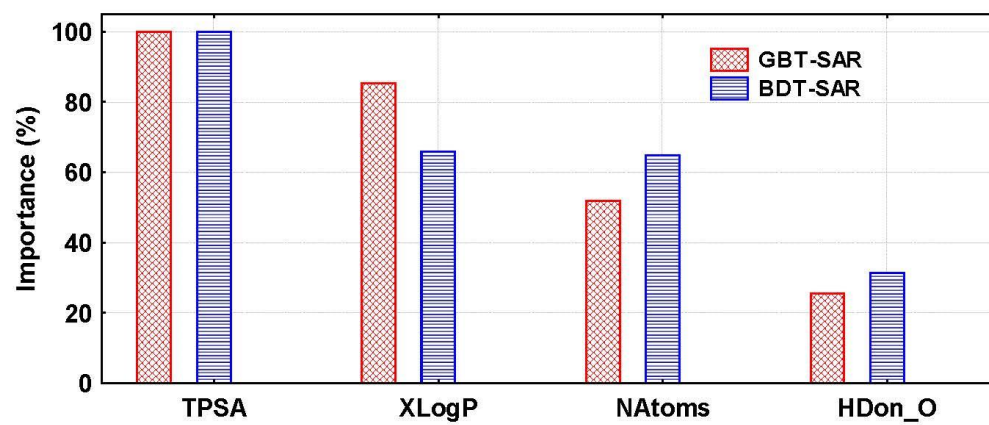


Figure 3

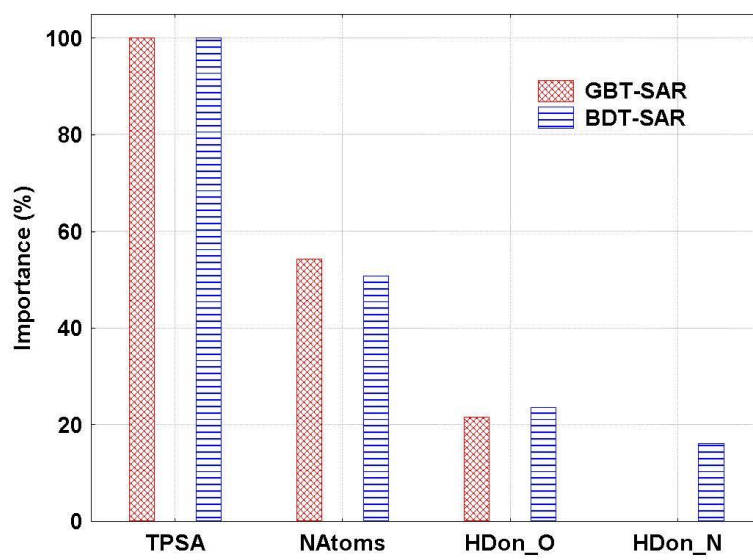


Figure 4

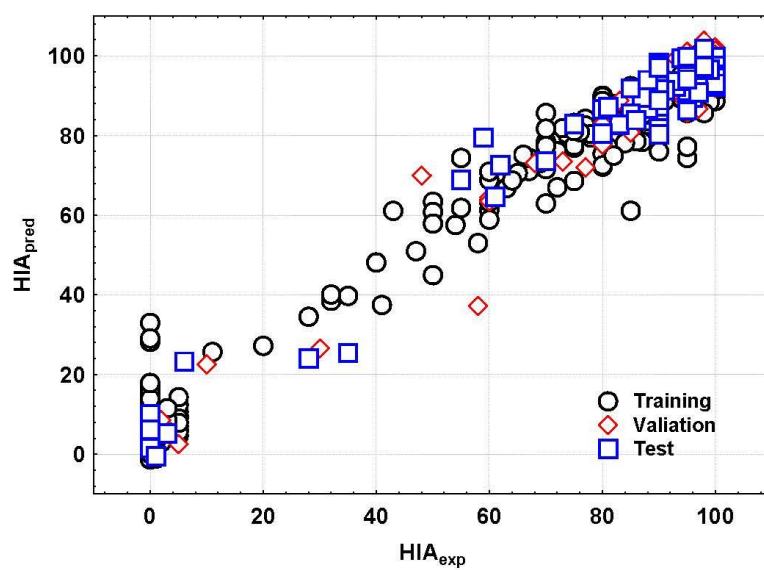


Figure 5a

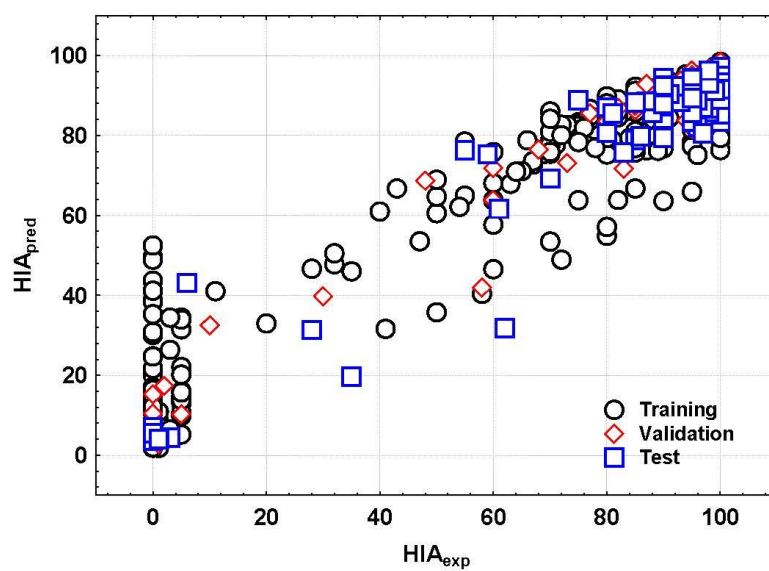


Figure 5b

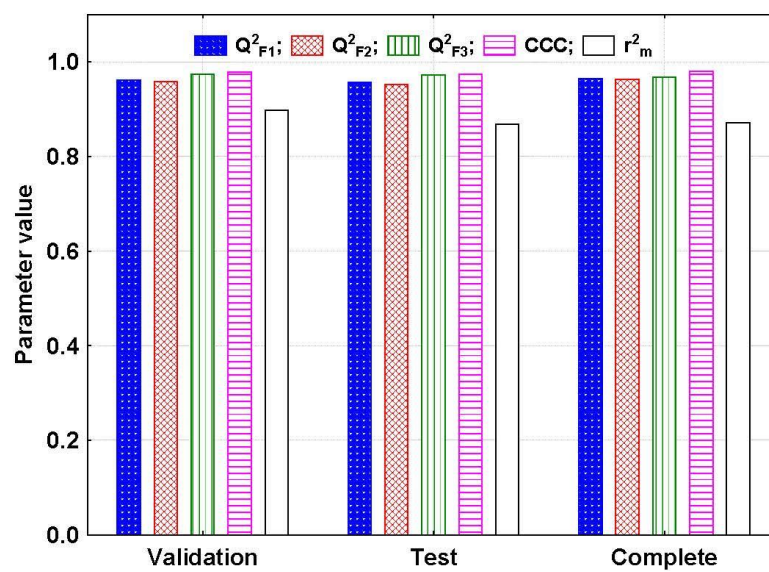


Figure 6a

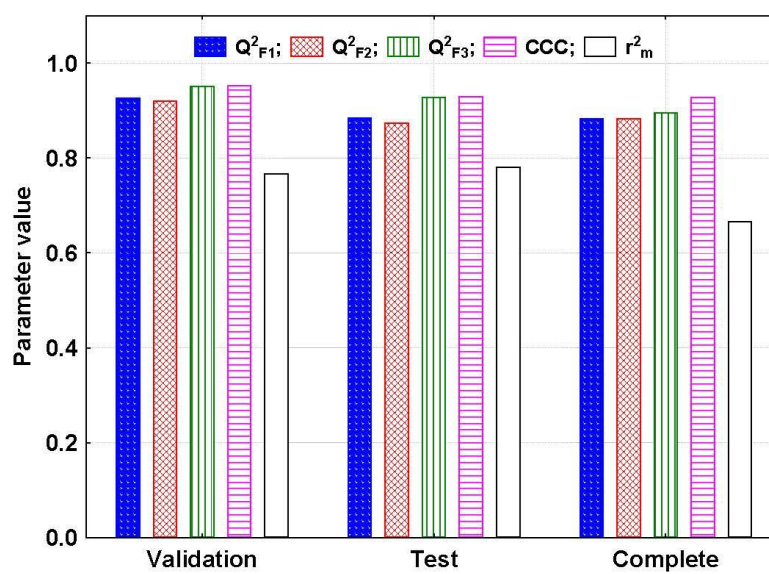


Figure 6b

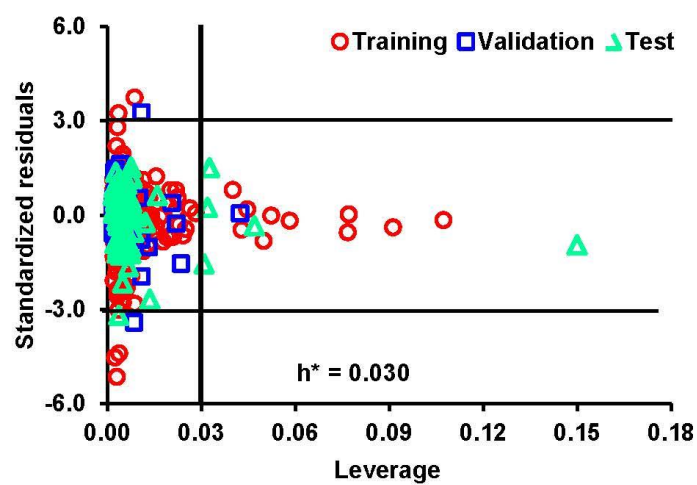


Figure 7a

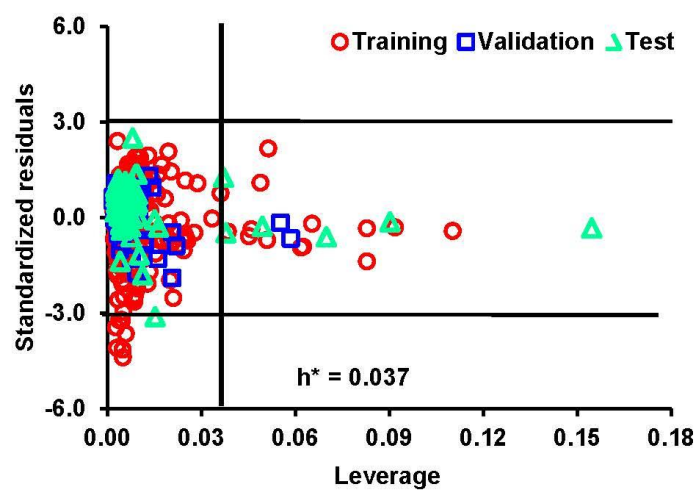


Figure 7b

Tables

Table 1: Description of selected descriptors in qualitative and quantitative SAR models

Descriptor Symbol	Model	Class	Description
HDon_O	R, C	Constitutional	Number of hydrogen bonding donors derived from the sum of O-H groups only in the molecule
