Classifying bio-concentration factor with random forest algorithm, influence of the bio-accumulative vs. non-bio-accumulative compound ratio to modelling result, and applicability domain for random forest model£

G. Piir, S. Sild* and U. Maran

Institute of Chemistry, University of Tartu, Tartu, Estonia

(Received 10 June 2014; in final form 3 August 2014)

In environmental risk assessment, the bio-concentration factor (BCF) is a widely used parameter in the estimation of the bio-accumulation potential of chemicals. BCF data often have an uneven distribution of classes (bio-accumulative vs. non-bio-accumulative), which could severely bias the classification results towards the prevailing class. The present study focuses on the influence of uneven distribution of the classes in training phase of Random Forest (RF) classification models. Three different training set designs were used and descriptors selected to the models based on the occurrence frequency in RF trees and considering the mechanistic aspects they reflect. Models were compared and their classification performance was analysed, indicating good predictive characteristics (sensitivity = 0.90 and specificity = 0.83) for the balanced set; also imbalanced sets have their strengths in certain application scenarios. The confidence of classifications was assessed with a new schema for the applicability domain that makes use of the RF proximity matrix by analysing the similarity between the predicted compound and the training set of the model. All developed models were made available in the transparent, accessible and reproducible way in QsarDB repository (http://dx.doi.org/10.15152/QDB.116).

Keywords: bio-concentration factor; random forest; classification; applicability domain; imbalanced data set; proximity matrix

1. Introduction

Bio-accumulation is the process where chemicals from the environment accumulate in an organism and reach concentrations greater than in the environment. The laboratory equivalent for this process is bio-concentration, which is characterized by the bio-concentration factor (BCF), defined as the ratio of the concentration of a chemical in an organism and that in the surrounding environment at steady state. BCF is widely used to assess the environmental risks of chemicals. The current European regulation on Registration, Evaluation, Authorisation and Restriction of Chemical substances (REACH) [1] makes it mandatory, among other properties, to assess the rate of bio-accumulation for chemicals that are produced or imported in amounts greater than 10 tonnes in a year. Only for those chemicals produced or imported in amounts greater than 100 tonnes in a year are expensive and time-consuming experimental

*Corresponding author. Email: sulev.sild@ut.ee
£Presented at the 16th International Workshop on Quantitative Structure-Activity Relationships in Environmental and Health Sciences (QSAR2014), 16–20 June 2014, Milan, Italy.

© 2014 Taylor & Francis
bio-concentration measurements are required. This forms the basis for the use of faster and cost-effective quantitative structure–activity relationship (QSAR) models to predict the bio-accumulation potential of chemicals and nurtures the research of predictive models.

Over the years many theoretical models have been developed to predict BCF [2–4]. Most of them are regression models, though in recent years, a few classification models have been published. One possible reason might be that, for the regression models, normal distribution of the data set is expected and most of the BCF data sets have normal distribution. At the same time, the distribution of bio-accumulative (B) and non-bio-accumulative (nB) within normal distribution of compounds is not equal. This makes building classification models for BCF problematic because it most likely influences the development of classification models.

To the best of authors’ knowledge the first classification model for BCF was developed fairly recently when Sun et al. [5] used a data set of 238 equally distributed diverse organic molecules to separate bio-accumulative and non-bio-accumulative molecules with four statistical learning approaches. Most of the tested algorithms correctly classified B-compounds slightly better (86.6–89.6%) than nB-compounds (81.8–87.4%) and had relatively similar overall prediction accuracy (84.5–87.7%). This was followed by work by Nendza and Müller [6] who developed a binary classification schema to identify nB-compounds based on the cut-off criteria of molecular weight and octanol–water partition coefficient (log P). The proposed schema securely classified 30–40% of the compounds with low bio-accumulation potential. Later Nendza and Herbst [7] improved the schema by developing a decision tree based on five physico-chemical properties (log P, ionisation at pH 7, Henry’s law constant, second order hydrolysis rate and biodegradability). The enhanced schema securely identified more than 50% of the compounds with low bio-accumulation potential. Very recently Strempel et al. [8] utilised two physico-chemical properties (octanol–water distribution coefficient and biotransformation half-life) with conditional inference trees to separate B-compounds and nB-compounds. They managed to identify 87% of nB-compounds and 39.2% of B-compounds from the entire data set with high certainty using four distinct rules. Additionally they built a Random Forest (RF) based regression model and used quantitative values from this model and cut-off values for classification. They were able to correctly classify 98% of the nB-compounds and 86% of the B-compounds from the test set of 140 compounds. Both methods classified nB-compounds better because their data set was imbalanced towards nB-compounds. Quantitative values for classification of BCF were also used by Fernández et al. [9]. They used the consensus of five different regression models with a continuous Bayesian formulation to improve the classification of compounds to nB-compounds, B-compounds and very bio-accumulative (vB) compounds.

