# Modeling and predicting chiral stationary phase enantioselectivity: An efficient random forest classifier using an optimally balanced training dataset and an aggregation strategy

Patrick Piras, Robert Sheridan, Edward Sherer, Wes Schafer, Christopher Welch, Christian Roussel

# Modeling and predicting chiral stationary phase enantioselectivity: An efficient random forest classifier using an optimally balanced training dataset and an aggregation strategy

Patrick Piras[1] | Robert Sheridan[2] | Edward C. Sherer[3] | Wes Schafer[4] | Christopher J. Welch[5] | Christian Roussel[1]

[1]Aix Marseille Université, CNRS, Centrale Marseille, iSm2, Marseille, France

[2]Department of Structural Chemistry, Merck Research Laboratories, Rahway, USA

[3]Modeling and Informatics Process Research and Development, Merck Research Laboratories, Rahway, USA

[4]Department of Process & Analytical Chemistry, Merck Research Laboratories, Rahway, NJ, USA

[5]Welch Innovation, LLC, Cranbury, USA

Predicting whether a chiral column will be effective is a daily task for many analysts. Moreover, finding the best chiral column for separating a particular racemic compound is mostly a matter of trial and error that may take up to a week in some cases. In this study we have developed a novel prediction approach based on combining a random forest classifier and an optimized discretization method for dealing with enantioselectivity as a continuous variable. Using the optimization results, models were trained on data sets divided into four enantioselectivity classes. The best model performances were achieved by over-sampling the minority classes ($\alpha \leq 1.10$ and $\alpha \geq 2.00$), down-sampling the majority class ($1.2 \leq \alpha < 2.0$), and aggregating multicategory predictions into binary classifications. We tested our method on 41 chiral stationary phases using layered fingerprints as descriptors. Experimental results show that this learning methodology was successful in terms of average area under the Receiver Operating Characteristic curve, Kappa indices and F-measure for structure-based prediction of the enantioselective behavior of 34 chiral columns.

**KEYWORDS**

aggregated classification, chiral column selection, chiral stationary phases, imbalanced data sets, random forest classifier

## 1 | INTRODUCTION

With the advent of automated multi-column parallel screening [1], finding the optimal separation of a new chiral compound has become increasingly fast [2]. However, identifying the most promising chiral columns given a particular chemical structure remains a challenge. Nowadays, the column choice is mostly based on the trial and error method, and thus it may sometimes be necessary to screen a great variety of chiral stationary phases (CSP). A brief history and description of the most well-known commercial chiral columns have been provided in the Supplementary Information.

Machine learning methods are especially well suited for the challenge of structure-based selection of appropriate CSPs. Recent machine learning approaches such as random forest (RF) can recognize multiple non-linear interactions that might occur in chiral separations. RF has been successfully applied to an increasing number of problems. The algorithm introduced by Breiman in 2001 [3] was first applied to chemistry in 2003 for predicting a compound's quantitative or categorical biological activity [4]. Thereafter, the number of applications began to grow including a variety of chemistry topics such as aqueous solubility, aquatic toxicity as well as drug discovery [5–7]. Other main applications of RF concern gene classification, mass spectrum protein analysis or protein-ligand interaction prediction [8–11].

Abbreviations: AUC, area under the receiver operating characteristic curve; CSP, chiral stationary phases; FN, false negative; FP, false positive; FPR, false positive rate; RF, random forest; ROC, receiver operating characteristic; SVM, support vector machine; TN, true negative; TNR, true negative rate; TP, true positive; TPR, true positive rate

Chemoinformatic approaches have emerged as one of the most powerful tools for predicting enantioselectivity of chiral HPLC separations [12–14]. Among them, machine learning approaches represent the latest evolution for computational chemistry [15].

In a previous work, we used a RF regression approach to predict the enantioseparation value for untested compounds on 19 different CSPs [16]. A certain level of success was obtained for only four CSPs that we interpreted as coming from the disparity of data. As data were extracted from a literature database (ChirBase [17]), unsuccessful experiments constitute only a small minority of the data sets. Thus the difficulty of carrying out successful RF training on chiral HPLC data could lie with the imbalanced distribution of data sets.

Our objective in this study was to continue our recent work on predicting chiral separations using regression models [16] or matched molecular pair analysis [18]. In some sense, it is less ambitious than our aforementioned regression study because in this work we intend to address the problem of identifying the potentially unsuccessful and the most promising CSPs for the purpose of achieving an excellent chiral separation chromatography. To that end, an RF classifier approach was designed to tackle the problem of imbalanced chiral separation data and improve the previously reported predictive performance of the models.