HDon_N	R	Constitutional	Number of hydrogen bonding donors derived from the sum of N-H groups only in the molecule
NAtoms	R, C	Constitutional	Number of all atoms in the molecule
XLogP	C	Constitutional	Octanol partition coefficient [in log units] of the molecule following the XLogP approach
TPSA	R, C	Topological	Topological polar surface area in [\AA^2] of the molecule derived from the polar 2D fragments

C- Classification; R-regression

Table 2: Basic statistics of the dataset

Descriptors	Min	Max	Mean	^aSD	^bCoV
HDon_N	0.00	12.00	1.07	1.68	157.47
HDon_O	0.00	13.00	1.00	1.56	156.01
XlogP	[-7.48]	7.82	2.10	2.19	104.20
TPSA	0.00	530.49	75.02	54.76	72.99
Natoms	9.00	176.00	43.87	18.60	42.40
HIA (%)	0.00	100.00	78.19	32.19	41.18

^aSD-standard deviation; ^bCoV-coefficient of variation

Table 3: Performance parameters for qualitative SAR models

Model	Data sub-sets	Sensitivity (%)	Specificity (%)	Accuracy (%)	MCC
LDA-SAR	Training	75.00	90.24	89.33	0.45
	Validation	100.00	95.00	95.40	0.78
	Test	100.00	93.75	94.25	0.74
	Complete	84.21	91.47	90.99	0.55
GBT-SAR	Training	100.00	99.71	99.75	0.99
	Validation	100.00	98.70	98.85	0.95
	Test	100.00	97.40	97.70	0.90
	Complete	100.00	99.20	99.31	0.97
BDT-SAR	Training	100.00	99.71	99.75	0.99
	Validation	100.00	98.70	98.85	0.95
	Test	100.00	97.40	97.70	0.90
	Complete	100.00	99.20	99.31	0.97

Table 4: Basic statistics of the selected descriptors in two different categories of chemicals used in classification models.