Most previous studies on the classification of BCF show better classification results for nB-compounds than for B-compounds because of the imbalance of the data sets. There are several methods to deal with the data set imbalance in classification modelling and many of them are reviewed by Galar et al. [10]. The main goal of this study was to develop a classification model for BCF by addressing the imbalance of the data set. For this a large imbalanced data set (four to one in favour of nB-compounds) and the RF algorithm, together with descriptor selection, were used to study the influence of balance between classes in the context of BCF. Also a new applicability domain (AD) approach is introduced for the RF based models, providing quantitative AD measures based on the proximity matrix of the RF model, where expert judgement can introduce a suitable confidence level for predictions. Finally the robustness of models was tested to make sure that the obtained results were not accidental and mechanistic interpretation of descriptors was provided.
2. Methods

2.1 Data set

The initial data set was adapted from the literature [11] and contained 1036 compounds where experimental values were measured inside the pH limits defined by guidelines in the REACH legislation [1]. All compounds were converted from SMILES to InChIKeys using JChem for Excel [12]. The data set was filtered in three rounds removing compounds that did not lend themselves for modelling.

In the first round all CAS registry numbers from the original data sources were compared against the CAS numbers in the PubChem database [13]. If a match was found, the InChIKeys from PubChem database were compared against generated InChIKeys. A total of 967 compounds matched in both comparisons and they remained in the data set.

In the second round compounds that did not pass the first round (69 compounds) were manually checked with a SciFinder [14] substance search. This helped to identify mixtures (ID: 503, 808) and compounds (ID: 238, 521, 824) with incomplete definition that were discarded. Non-matching compounds included also those with no CAS registry number (ID: 262, 263, 702, 902) and compounds where the CAS registry number and structure did not match (ID: 404, 726) that were discarded. The second round cleared 58 compounds and they were reinstated to the data set.

In the third filtering round the resulting dataset of 1025 compounds was refined further by implementing the following adjustments. Firstly, two duplicates (ID: 341 versus 495, and 493 versus 715) were identified by comparing InChiKeys and removed. Secondly, compounds containing tin (ID: 6, 160, 175, 177, 311, 397, 501, 517, 636, 725, 817) and silicon (ID: 338, 470, 613, 627, 695) were removed from data set because log P calculations for those compounds can be erroneous. Thirdly, salts (ID: 216, 252, 640, 668, 671, 966) in the data set were converted to their acidic form. As a consequence the final data set used for this study contained 1007 compounds and the range of logarithmic BCF (log BCF) values was from -2.00 to 6.43.

The threshold for a compound being bio-accumulative is not uniquely defined in different regulations. For example, the REACH legislation [1] defines compounds as bio-accumulative when the log BCF threshold value is 3.3 and more. The US Environmental Protection Agency sets the corresponding threshold at 3.0 [15]. Both regulations assign log BCF values for highly bio-accumulative compounds to 3.7.

According to different sources [16,17], measurements of BCF have an error from 0.35 to 0.75 log units. Considering this we decided to group our data set into two: (i) nB-compounds with log BCF values less than 3.0; and (ii) B-compounds with log BCF values equal or greater than 3.0. This lead to the data set with 798 nB-compounds and 209 B-compounds.

The pruned data set was divided into training and validation sets. For this the entire data set was sorted according to experimental log BCF values and then every third compound was inserted to the validation set and remaining compounds were considered as training set. This made a final training set of 673 compounds and a final validation set with 334 compounds.

The training set consisted of 141 B-compounds and 532 nB-compounds. Distribution of the B- and nB-compounds in the validation set was 68 and 266, respectively.

2.2 Calculation of molecular descriptors

Chemical structures are numerically characterized with the molecular descriptors. Altogether 770 one-dimensional (1D) and two-dimensional (2D) descriptors were calculated with the
PaDEL-Descriptor (version 2.18) descriptor calculator. This set was complemented with three \( n \)-octanol–water partition coefficients (log P) calculated using XLogP3 (version 3.2.2) [19], EPISuite (version 4.11) [20] and VCCLab’s AlogPS 2.1 web service [21,22].

2.3 Model development

In our previous study [16], it was shown that consensus modelling improves prediction accuracy. The downside of consensus modelling is the fact that one must build many different models which can be time-consuming and makes it difficult to achieve consistent interpretation. Keeping the classification problem in mind and looking for suitable approaches led to the RF algorithm which, by design, can be also considered a consensus approach.

2.3.1 Random Forest algorithm

The Random Forest [23] algorithm is a tool for regression and classification analysis, and is gaining popularity in the field of QSAR. The RF algorithm builds on the ensemble of decision trees, where final quantitative or qualitative decision is made by consensus of all the trees. A brief overview of the working principles of the RF algorithm is given below.