The three main strategies used to achieve reasonably accurate models are summarized as follows:

*Optimized discretization*: Discretization was used for grouping enantioselectivity values into $N$ discrete groups called bins. The advantage of the binning strategy is to minimize the probability of error. In chiral chromatography, binning data can reduce the effects of enantioselectivity shifting which can occur on some chiral columns due to changes in mobile phase, pH or other factors.

*Balancing data sets*: We combined over-sampling the minority of samples with down-sampling the majority of samples. This approach allowed us to raise the weight of the minority classes e.g those with lowest and highest enantioselectivity values by randomly adding replicated samples [19]. On the other hand, down-sampling randomly eliminates samples from the majority class. Both are useful in forming balanced reduced training sets.

*Binary classification*: A simple aggregation strategy was applied to aggregate the RF multicategory predictions into a binary classification problem.

## 2 | METHODS

### 2.1 | Data preprocessing

Data for all CSP model studies were selected from ChirBase database. Forty-one chiral column data sets were extracted from 134 000 chiral HPLC/SFC separations. Data pretreatment processing was conducted according to the same procedure and selection criteria as detailed in our previous regression study [16]: only a single stereocenter is present in the molecules, the enantioselectivity value is available or can be estimated from retention times, and HPLC was achieved in the isocratic mode on a commercially available CSP. In ChirBase, it often happens that multiple separations of a unique molecule are recorded using different conditions for each unique CSP. In our study, for each CSP data set, we have identified the unique molecules and for each unique molecule we have taken the best recorded conditions, e.g. the 95th percentile enantioselectivity value. Similar to the previous study, we retained the CSP data sets that contain more than one hundred unique molecules.

### 2.2 | Descriptor generation

Compounds were encoded with RDKit layer fingerprints [20]. Layer fingerprints are subgraph-based 2D topological descriptors similar to Daylight fingerprints. Atom-types and aromaticity state are combined with bond types to hash all branched and linear molecular subgraphs up to a particular size. Unlike molecular keys with predefined patterns as Molecular Access System format Keys, the fingerprints are generated from the molecule itself. The algorithm screens the entire molecule and generates fingerprints:–

- For each atom
- Each atom and its nearest neighbors (with the bonds that connect them)
- Each group of atoms and bonds connected by paths up to two bonds long
- Each group of atoms and bonds connected by paths up to three bonds long and so on with paths up to four, five, six, etc. bonds long.

In this study, the size of the fingerprints was set to 1024 molecular bits and the maximum radius of the atomic environments considered was set to eight bonds.

### 2.3 | Random forest classifier

Four different learning approaches were evaluated: a support vector machine (SVM) as binary classifier, Naïve Bayesian, probabilistic neural network, and random forest classifier. SVM was tested with a linear and a polynomial kernel function. In all cases, the best performances were reached by the random forest classifier.

Random forest is an ensemble learning method that produces several decision trees. The method has few parameters to tune: the number of selected features ($n$) and the number of trees ($k$). The approach includes four steps:

- Generate a subset of training samples through replacement (bagging algorithm) on the training set. Typically, two-thirds of the samples are kept in the construction of each tree. The remaining samples, about one-third, are used in an internal cross-validation technique.
- Build a decision tree from the re-sampled training set with a subset of the features randomly selected at each individual branching of the tree. Repeat Steps 1, 2 and 3 for k times and generate k decision trees.
- Generate final predictions using the mean prediction over all the k decision trees

An ensemble of $k = 100$ trees without limiting the number of levels was sufficient for our classification purpose since the classification did not improve when the number of trees was increased. The Gini index was used as the splitting criterion.

## 2.4 | Validation

The models were validated by $k$-fold cross-validation. In $k$-fold cross-validation, we split the data into $k$ parts and use $k-1$ for training and the remaining for testing. Cross-validation is iterated $k$ times. We have estimated the prediction performances of the classifier with a $k$ value of 5. For some large data sets (number of molecules > 1000), the prediction errors were similar when using a stratified two-fold or five-fold cross-validation approach. For six CSP data sets dealing with very small data sets (number of molecules < 150), increasing the number of folds to ten leads to more stable performance estimations and thus reduces the variance of the error. An external validation was performed by testing all the classification models against molecules not included in training data.

## 2.5 | Discretization procedure

The chromatographic enantioselectivities were discretized into $k$ bins using an automatic optimization procedure. To determine the optimal number of bins, all numbers between $k = 3$ and 6 were tested as bin counts. Starting with an equal size binning, we performed learning with different random bin boundary placements iteratively until an optimal combination is achieved.

At each iteration, we selected half of the data for RF model learning and the remainder data was used as test to determine the classification error. A low classification error rate was the criterion to determine the best splitting. This discretization strategy leads to bin the enantioselectivity into the following four categories.