Descriptors	Poor HIA absorption		Good HIA absorption	
	Range	Mean	Range	Mean
HDon_O	0.00 - 13.00	2.78	0.00 - 6.00	0.72
XlogP	[-7.48] - 6.78	-0.16	[-3.34] - 7.82	2.45
TPSA	0.00 - 530.49	144.75	3.24 - 201.85	64.13
NAtoms	15.00 - 176.00	56.77	9.00 - 136.00	41.85

Table 5: Comparison with previously published qualitative SAR models

Models	Number of compounds	Number of descriptors	Type of descriptors	Accuracy (%)	References
SVM	480 (Tr) 98 (T)	-	CD, TD	97.3 (Tr) 98.0 (T)	Hou et al., 2007a
ANN	202 (Tr) 165 (T)	-	-	79.0 (Tr) 75.0 (T)	Guerra et al., 2010
Probit-PLS	158 (Tr) 67 (T)	166	2D	89.0 (Tr) 84.0 (T)	Obrezanova and Segall , 2010
DT	158 (Tr) 67(T)	166	2D	96.0 (Tr) 85.0 (T)	Obrezanova and Segall , 2010
SVM-RBF	480 (Tr) 98 (T)	307	FP4	98.5 (Tr) 99.0 (T)	Shen et al., 2010
DTI	458	14	CD, TD, GD, OH	91.0	Suenderhauf et al., 2010
ANN	458	14	CD, TD, GD, OH	79.0	Suenderhauf et al., 2010
GBT-SAR	403 (Tr) 87 (V) 87 (Tt)	4	CD, TD	99.75 (Tr) 98.85 (V) 97.70 (T)	Present work
BDT-SAR	403 (Tr) 87 (V) 87 (T)	4	CD, TD	99.75 (Tr) 98.85 (V) 97.70 (T)	Present work

SVM support vector machine; ANN artificial neural networks; PLS partial least square; DT decision tree; DTI decision tree induction; RBF radial basis function; CD-constitutional descriptors ; TD Topological descriptors; 2D two dimensional; FP4 finger prints patterns ; GD geometrical descriptors; OH others; Tr training; V validation; T test.

Table 6: Performance parameters for quantitative SAR models

Experimental/ Model	Data sub-sets	Mean	Standard deviation	RMSE	R²
Experimental	Training	75.80	34.17	-	-
	Validation	83.40	26.84	-	-
	Test	84.05	26.01	-	-
	Complete	78.19	32.19	-	-
GBT-SAR	Training	75.82	30.87	6.39	0.972
	Validation	83.54	25.84	5.50	0.958
	Test	84.06	24.23	5.71	0.953
	Complete	78.23	29.43	6.16	0.968
BDT-SAR	Training	75.52	26.60	12.00	0.906
	Validation	81.46	22.38	7.53	0.945
	Test	80.32	22.96	9.22	0.898
	Complete	77.14	25.56	11.04	0.908
MLR-SAR	Training	75.80	17.40	29.37	0.259
	Validation	78.17	18.42	20.53	0.447
	Test	72.99	29.93	20.11	0.683
	Complete	75.73	19.94	26.96	0.308

Table 7: Comparison with previously published quantitative models correlating HIA of the chemicals