The RF algorithm utilizes predictions of \( N \) decision trees (in the current work 501). By default, RF draws from the given data set of \( n \) compounds a bootstrap sample, which usually contains two-thirds of the compounds from the initial data set. A third of the compounds that are left out are called the ‘out-of-bag’ (OOB) sample. For each drawn sample a tree is grown as follows. To select the best descriptor to split the sample, a randomly selected subset of descriptors \( m_{try} \) is used and choice is made among those descriptors. The default value for \( m_{try} \) is different in regression and classification analysis. For the regression analysis, RF selects randomly \( p/3 \) of the descriptors and for the classifications analysis the number of descriptors is equal to square root of \( p \), where \( p \) is total number of descriptors. Each tree is grown until it is impossible to make another split. The final prediction for the regression model is made by averaging predictions from all trees. In the classification model the final decision is made by using the majority of votes in all trees.

All the compounds that did not take part of the tree growing process (OOB compounds) are used as an internal validation set to get the estimated error of the model. Each OOB compound is entered into the grown tree to obtain a prediction. These predictions are kept and the majority of predictions are compared with real values to get the estimated error. All statistical parameters in the tables for training sets are based on the OOB predictions.

These OOB compounds are also used to estimate the importance of each descriptor in the model. This procedure is based on a permutation test and uses a mean decrease in the accuracy to estimate descriptor importance. The procedure works as follows. Firstly, for each tree, prediction for each OOB compound is made and correct predictions are counted. In the second step, values of the descriptor are permuted and predictions for each OOB compound are repeated. Finally, the importance score for descriptor is calculated by averaging the difference between correctly predicted compounds from permuted and original OOB compounds over all trees.

All the RF (version 4.6-7) calculations were made with the R (version 3.0.2) [24] statistical package. Since RF algorithm uses random sampling in different calculation steps, one gets different results in each run. This may be problem when the reproducibility is a concern, but in general, the RF algorithm gives similar results irrespective of random number. To make sure the model is reproducible one must verify it by repeating the modelling process multiple
times and comparing the results for the variance. The problem of exact reproduction of the final published model can be solved by setting the initial seed number of the random number generator to a fixed number (so-called ‘random seed’) to regenerate exactly the same sequence of random numbers for the reproduction of the model. In this work the final models were built using an arbitrary number 1234 as the random seed.

2.3.2 Selection of descriptors
All the developed models went through a multistep descriptor selection procedure. Firstly, experimental log P values were gathered using EPISuite and these values were correlated with all calculated log P values. Based on those correlations XlogP3 was chosen because it had highest squared correlation coefficient ($R^2 = 0.96$). Secondly, descriptors with zero variance and highly correlated ($R^2 > 0.9$) descriptors were removed from the descriptor pool. The remaining 248 descriptors were analysed for the relevance to the tree growing and the ones that did not contribute were excluded according to the following procedure. At first 100 models were built with randomly split data set. From every model, the 20 most important descriptors were selected and the occurrence of the descriptors were analysed. Descriptors present in the top 20 more than two-thirds of the times were chosen. From those descriptors only descriptors that had mechanistic relevance regarding the BCF were selected for the model building. For final models, $m_{tr}$ was set to match the number of descriptors in the final selection. This means that, in the tree growing process, the best descriptor was chosen at each split from all the remaining descriptors.

2.3.3 Distribution of log BCF in the data set
In the real world the distribution of compounds between classes is very likely imbalanced, making solving classification problems difficult. This is because imbalanced data can severely bias the results towards the majority class and this also influences the performance of the RF algorithm. The data set in this study is also strongly imbalanced having a four to one ratio in the favour of nB-compounds. In order to understand how class distributions influence classification results, three different class distribution schemas were studied. For all three schemas in the training process, the RF algorithm was forced to pick only certain number of compounds from both classes as an input sample. The first schema is imbalanced towards nB-compounds and uses default parameters for random sample selection (results in section 3.1). The second schema is balanced, where an equal amount of B and nB-compounds were selected in each tree growing (results in section 3.2). The third schema is imbalanced towards B-compounds, which means that more B-compounds than nB-compounds were selected to the growing of each tree (see section 3.3).