For Pirkle CSPs:

- Class 1 (10–15% of the data sets)

$$1 \leq \alpha < 1.05$$

- Class 2 (20–30% of the data sets)

$$1.05 \leq \alpha < 1.15$$

- Class 3 (45–55% of the data sets)

$$1.15 \leq \alpha < 2.00$$

- Class 4 (10–15% of the data sets)

$$\alpha \geq 2.00$$

For polysaccharide CSPs:

- Class 1 (10–15% of the data sets)

$$1 \leq \alpha < 1.10$$

- Class 2 (20–30% of the data sets)

$$1.10 \leq \alpha < 1.20$$

- Class 3 (45–55% of the data sets)

$$1.20 \leq \alpha < 2.10$$

- Class 4 (10–15% of the data sets)

$$\alpha \geq 2.10$$

Using this setting (four bins), we obtained the best performances for most of the CSP's data sets (46–60% accuracy). This optimal split is quite consistent with our experience of chiral HPLC:

- Class 1: “no separation” or “poor separation”
- Class 2: “separation is achieved or almost achieved”
- Class 3: “excellent separation”
- Class 4: “large separation”

## 2.6 | Dealing with multi-class imbalanced data

If best overall prediction rates of the classifier were obtained with this optimal split {Class 1, Class 2, Class 3, Class 4} of the data sets, the accuracies of each individual class were found to vary considerably (30 to 80%). This may be due to a negative effect of the imbalanced class distribution on prediction accuracy. We tried balancing data sets by randomly oversampling the two minority classes {Class 1, Class 4} and down-sampling the majority class {Class 3}. Unfortunately, we observed only a slight improvement of overall accuracy.

One can note that the automatic generated classes {Class 3, Class 4} include all the samples that can be separated at preparative scale (high enantioselectivity). A principal component analysis using molecular fingerprints as descriptors showed that samples of {Class 3, Class 4} (high

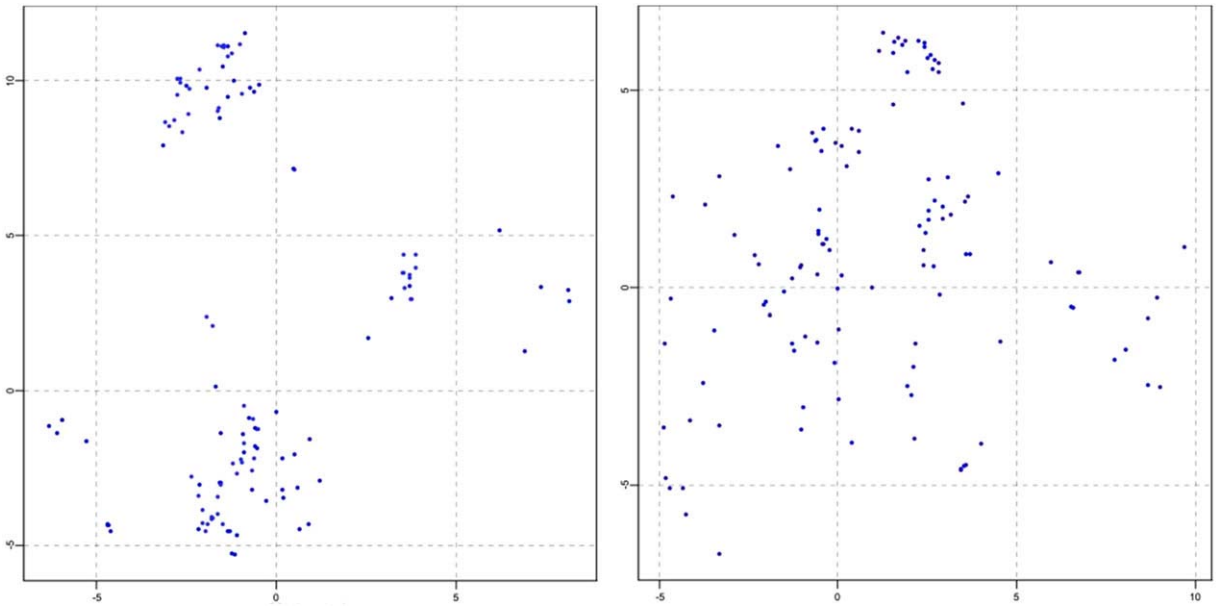

**FIGURE 1** Comparison of {Class 3, Class 4} samples ($\alpha \geq 1.18$, left picture) and {Class 1, Class 2} samples ($\alpha < 1.18$, right picture) using a principal component analysis and RDKit layered fingerprints as descriptors. Samples were separated on a Pirkle CSP and plotted against the principal component analysis scores

enantioselectivity) form distinct clusters, whereas samples of {Class1, Class 2} (none to moderate enantioselectivity) are less structured and more dispersed (Figure 1).

Clearly, compounds of {Class 3, Class 4} hold key information toward modeling chiral separation and thus may be an optimal split in a binary classification scenario.