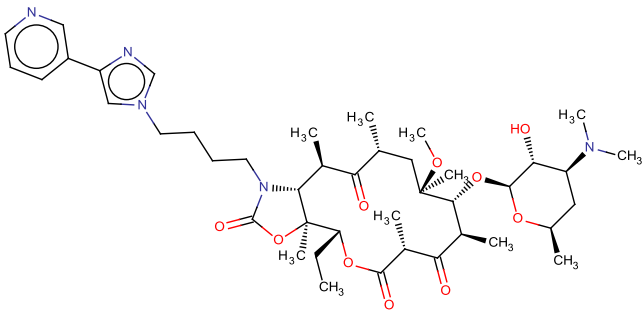
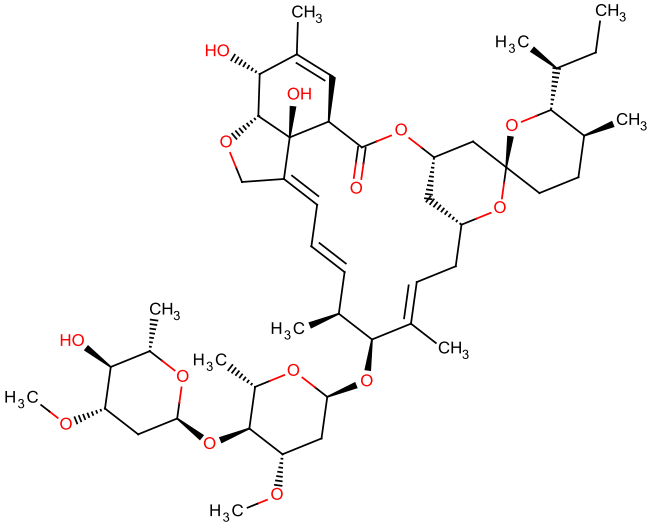
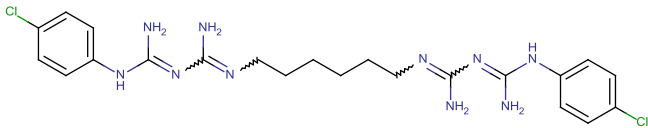
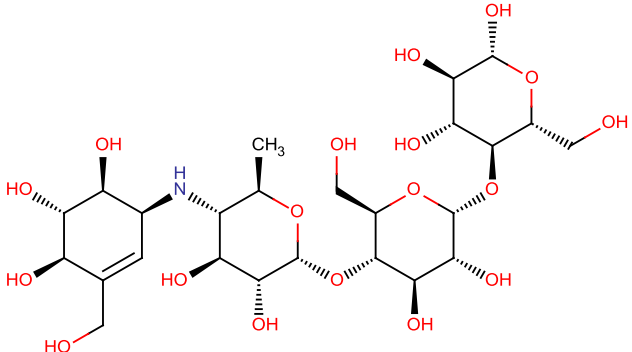
Models	Number of compounds	Number of descriptors	Type of descriptors	R ²	References
GFA	455 (Tr) 98 (T)	-	CD, OH	0.706 (Tr) 0.810 (T)	Hou et al., 2007
MLR	67	8	CD, TD, OH	0.430 (Tr)	Deconinck et al., 2007
Nonlinear-MARS	67	4	CD, ChrD, OH	0.820 (Tr)	Deconinck et al., 2007
MI-QSAR	188	-	CD, OH	0.730	Iyer et al., 2007
PLS	380 (Tr) 172 (T)	9	Adriana, CD, GD, TD, PD	0.548 (Tr) 0.689 (T)	Yan et al., 2008
SVM	380 (Tr) 172(T)	9	Adriana, CD, GD, TD, PD	0.656 (Tr) 0.774 (T)	Yan et al., 2008
QSAR	567(Tr)	4	CD, OH	0.930 (Tr)	Reynolds et al., 2009
ANN	37	3	TD	0.930 (Tr)	Guerra et al., 2010
MLR	90(Tr) 30 (V) 40 (T)	4	FgD, PD, ChD, OH	0.659 (Tr) 0.654 (V) 0.551 (T)	Talevi et al., 2011
MLR	496	5	CD	0.533	Ghafouriana et al., 2012
GBT-SAR	403(Tr) 87(V) 87 (T)	3	CD, TD	0.972 (Tr) 0.958 (V) 0.953 (T)	Present work
BDT-SAR	403(Tr) 87(V) 87 (T)	4	CD,TD	0.906 (Tr) 0.945 (V) 0.898 (T)	Present work

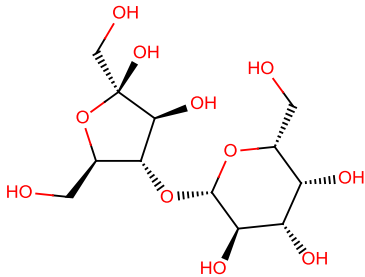
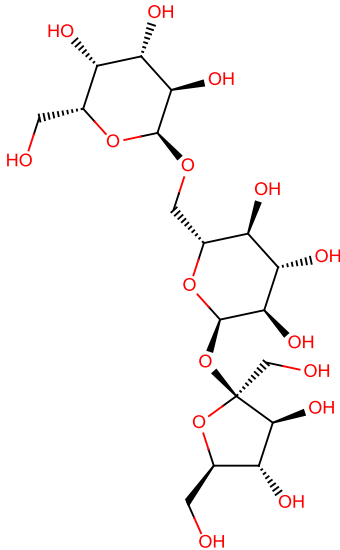
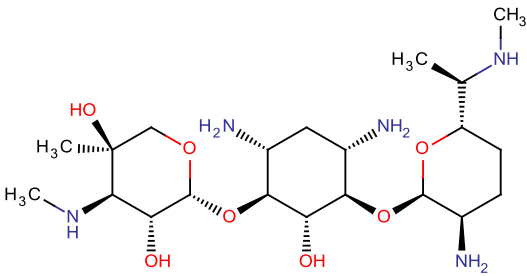
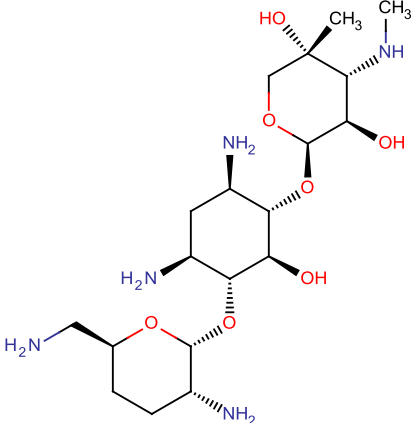
GFA genetic function approximation; MLR multiple linear regression; PLS partial least square; SVM support vector machine; ANN artificial neural networks; CD-constitutional descriptors; TD Topological descriptors; ChrD chromatographic descriptors; GD geometrical descriptors; PD Physico-chemical descriptors; FgD fragment descriptors; ChD charge descriptors; MARS multivariate adaptive regression splines; MI Membrane-interaction; OH others; Tr Training; V validation; T test.