2.4 Performance metrics
The performance of the classification models is measured with: sensitivity (Equation (1)); specificity (Equation (2)); accuracy (Equation (3)); positive predictive value (PPV) (Equation (4)); and negative predictive value (NPV) (Equation (5)). Values for these terms are calculated according to the number of true positive (TP), true negative (TN), false positive (FP) and false negative (FN) predictions. For more detailed analysis of the TP, TN, FP and FN predictions, the so-called confusion matrix (also known as contingency table or an error
matrix) is used, which simplifies information gathering of the performance of an algorithm. Each row in the confusion matrix shows how many instances are in the actual class and each column describes instances in a predicted class. In this work positive predictions correspond to the B-compounds and negative predictions to the nB-compounds. Figure 1 shows how these calculations are related to the confusion matrix of a classification model.

\[
\text{Sensitivity} = \frac{TP}{TP + FN} \tag{1}
\]
\[
\text{Specificity} = \frac{TN}{TN + FP} \tag{2}
\]
\[
\text{Accuracy} = \frac{(TP + TN)}{(TP + FP + TN + FN)} \tag{3}
\]
\[
\text{PPV} = \frac{TP}{(TP + FP)} \tag{4}
\]
\[
\text{NPV} = \frac{TN}{(TN + FN)} \tag{5}
\]

The sensitivity shows the proportion of the B-compounds that are correctly classified. Contrarily to the sensitivity, the specificity measures the proportion of correctly classified nB-compounds. All correctly classified compounds (TP and TN) divided by the number of compounds in the data set gives the overall accuracy of the model. PPV and NPV show the proportion of how many compounds are classified correctly within all compounds classified as certain class (here B- and nB-compounds). In summary the higher the performance values, the better the performance of the classification.

2.5 Definition of applicability domain using proximity values

The applicability domain (AD) of an in silico predictive model has been important research topic in recent years. For the QSAR model, AD can be determined using a variety of methods, as described in the overview by Sahigara et al. [25], and many of them are applicable to RF classification models. In this work the usage of proximity matrix, as provided by the RF algorithm, is studied to determine the AD of the models. Proximity matrix is a square matrix containing proximity values for each pair of compounds which are calculated by the following procedure. The initial proximity value for each pair is zero. If two compounds in the tree end up at the same terminal node, one is added to the proximity value. The same comparison is made with each tree in the model and finally the sum is divided by the number of the trees. Proximity value shows how similar compounds are to each other based on descriptor values. The closer to 1 the proximity value is the more similar comparable compounds are.

![Figure 1. Illustrative representation of the confusion matrix (performance metrics).](image-url)
The proximity value is calculated for every compound in the validation set and, in order to fall into the AD of the corresponding model, the compound must meet requirements of two user-defined rules. The first rule is a proximity cut-off value (e.g. 0.5) selected by the user. This concerns all compounds from the training set that have proximity value less than the cut-off and are considered too dissimilar to the compound under investigation. Consequently all compounds having proximity value less than cut-off are considered as an unreliable source for the comparison. The second rule (defined by the user) indicates how many training set compounds (e.g. 3) must have the proximity value equal or greater than selected cut-off value. This rule determines how many compounds in the training set must be similar to the compound under investigation. Using those two rules helps to identify dissimilar compounds in relation to the training set and makes it easier to compare the compound under investigation with remaining similar compounds. For example, if all the remaining compounds are predicted correctly, then one can assume that compound under the investigation is predicted correctly as well. But if most of the remaining compounds are predicted incorrectly, then one must consider that prediction for the investigated compound might be unreliable. Since there is no fixed value for either rule, the users of the model can decide how loose or strict the rules are while analysing the AD.

3. Results and discussion
All compounds, their classifications based on experimental values of log BCF, calculated classifications and all the descriptors used in the models can be found in Table S1 and Table S2, respectively, in the supplementary material which is available via the multimedia link on the online article webpage. The results (Table 1) and discussions are grouped according to the distribution schemas of compounds and the AD section gives example for estimating RF classification performance and how one can adjust this according to the prediction needs and expert judgement.

3.1 Imbalanced model towards nB-compounds
The situation when the balance of the training set is tilted towards nB-compounds is common for BCF datasets, because one usually has more non-bio-accumulative than bio-accumulative compounds. Since the data set was already tilted towards nB-compounds there was no need to force the algorithm to use a specific number of compounds from both classes. Therefore default parameters of the RF algorithm were used where two-thirds of the compounds were included to grow each individual tree. From 100 random models (see section 2.3.2), the most frequent 20 descriptors were analysed and among them 12 occurred in the top two-thirds (Table S2). The final model (Table 1: Model 1) was built using three descriptors: logarithm of octanol–water partition coefficient (XlogP3), Wiener index (WPATH) and overall solute hydrogen bond basicity (MLFER_BH). The model and corresponding data are accessible at the QsarDB [26] repository (http://dx.doi.org/10.15152/QDB.116).