This result led us to rather focus on a binary classification that identifies the lowest and the highest enantioselective chiral separations.

To this end, we have compared two strategies:

- The first strategy is a direct binary classification. It involves training the RF classifier on data sets grouped into two classes: none-to-moderate and high enantioselectivity.
- The second strategy is a two-step binary classification. It involves first training the RF classifier on data sets divided into the four optimized classes (same procedure as above). The second step involves aggregating the prediction classes {Class 1, Class 2, Class 3, Class 4} to their corresponding classes of none-to-moderate, or high enantioselectivity, to obtain a binary classification. The procedure consists of translating the prediction results simply by grouping output classes together in a $2 \times 2$ confusion matrix (binary classification) as detailed further in Section 2.7.1.

To determine which of these two strategies provides the best classification, we trained a RF classifier implementing each approach on the full original imbalanced data sets and on balanced randomly stratified data sets. For comparison purposes, we also studied the effect of the two strategies on a SVM classifier, a non-parametric supervised learning technique. To measure the level of performance of the models, we used the overall accuracy, Cohen'Kappa, F-measure (F1). and the area under the ROC curve (AUC) metrics as defined in Section 2.7.2. For all the measures, a value of 1 indicates a perfect classification. As seen in Table 1, RF models are significantly influenced by the chosen strategy when balanced training sets are applied. Interestingly, the same effect is observed with the SVM classifier. For both classifiers, the aggregation strategy provided the best performance. No differences between the two strategies are revealed when the full original imbalanced data sets are used as training sets. In all cases, whatever the strategy, balanced data sets always provided the best results.

In a literature search, we found no systematic study comparing the effects of such an aggregation strategy on the performance of RF classifiers.

Based on these results, the present study was carried out applying a RF classifier to balanced training data sets following the aggregation strategy outlined above.

## 2.7 | Model performance evaluation

### 2.7.1 | Confusion matrix

Training the RF classifier to the four class data sets described above generated classification results represented by a $4 \times 4$ confusion matrix that gather correct (true) and incorrect (false) class identification of samples (Table 2). The aggregation strategy was then applied by converting the $4 \times 4$ confusion matrix into a $2 \times 2$ confusion matrix (Table 3) grouping at first the four classes together:

**TABLE 1** 2X ten-fold-Cross-validated predictive performances of the direct binary classification and aggregation strategies obtained on Chiralpak IA column. Comparable values were found on other Daicel CSPs

| Training data set | Method | Classifier | Accuracy | F-measure $\alpha < 1.2$ | F-measure $\alpha \geq 1.2$ | Kappa | AUC |
|---|---|---|---|---|---|---|---|
| Imbalanced | Direct | RF | 0.7 | 0.66 | 0.73 | 0.4 | 0.77 |
| Imbalanced | Aggregation | RF | 0.69 | 0.61 | 0.74 | 0.35 | 0.78 |
| Imbalanced | Direct | SVM | 0.52 | 0.22 | 0.65 | 0.04 | X |
| Imbalanced | Aggregation | SVM | 0.5 | 0.1 | 0.66 | 0.02 | X |
| Balanced | Direct | RF | 0.68 | 0.69 | 0.68 | 0.37 | 0.74 |
| **Balanced** | **Aggregation** | **RF** | **0.78** | **0.78** | **0.79** | **0.57** | **0.86** |
| Balanced | Direct | SVM | 0.59 | 0.64 | 0.52 | 0.18 | X |
| **Balanced** | **Aggregation** | **SVM** | **0.7** | **0.69** | **0.71** | **0.4** | **X** |

**TABLE 2** $4 \times 4$ confusion matrix

| | | Predicted class | | | |
|---|---|---|---|---|---|
| **4 × 4 Matrix** | | **Class 1** | **Class 2** | **Class 3** | **Class 4** |
| Actual class | Class 1 | **True** | False | False | False |
| | Class 2 | False | **True** | False | False |
| | Class 3 | False | False | **True** | False |
| | Class 4 | False | False | False | **True** |

**TABLE 3** 2×2 confusion matrix obtained from conversion of a 4×4 confusion matrix

| 2 × 2 Matrix | Predicted negative {Class 1, Class 2} | Predicted positive {Class 3, Class 4} |
|---|---|---|
| Actual negative {Class 1, Class 2} | A (**true** negative) | C (false negative) |
| Actual positive {Class 3, Class 4} | B (false positive) | D (**true** positive) |

{Class1, Class 2} → Negative (no-to-moderate enantioselectivity)
{Class 3, Class 4}→ Positive (high enantioselectivity).

and counting the number of true positive, false positive, false negative, and true negative cases predicted by the $4 \times 4$ model, e.g. the number of samples correctly predicted to have a high α value (A-true positive (TP)), the number of samples predicted to have a high α value but do not actually have (B-false positive (FP)), the number of samples that do not have a high α value, but were not predicted by the model (C-false negative (FN)) and the number of samples correctly predicted to not have a high α value (D-true negative (TN)).