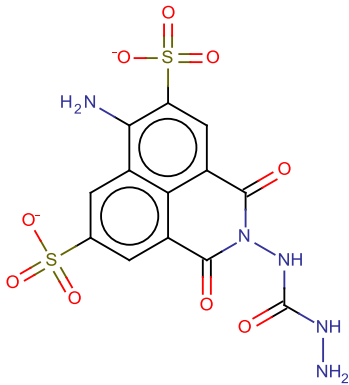
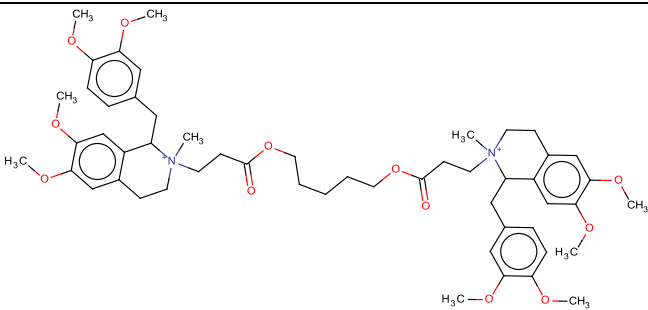
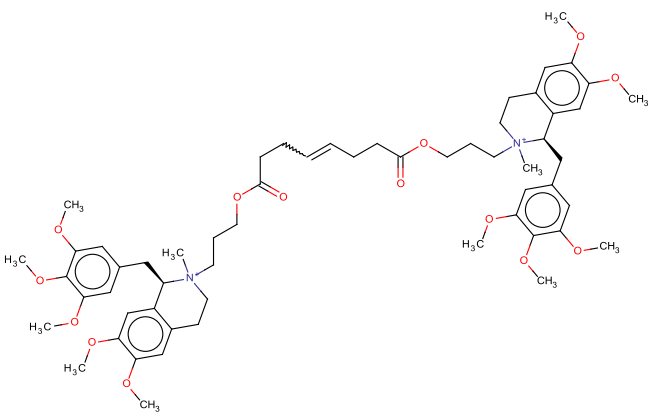
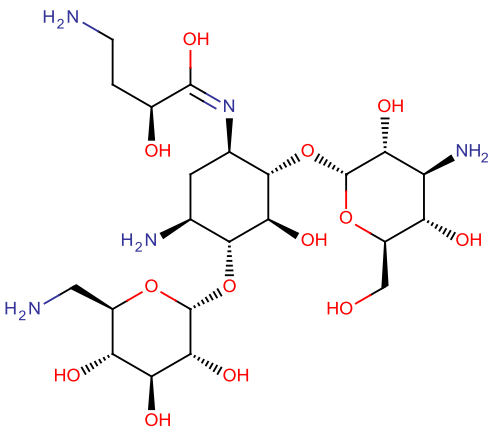
Table 8: AD of the qualitative and quantitative SAR models

Descriptors	Training set	Validation set	Test set
Qualitative model			
HDon_O	0.00 - 13.00	0.00 - 11.00	0.00 - 7.00
XLogP	[-7.48] - 7.82	[-6.16] - 6.61	[-6.14] - 6.42
TPSA	0.00 - 530.49	3.24 - 336.43	0.00 - 282.61
NAtoms	9.00 - 176.00	18.00 - 100.00	15.00 - 139.00
Out of AD	-	-	-
Quantitative model			
HDon_N	0.00 - 10.00	0.00 - 10.00	0.00 - 12.00
HDon_O	0.00 - 13.00	0.00 - 7.00	0.00 - 10.00
TPSA	0.00 - 335.43	0.00 - 336.43	6.48 - 530.49
NAtoms	9.00 - 154.00	17.00 - 97.00	12.00 - 176.00
Out of AD	-	1	3