The logarithm of the octanol–water partition coefficient is the most commonly used descriptor in modelling BCF. It describes lipophilicity and accounts for cell permeability. But unfortunately log P alone cannot be used to distinguish between nB- and B-compounds. The Wiener index measures the size and branching of the compound, being the second parameter related to the permeability, because larger and more branched compounds have more difficulty penetrating the cell membrane. Therefore it is reasonable that log P is complemented by the size-related descriptor. The third descriptor is the Molecular Linear Free
Energy Relation descriptor [27], which denotes overall solute hydrogen bond basicity. This descriptor quantifies the hydrogen bond acceptor capabilities of a compound. It helps separate nB-compounds from B-compounds since the hydrogen bonding ability lowers the BCF value by keeping the compound in the aqueous phase or causing binding to the cell membrane.

The accuracy for Model 1 was very similar for both training and validation sets, with values of 0.85 and 0.87, respectively (Table 1). Despite this, the model classified nB-compounds more precisely than B-compounds. In more detail, the sensitivity for the training and validation sets were 0.64 and 0.72, respectively; this means that from 209 B-compounds in the whole data set, 70 B-compounds were misclassified (33%). On the contrary much higher specificity was achieved for the training set (0.91) and for the validation set (0.91), showing that more than 90% of the nB-compounds were classified correctly. Taking into consideration the environmental risk of misclassification of both classes one can say that classifying B-compounds as nB-compounds is more severe. Because of this it is more important to find a model with higher sensitivity and reasonable specificity for the classification of BCF.

### 3.2 Balanced model

In the case of the balanced model the same number of compounds from both classes was selected into the random sample; this selection was driven by the B-compounds since it was the smaller class. Two-thirds of the B-compounds, which is 94 B-compounds, and the same amount of nB-compounds were selected to take part in tree growing. From the preliminary models, the 12 most frequent descriptors were analysed and three descriptors were selected to the final model (Table 1: Model 2). The first two descriptors in Model 2 were exactly the same as in Model 1 (Table 1). The third descriptor, topological polar surface area (TopoPSA), in Model 2 (Table 1) is defined as the sum of surface areas of the oxygen, nitrogen, sulphur and phosphorous atoms, and also including surface areas of hydrogens connected to those atoms. This descriptor brings up an important aspect related to the polarizability and other...
electrostatic effects in the molecules caused by the surrounding condensed media. Because of the surface area of electronegative atoms (and attached hydrogens) in the calculation the descriptor, it bears some similarity to MLFER_BH descriptor from Model 1 and also brings up hydrogen bond acceptor capabilities of the molecules.

From the confusion matrix of the Model 2 one can see that performance measurements for both training set and validation set were consistent, being almost identical. The training set’s accuracy for the Model 2 is comparable with that of training set of the Model 1 (0.84 vs. 0.85), but unlike Model 1, sensitivity is higher than specificity in Model 2. Ideally sensitivity and specificity could be equal, but it is more important to have as little as possible of B-compounds that are predicted as nB-compounds in the model. From this point of view it is more important to have higher sensitivity rather than high specificity. From Table 1 one can see that balancing of input data achieved situation (sensitivity = 0.94, specificity = 0.82) where only 6% of the B-compounds in the training set were misclassified. A comparable situation was present in the validation set (sensitivity = 0.90, specificity = 0.83) where 10% of the B-compounds were misclassified.

3.3 Imbalanced model towards B-compounds

From Model 1 and Model 2, one can see that the ratio of the B and nB-compounds in the model’s training phase has an effect on prediction precision. Therefore the opposite situation to the Model 1 was created artificially by setting balance towards B-compounds by using the 4:1 ratio of B- and nB-compounds. For this two-thirds of random B-compounds were selected (94 compounds) and the number of nB-compounds to achieve the ratio above was a quarter from all B-compounds (24 compounds). From the five most frequent descriptors (Table S2) in the preliminary models, three descriptors were selected to Model 3 (Table 1; Model 3). The first two descriptors were identical to the descriptors in previous models. The third descriptor is the number of hydrogen bond acceptors (nHBAcc), considering hydrogen bonding ability, similarly to Model 1 (Table 1).

Model 3 (Table 1) predicted nearly perfectly (140 TP vs. 1 FN) training set’s B-compounds, but had high misclassification rate for the nB-compounds (372 TN vs. 160 FP). It is clear that Model 3 cannot be used as a prediction tool for B-compounds but it can be still useful to determine precisely if a compound is an nB-compound due to its high sensitivity and NPV. As one can see, in Table 1, only one B-compound from the training set was misclassified (note that log BCF for this compound is 3.01). Analysis of the NPV (0.997 for the training set and 1.00 for the validation set) of Model 3 shows that all compounds (except one) that were classified as nB-compounds actually were nB-compounds. Additional analysis of the specificity (0.70 for the training set and 0.67 for the validation set) shows that almost 70% of the nB-compounds from the whole data set were correctly classified as nB-compounds. In addition the sensitivity for the training set is 0.99 and for the validation set 1.00. Comparable results were achieved by Nendza and Herbst [7] when they developed a classification model for BCF with perfect sensitivity.