### 2.7.2 | Performance metrics

The $2 \times 2$ confusion matrix generated above gave us access to a number of $2 \times 2$ performance metrics, such as precision and recall, false positive rate and false negative rate, F-measure and the receiver operating characteristic (ROC). The following formulas were applied:

$$\text{Sensitivity or Recall or true positive rate (TPR)} = \frac{TP}{TP+FN} \quad (1)$$

$$\text{Sensitivity or Recall or true positive rate (TPR)} = \frac{TP}{TP+FN} \quad (2)$$

$$\text{Specificity or true negative rate (TNR)} = \frac{TN}{TN+FP} \quad (3)$$

$$\text{False positive rate (FPR)} = \frac{FP}{TN+FP} \quad (4)$$

$$\text{Precision} = \frac{TP}{TP+FP} \quad (5)$$

$$\text{Accuracy} = \frac{TP+TN}{TP+FP+TN+FN} \quad (6)$$

$$\text{F-measure (F1)} = \frac{2 \times \text{Recall} \times \text{Precision}}{\text{Recall} + \text{Precision}} \quad (7)$$

F-measure can be defined as a weighted harmonic mean of the precision and recall. It is high when both recall and precision are high.

Another common way to measure the performance of a binary classification is to create a ROC curve by plotting the true positive rate versus false positive rate at various discrimination thresholds. By computing the area under the ROC curve (AUC), one can measure the discriminating ability of the models [21]:

AUC > = 0.9 → excellent prediction
AUC > = 0.8 → good prediction
AUC < = 0.7 → mediocre prediction
AUC < = 0.5 → random prediction

The Cohen's kappa statistic is another well-known statistic measure for assessing the reliability of models by taking