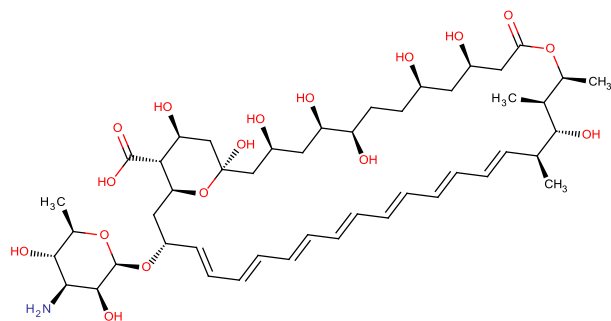
Table 9: Out of AD compounds in quantitative SAR models

S.No.	Model	Chemical Name	Chemical Structure
High Leverage Compounds			
1	GBT, BDT	Telithromycin	
2	GBT, BDT	Ivermectin	
3	BDT	Chlorhexidine	
4	GBT, BDT	Acarbose	

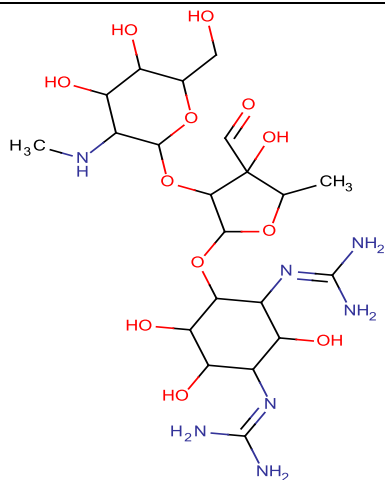
5	GBT, BDT	Lactulose	
6	GBT, BDT	Raffinose	
7	BDT	Gentamicin_C1	
8	BDT	Gentamicin_C1a	

9	GBT, BDT	Lucifer_Yellow_CH	Li^+ Li^+	
10	GBT, BDT	Atracurium		
11	GBT, BDT	Mivacurium		
12	GBT, BDT	Amikacin		

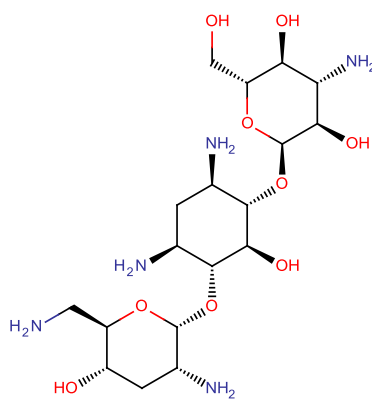
13	GBT, BDT	Amphotericin_B
----	----------	----------------



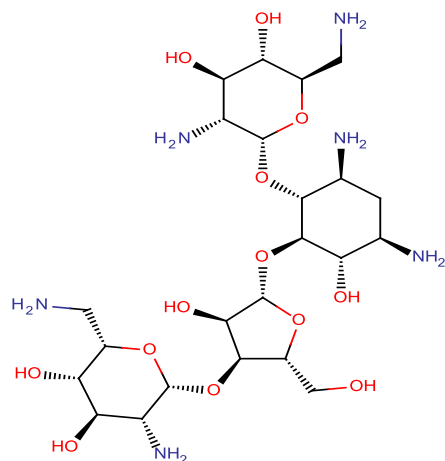
14	GBT, BDT	Streptomycin
----	----------	--------------



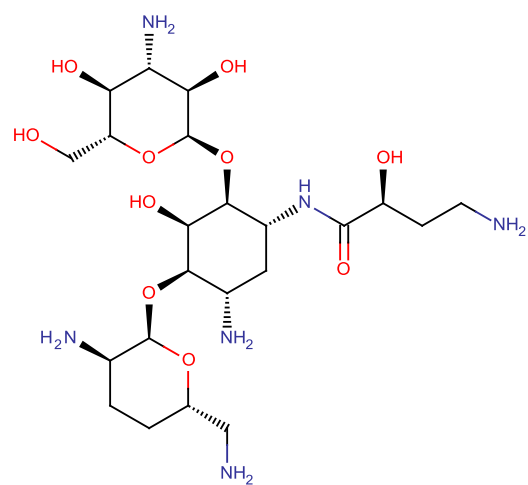
15	BDT	Tobramycin
----	-----	------------