3.4 Robustness of the models

Robustness of the models was evaluated by repeating the model building process 100 times with randomly divided training and validation sets. Each randomly divided training set contained two-thirds of the compounds and the validation set one-third of the compounds. In each model three descriptors used to develop original model were used and average
sensitivity, specificity and accuracy were calculated using results from all 100 models. Average performance measurements of those models were compared against performance measurements from original models (Table 2). The comparison shows that there are only minor variations in the measurements of performance. From that we can conclude that the developed models were robust and that the selected descriptors were able to classify compounds regardless of the distribution of the compounds in the data sets.

3.5 Proximity-based applicability domain

From the classification results corresponding to the different data distributions it is clear that Model 2 is most suitable for classification purposes. Therefore the new AD concept is analysed only for Model 2 and the 334 compounds from the validation set. One example is shown where the optimal set of rules was obtained by considering the balance between sensitivity, specificity, PPV and FPV (Figure 2); the results are shown in Table 3.

For the optimal rules proximity cut-off was set to 0.7. For each compound in the validation set the training set had to contain at least four compounds with proximity value 0.7 or more. Utilization of those rules showed that, from the 334 compounds, 232 fit into the applicability domain. Comparison of the confusion matrices before and after applying AD rules (Table 3) shows that, from 102 compounds, which were out of AD, 32 were B-compounds and 70 nB-compounds. Comparison between confusion matrices (Table 3) shows that 232 compounds in the AD were predicted very well. From 36 B-compounds 35 were predicted correctly and, from 196 nB-compounds, nine were misclassified. In contrast all statistical parameters degrade for the compounds out of the AD. Dependent on one’s need the rules for AD can be adjusted to be looser or stricter. One must bear in mind that looser rules include more compounds in the AD at the cost of precision. Stricter rules, however, can give more precise results at the cost of excluding more compounds.

Using the proximity-based AD concept for a single unknown compound can be considered as finding the source compound(s) in read-across analysis and using this to infer properties on the target compound, that is, omitting compounds that have lower proximity value than cut-off from the training set. Reduction of such compounds leaves in the training set only compounds that have a certain similarity to the compound under investigation. This gives a better overview of the compounds in the training set and their relation to the compound under investigation. This makes it easier to check which kind of compounds are classified correctly or misclassified.

Explanation of the AD analysis for the two compounds is depicted in Figure 3. In both cases the proximity cut-off value was set to 0.7. For the first compound (Figure 3A, ID: 266), no compounds in the training set exceeded the proximity cut-off value. The calculated class

Table 2. Results from testing the robustness of model building process.

|                  | M1 | M1* | M2 | M2* | M3 | M3* | M1 | M1* | M2 | M2* | M3 | M3* |
|------------------|----|-----|----|-----|----|-----|----|-----|----|-----|----|-----|
| Sensitivity      | 0.64| 0.64| 0.94| 0.92| 0.99| 0.99| 0.72| 0.64| 0.90| 0.92| 1.00| 0.99|
| Specificity      | 0.91| 0.92| 0.82| 0.82| 0.70| 0.68| 0.91| 0.92| 0.83| 0.82| 0.67| 0.68|
| Accuracy         | 0.85| 0.86| 0.84| 0.84| 0.76| 0.75| 0.87| 0.86| 0.84| 0.84| 0.74| 0.75|

*the average of 100 models with randomly selected training and validation sets.
for this compound was nB, but with no similar compounds in the training set, the prediction for this compound might be unreliable. The experimental class for this compound is B and therefore the decision not to trust this prediction was correct. For the second compound

![Figure 2. Matrices to determine the applicability domain rules for Model 2.](image)

**Table 3. Examples of the applicability domains.**

|               | Without AD | In the AD (0.7 ≥ 4) | Out of the AD |
|---------------|------------|---------------------|---------------|
|               | B          | nB | Perf.* | B          | nB | Perf. | B          | nB | Perf.* |
| B             | 61         | 7  | 0.90*  | 35         | 1  | 0.97* | 26         | 6  | 0.81*  |
| nB            | 45         | 221 | 0.83*  | 9          | 187 | 0.95  | 36         | 34 | 0.51*  |
| Perf.*        | 0.58*d     | 0.97*e | 0.84*c  | 0.80*d     | 0.99*e | 0.96*e | 0.42*d     | 0.85*e | 0.59*c  |
| TNC*f         | 334        |     |         | 232        |      |      | 102        |      |        |

*aPerformance metrics.
*bSensitivity.
*cSpecificity.
*dAccuracy.
*ePPV.
*fNPV.