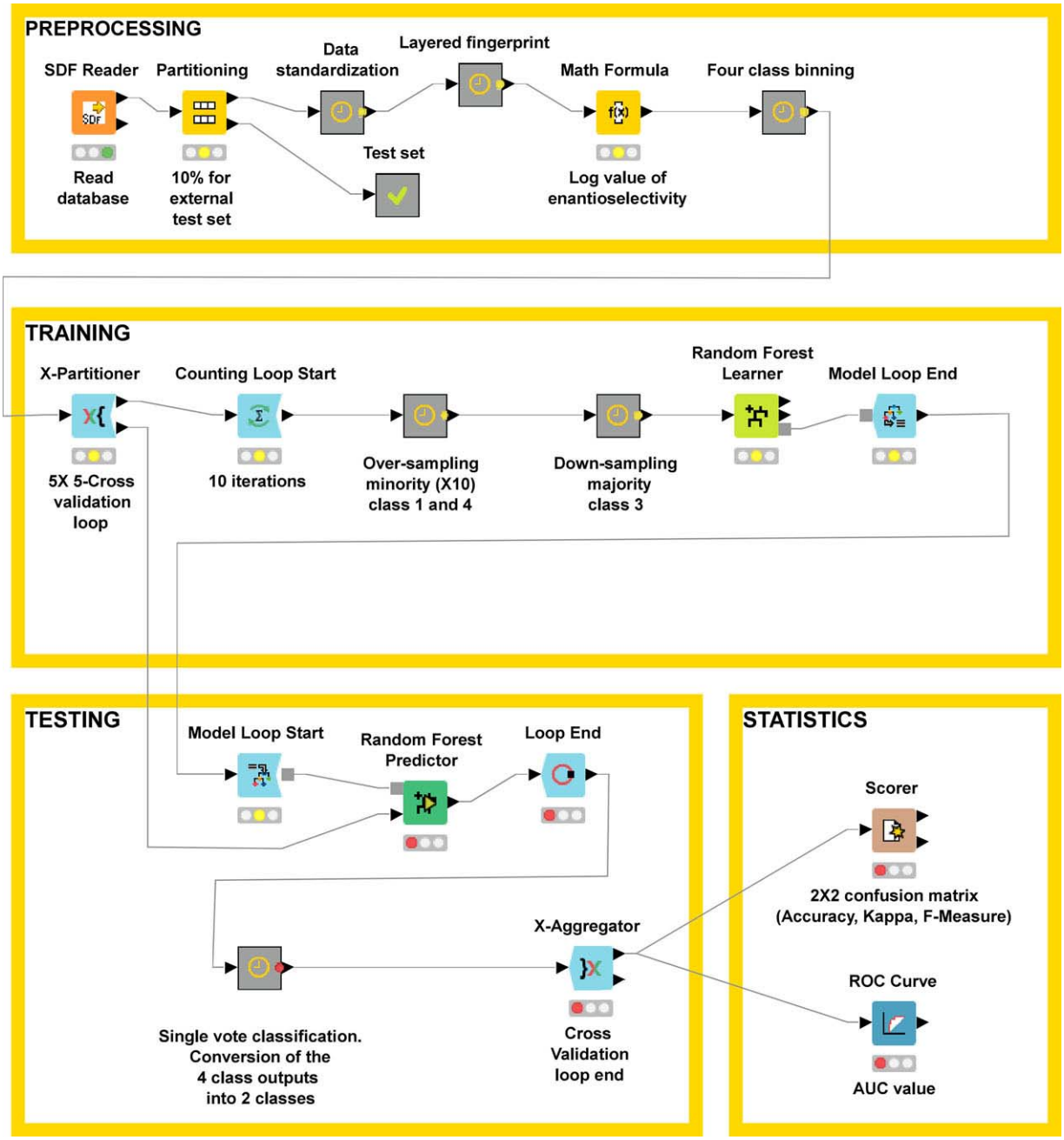


**FIGURE 2** Flowchart of the successful proposed classification method (five-fold cross-validation Knime workflow)

into account if a given correct prediction could be obtained by chance alone:

$$\text{Kappa} = \frac{P(A) - P(E)}{1 - P(E)} \tag{8}$$

where $P(A)$ is the total matching probability and $P(E)$ is the probability to match by chance.

A common interpretation is to consider kappa $< 0.3$ as a poor prediction, 0.3–0.4 as moderate, 0.4–0.6 as significant to good, and >0.60 as very good to excellent (>0.75).

For all these performance metrics, a maximum value of 1 is related to a perfect agreement.

## 3 | MODEL DEVELOPMENT

We constructed the final models in KNIME [22]. The data processing workflow is shown in Figure 2. For a given CSP, we first divided the available data into a stratified training set and a stratified test set for external validation (10% of the original data set depending on the CSP data set size). After applying all preprocessing steps, enantioselectivity values were binned into four classes as described in Section 2.5. For each CSP target, ten optimally balanced and diversified reduced datasets are constructed by means of under-sampling (e.g. randomly eliminating entries in class 3) and over-sampling

**TABLE 4** RF Classifier averaged results for all individual chiral columns using balanced data sets. Performance metrics were ranked according to AUC and Kappa values. Results are based on a 5X five-fold cross-validation