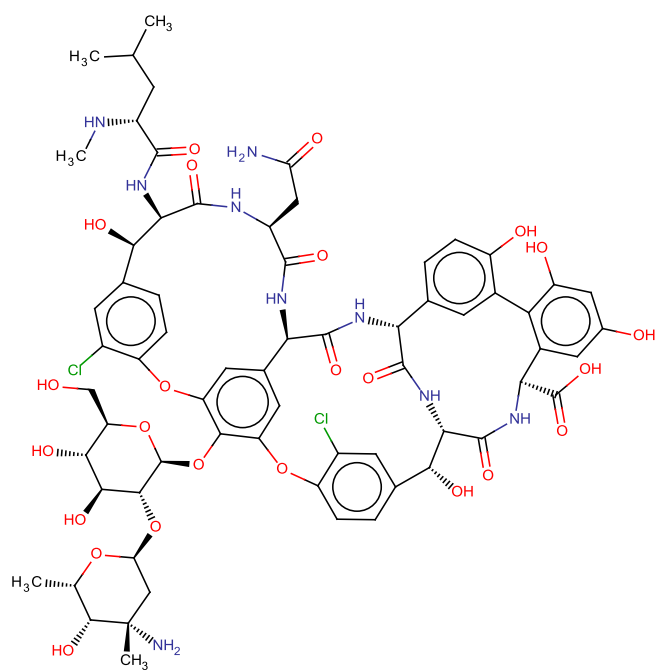
16 GBT, BDT Neomycin

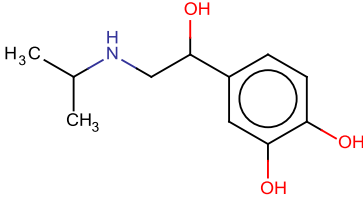
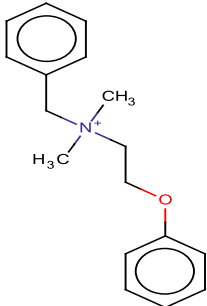
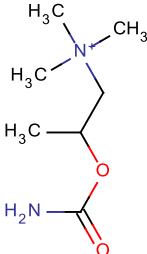
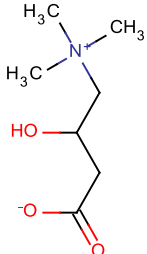
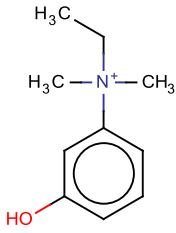
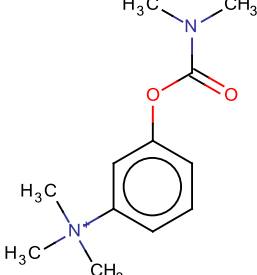


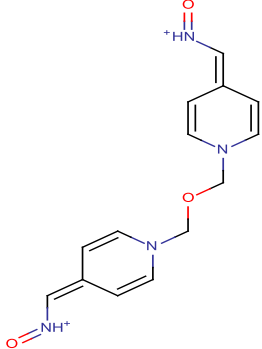
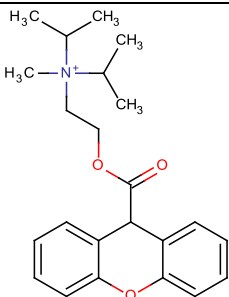
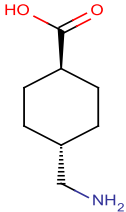
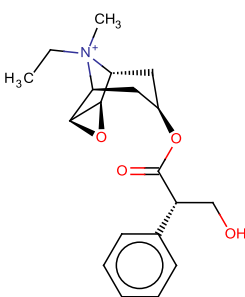
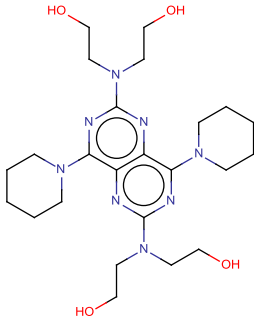
17 GBT, BDT Arbekacin



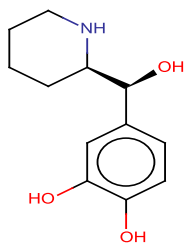
18 GBT, BDT Vancomycin



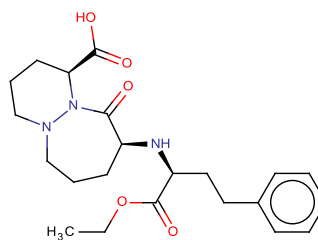
2	GBT	Isoproterenol	
3	BDT	Bephenium	
4	BDT	Bethanechol	
5	BDT	Carnitine	
6	BDT	Edrophonium	
7	GBT, BDT	Neostigmine	

8	GBT, BDT	Obidoxime	
9	BDT	Propantheline	
10	GBT	Tranexamic_acid	
11	GBT, BDT	Oxitropium	
12	GBT	Dipyridamole	

13 GBT Rimiterol



14 GBT Cilazapril



15 BDT Tiludronic_acid