Total number of compounds.
there was eight compounds in the training set with a proximity value higher than cut-off. This compound was predicted as B-compound and the reliability of this prediction was evaluated using classification results of those eight compounds. From eight compounds seven were predicted correctly as B-compounds, which increased confidence that prediction was reliable. In addition, the structural similarity of compounds was examined. Six of the eight compounds were polychlorinated biphenyls (PCBs) and all of them were predicted correctly. The compound under investigation was also a PCB, which gave extra confidence that the prediction (experimental class is B) was reliable. Using this step-by-step analysis of AD makes decision about the credibility of the predictions easier.

4. Conclusions
The present study uses the Random Forest approach to classify an imbalanced BCF data set into B-compounds and nB-compounds. The research compares three model building strategies based on the balance of classes (towards nB-compounds, balanced, and towards B-compounds) in training phase to find the most suitable model. The proximity-based AD
approach was proposed for the RF models to provide quantitative measures for assessing the similarity of the target compound relative to the compounds in the training set. To the best of the authors’ knowledge the proximity matrix and information it contains has not been used before to determine an AD for a QSAR model. As an important addition, depending on expert judgement, one can easily change the AD parameters to choose the confidence level to their predictions.

The systematic descriptor selection procedure used for the model development is based on the frequency of occurrence of descriptors in RF trees and their mechanistic interpretations. The robustness analysis of the models showed that the results obtained were not accidental. All the descriptors used in the final models are discussed from the point of the bio-accumulation process. In all models, lipophilicity was the most significant descriptor, complemented by descriptors related to the size of the compound and polarizability/electrostatic effects introduced by the condensed medium and/or the hydrogen bonding ability of the compounds.

Among the three final models, the balanced model (Model 2) was the most suitable to classify both classes, although the models from imbalanced datasets have their strengths in certain application scenarios. The balanced model was able to correctly classify 90% of the B-compounds and 83% of the nB-compounds from the validation set. By applying the current AD schema with the proposed rules to the model, a third of the compounds were discarded from the validation set. From compounds inside the AD, 97% of the B-compounds and 95% of the nB-compounds were correctly classified. The imbalanced model towards B-compounds (Model 3) was able to correctly classify all the B-compounds. However, it predicted 33% of the nB-compounds as B-compounds. Despite that this model can be used to securely identify 67% of the nB-compounds.

All proposed models have their strengths in different application and decision-making scenarios for the regulatory assessment of compounds. Combining them with mechanistic knowledge about the descriptors gives an additional assurance to the decision-making process.

Supplemental material
Supplemental material for this article is available at http://dx.doi.org/10.1080/1062936X.2014.969310

The underlying research materials for this article can be accessed at http://dx.doi.org/10.15152/QDB.116 both for downloading and usage in the QsarDB repository web environment.

Funding
This work was supported by the Estonian Ministry for Education and Research [grant number SF0140031Bs09] and the European Union Regional Development Fund [grant number 3.2.1201.13-0021].

References

[1] European Parliament, Regulation (EC) No 1907/2006 of the European Parliament and of the Council of 18 December 2006 concerning the Registration, Evaluation, Authorisation and Restriction of Chemicals (REACH) establishing a European Chemicals Agency, amending Directive 1999/45/EC and repealing Council Regulation (EEC) No 793/93 and Commission Regulation (EC) No 1488/94 as well as Council Directive 76/769/EEC and Commission Directives 91/155/EEC, 93/67/EEC, 93/105/EC and 2000/21/EC, Off. J. Eur. Union L396 (2006), pp. 1–849.
J.D. Dearden, *QSAR modeling of bioaccumulation*, in *Predicting Chemical Toxicity and Fate*, M.T.D. Conin and D.J. Livingstone, eds., CRC Press, Boca Raton, FL., 2004, pp. 333–356.

M. Pavan, A.P. Worth, and T.I. Netzeva, *Review of QSAR Models for Bioconcentration*, EUR 22327EN, European Commission Joint Research Centre, Ispra, Italy, 2006.

G. Piir, S. Sild, A. Roncaglioni, E. Benfenati, and U. Maran, *QSAR model for the prediction of bio-concentration factor using aqueous solubility and descriptors considering various electronic effects*, SAR QSAR Environ. Res. 21 (2010), pp. 711–729.

X. Sun, Y. Li, X. Liu, J. Ding, Y. Wang, H. Shen, and Y. Chang, *Classification of bioaccumulative and non-bioaccumulative chemicals using statistical learning approaches*, Mol. Diversity 12 (2008), pp. 157–169.

M. Nendza and M. Müller, *Screening for low aquatic bioaccumulation (1): Lipinski’s ‘Rule of 5’ and molecular size*, SAR QSAR Environ. Res. 21 (2010), pp. 495–512.