| CSP | True positive rate | True negative rate | F1 | Kappa | AUC | Accuracy | Kappa[b] | AUC[b] |
|---|---|---|---|---|---|---|---|---|
| DNB_LEU | 0.93 | 0.81 | 0.88 | **0.77** | **0.96** | 0.89 | 0.83 | 0.98 |
| Chirobiotic T | 0.86 | 0.89 | 0.85 | **0.74** | **0.94** | 0.86 | 0.65 | 0.92 |
| Crownpak-CR(+) | 0.86 | 0.88 | 0.85 | **0.73** | **0.92** | 0.87 | 0.70 | 0.93 |
| beta-GEM | 0.90 | 0.77 | 0.84 | **0.67** | **0.91** | 0.83 | 0.73 | 0.90 |
| Pirkle (R or S)-DNBPG | 0.79 | 0.88 | 0.82 | **0.68** | **0.90** | 0.83 | 0.51 | 0.87 |
| Whelk-O | 0.84 | 0.80 | 0.82 | **0.63** | **0.90** | 0.82 | 0.58 | 0.88 |
| Chirobiotic R | 0.77 | 0.84 | 0.80 | **0.61** | **0.90** | 0.80 | 0.48 | 0.86 |
| Cyclobond I RN | 0.83 | 0.80 | 0.82 | **0.62** | **0.88** | 0.82 | 0.50 | 0.86 |
| Kromasil_CHI-TBB | 0.96 | 0.70 | 0.86 | **0.52** | **0.87** | 0.80 | 0.62 | 0.89 |
| Kromasil CHI-DMB | 0.73 | 0.83 | 0.77 | **0.56** | **0.86** | 0.78 | 0.46 | 0.89 |
| Chiralpak IA | 0.77 | 0.79 | 0.78 | **0.56** | **0.86** | 0.78 | 0.50 | 0.83 |
| (R)-alpha-Burke 1 | 0.81 | 0.72 | 0.79 | **0.49** | **0.86** | 0.75 | 0.70 | 0.87 |
| Pirkle-1J | 0.80 | 0.73 | 0.80 | **0.53** | **0.85** | 0.77 | 0.44 | 0.86 |
| Chirobiotic V | 0.78 | 0.77 | 0.78 | **0.51** | **0.85** | 0.78 | 0.61 | 0.88 |
| Chiral AX QN-1 | 0.85 | 0.75 | 0.74 | **0.54** | **0.84** | 0.79 | 0.80 | 0.95 |
| Chirobiotic TAG | 0.80 | 0.72 | 0.77 | **0.52** | **0.83** | 0.75 | 0.54 | 0.84 |
| Chiralpak IC (Sepapak 5) | 0.73 | 0.77 | 0.74 | **0.48** | **0.83** | 0.74 | 0.47 | 0.83 |
| Cyclobond RSP | 0.76 | 0.75 | 0.76 | **0.54** | **0.82** | 0.76 | 0.40 | 0.80 |
| Chiral-BSA | 0.75 | 0.77 | 0.76 | **0.53** | **0.82** | 0.76 | 0.60 | 0.80 |
| Chiralcel OZ (Lux Cellulose-2 ) | 0.78 | 0.74 | 0.76 | **0.50** | **0.82** | 0.75 | 0.47 | 0.83 |
| Chiralpak AD | 0.75 | 0.74 | 0.75 | **0.50** | **0.82** | 0.75 | 0.48 | 0.82 |
| Chiradex | 0.73 | 0.74 | 0.74 | **0.47** | **0.82** | 0.73 | 0.45 | 0.83 |
| Chiralcel OD (Lux Cellulose-1) | 0.74 | 0.74 | 0.74 | **0.48** | **0.81** | 0.74 | 0.50 | 0.82 |
| Chiralcel OJ (Lux Cellulose-3) | 0.72 | 0.75 | 0.73 | **0.47** | **0.81** | 0.73 | 0.40 | 0.80 |
| Chiralpak IB | 0.70 | 0.76 | 0.72 | **0.46** | **0.81** | 0.73 | 0.50 | 0.80 |
| DACH-DNB | 0.80 | 0.72 | 0.76 | **0.50** | **0.80** | 0.75 | 0.60 | 0.85 |
| Chiralcel OB | 0.72 | 0.76 | 0.74 | **0.47** | **0.80** | 0.71 | 0.56 | 0.85 |
| Chiralcel OF | 0.77 | 0.70 | 0.73 | **0.46** | **0.80** | 0.73 | 0.50 | 0.82 |
| (SS)-ULMO | 0.70 | 0.79 | 0.75 | **0.44** | **0.80** | 0.73 | 0.53 | 0.80 |
| Chiralpak AS | 0.72 | 0.71 | 0.72 | **0.43** | **0.80** | 0.71 | 0.51 | 0.86 |
| Chiralpak AY (Lux Amylose-2)[a] | 0.76 | 0.71 | 0.70 | **0.43** | **0.80** | 0.71 | 0.44 | 0.81 |
| Chiral-AGP | 0.78 | 0.70 | 0.76 | **0.42** | **0.80** | 0.74 | 0.47 | 0.81 |
| Cyclobond I AC | 0.75 | 0.70 | 0.74 | **0.40** | **0.80** | 0.74 | 0.50 | 0.82 |
| (SS or RR)-P-CAP | 0.70 | 0.78 | 0.74 | **0.42** | **0.80** | 0.72 | 0.43 | 0.80 |
| Chiralpak zwix[a] | 0.72 | 0.75 | 0.80 | 0.47 | 0.78 | 0.76 | 0.42 | 0.76 |
| Chiralpak ID[a] | 0.70 | 0.76 | 0.70 | 0.44 | 0.77 | 0.72 | 0.40 | 0.70 |
| Cyclobond I | 0.71 | 0.69 | 0.69 | 0.38 | 0.75 | 0.69 | 0.43 | 0.73 |
| Chiralcel OX (Lux Cellulose-4)[a] | 0.60 | 0.70 | 0.63 | 0.30 | 0.75 | 0.65 | 0.20 | 0.70 |
| Chiralcel CA | 0.60 | 0.71 | 0.67 | 0.32 | 0.72 | 0.66 | 0.23 | 0.64 |
| LARIHC_CF6-P[a] | 0.68 | 0.66 | 0.71 | 0.28 | 0.69 | 0.65 | 0.30 | 0.70 |
| Ultron-ES-OVM[a] | 0.60 | 0.69 | 0.58 | 0.34 | 0.63 | 0.58 | 0.22 | 0.64 |

[a]Ten fold-cross validation.

[b]External data sets (10% of the original set).

methods (e.g. randomly adding replicated entries in class 1 and class 4). The ten successive iterations contribute to refine the diversity of the reduced data sets and also compensate the loss of information due to the entry elimination during the under-sampling operation.

For most CSP training sets, performing five replications of a stratified five-fold cross validation was found to afford stable performance estimations. 20% of the training set was withheld for external predictions, and the remaining 80% of the training set was used for model construction. For six CSP data sets, a ten-fold cross validation was preferred to lower the bias and thus train on as many compounds as possible.