M. Nendza and T. Herbst, *Screening for low aquatic bioaccumulation (2): Physico-chemical constraints*, SAR QSAR Environ. Res. 22 (2011), pp. 351–364.

S. Strempel, M. Nendza, M. Scheringer, and K. Hungerbühler, *Using conditional inference trees and random forests to predict the bioaccumulation potential of organic chemicals*, Environ. Toxicol. Chem. 32 (2013), pp. 1187–1195.

A. Fernández, A. Lombardo, R. Rallo, A. Roncaglioni, F. Giralt, and E. Benfenati, *Quantitative consensus of bioaccumulation models for integrated testing strategies*, Environ. Int. 45 (2012), pp. 51–58.

M. Galar, A. Fernández, E. Barrenecha, H. Bustince, and F. Herrera, *A review on ensembles for the class imbalance problem: Bagging-, boosting-, and hybrid-based approaches*, IEEE Trans. Syst. Man. Cybern. C 42 (2012), pp. 463–484.

A.P. Toropova, A.A. Toropov, A. Lombardo, A. Roncaglioni, E. Benfenati, and G. Gini, *A new bioconcentration factor model based on SMILES and indices of presence of atoms*, Eur. J. Med. Chem. 45 (2010), pp. 4399–4402.

*JChem for Excel 6.0.0*. ChemAxon, 2013; software available at http://www.chemaxon.com/products/jchem-for-office/

E. Bolton, Y. Wang, P.A. Thiessen, and S.H. Bryant, *PubChem: Integrated platform of small molecules and biological activities*, Annu. Rep. Comput. Chem. 4 (2008), pp. 217–241.

*SciFinder*, Chemical Abstracts Service, Columbus, OH, 2014; software available at https://scifinder.cas.org.

US Environmental Protection Agency, *Persistent Bioaccumulative Toxic (PBT) Chemicals; Lowering of Reporting Thresholds for Certain PBT Chemicals; Community Right-to-Know Toxic Chemical Reporting*, Fed. Reg. 64 (1999), pp. 58666–58753.

G. Piir, S. Sild, and U. Maran, *Comparative analysis of local and consensus quantitative structure-activity relationship approaches for the prediction of bio-concentration factor*, SAR QSAR Environ. Res. 24 (2013), pp. 175–199.

S. Dimitrov, N. Dimitrova, T. Parkerton, M. Comber, M. Bonnell, and O. Mekenyan, *Base-line model for identifying the bioaccumulation potential of chemicals*, SAR QSAR Environ. Res. 16 (2005), pp. 531–554.

C.W. Yap, *PaDEL-Descriptor: An open source software to calculate molecular descriptors and fingerprints*, J. Comput. Chem. 32 (2011), pp. 1466–1474.

T. Cheng, Y. Zhao, X. Li, F. Lin, Y. Xu, X. Zhang, Y. Li, R. Wang, and L. Lai, *Computation of octanol-water partition coefficients by guiding an additive model with knowledge*, J. Chem. Inf. Model. 47 (2007), pp. 2140–2148.

*Estimation Programs Interface Suite™ for Microsoft® Windows v 4.11*, US Environmental Protection Agency, Washington DC, 2012; software available at http://www.epa.gov/opptintr/exposure/pubs/episuite.htm.

I.V. Tetko, J. Gasteiger, R. Todeschini, A. Mauri, D. Livingstone, P. Ertl, V.A. Palyulin, E.V. Radchenko, N.S. Zefirov, A.S. Makarenko, V.Y. Tanchuk, and V.V. Prokopenko, *Virtual computational chemistry laboratory - design and description*, J. Comput. Aid. Mol. Des. 19 (2005), pp. 453–463.
[22] VCCLAB, 2005; software available at http://www.vcclab.org.

[23] L. Breiman, *Random Forests*, Machine Learning 45 (2001), pp. 5–32.

[24] *R: A language and environment for statistical computing 3.0.2*, R Foundation for Statistical Computing, Vienna, 2013; software available at http://www.R-project.org/.

[25] F. Sahigara, K. Mansouri, D. Ballabio, A. Mauri, V. Consonni, and R. Todeschini, *Comparison of different approaches to define the applicability domain of QSAR models*, Molecules 17 (2012), pp. 4791–4810.

[26] V. Ruusmann, S. Sild, and U. Maran, *QSAR DataBank – An approach for the digital organization and archiving of QSAR model information*, J. Cheminf. 6 (2014), p. 25.

[27] J.A. Platts, D. Butina, M.H. Abraham, and A. Hersey, *Estimation of molecular linear free energy relation descriptors using a group contribution approach*, J. Chem. Inf. Comput. Sci. 39 (1999), pp. 835–845.