At each cross-validation cycle, a classification single vote from all balanced data set models is combined using a rule based approach that exploits the confidence of predicted classes and the final multi-class results of the RF models are converted into a two class results by following the aggregation strategy described in Section 2.7.1.

## 4 | RESULTS

Table 4 shows the ranked classification performance of 41 CSPs after applying the proposed classification method. Three metrics have been considered to evaluate the global level of performance of the models: AUC, Kappa, and F-measure (F1). If one of these three metric values is not greater than a sufficient threshold value e.g. AUC $\geq$ 0.8, Kappa $\geq$ 0.4, and F1 $\geq$ 0.7 then the model performance will be considered as unsatisfactory. Based on these criteria, using cross-validation and a test set for external validation, we found good predictions for 34 CSPs.

As illustrated in Table 4, the top ten are occupied by Pirkle CSPs which present the best overall prediction values indicating excellent model performances. An obvious interpretation of this result is the smaller number of binding sites offered by Pirkle CSPs in comparison with the polymeric polysaccharide CSPs. Also in many cases these are designed CSPs, intended to function by a given chiral recognition mechanism.

Among the four cyclodextrin-based chiral columns (Astec Cyclobond CSPs), Cyclobond I RN (cyclodextrin derivatized with (*R*)-naphthyl-ethylcarbamate) is ranked with the top ten Pirkle CSPs. This is consistent with the well-known trend of this column to act like a standard Pirkle CSP [23]. Another interesting result concerns the Chiralpak IA immobilized column which is the best ranked Daicel column by the AUC measure and thus provides a better predictive model than its corresponding coated version Chiralpak AD.

Our model was unable to correctly predict the enantioselectivity of seven out of the forty one CSP data sets. We can note that:

**TABLE 5** Comparison between previous RF regression study (5X two-fold cross-validated $R^2$) and the new RF Classifier (5X five-fold cross-validated AUC and Kappa indices)

| CSP | CV-$R^2$[16] | Kappa | AUC |
|---|---|---|---|
| (DNB)-Leu | **0.67** | **0.77** | **0.96** |
| Chirobiotic-T | **0.62** | **0.74** | **0.94** |
| Crownpak-CR(+) | **0.57** | **0.73** | **0.92** |
| Whelk-O | **0.49** | **0.63** | **0.90** |
| Chirobiotic-R | **0.43** | **0.61** | **0.90** |
| Chirobiotic-TAG | 0.4 | 0.52 | 0.83 |
| Chirobiotic-V | 0.38 | 0.51 | 0.85 |
| Chiralpak-IA | 0.35 | 0.56 | 0.86 |
| Chiralcel-CA-1 | 0.3 | 0.32 | 0.72 |
| Chiralcel-OD | 0.27 | 0.48 | 0.81 |
| Chiralpak-AD | 0.26 | 0.50 | 0.82 |
| Chiralcel-OJ | 0.24 | 0.47 | 0.81 |
| Chiralpak-AS | 0.21 | 0.43 | 0.80 |
| Chiralcel-OZ | 0.21 | 0.50 | 0.82 |
| Chiralcel-OF | 0.18 | 0.46 | 0.80 |
| Chiralpak-IB | 0.18 | 0.46 | 0.81 |
| Chiral-AGP | 0.06 | 0.42 | 0.80 |
| Chiralpak-AY[a] | 0.04 | 0.43 | 0.80 |
| Ultron-ES-OVM[a] | 0.02 | 0.34 | 0.63 |

[a]Ten-fold cross-validated results

- Chiralcel CA CSPs are made of bulk microcrystalline cellulose and always used in the past as preparative columns. Regularly overloaded, they are not ideal columns for measuring accurate enantioselectivities.
- Cyclobond I and LARIHC columns are macrocyclic CSPs mainly governed by host–guest inclusions. Classification failures may be due to the difficulty of the classifier to characterize and generalize these processes.
- Ultron-ES-OVM, an ovomucoid protein based CSPs, is the worst CSP to predict enantioselectivity. It is not clear how to interpret this observation since the other protein CSPs Chiral-AGP and BSA offer better predicting results.
- Chiralpak ID and Chiralcel OX (Lux Cellulose-4) correspond to the sparsest data sets among the polysaccharide CSPs (<200 entries). This may explain why the RF approach failed to obtain the same good performances as other Daicel CSPs.

It is worth noting that cross-validation results were confirmed by comparing ROC curves on the test sets. Furthermore, we also found that our prediction ranking is consistent with our previous RF regression study for the top five CSPs and for the degree of difficulty of predicting the Ultron-ES-OVM as seen in Table 5.

Finally, we could bring up some interesting comparisons of CSPs by computing the variable importance scores to identify

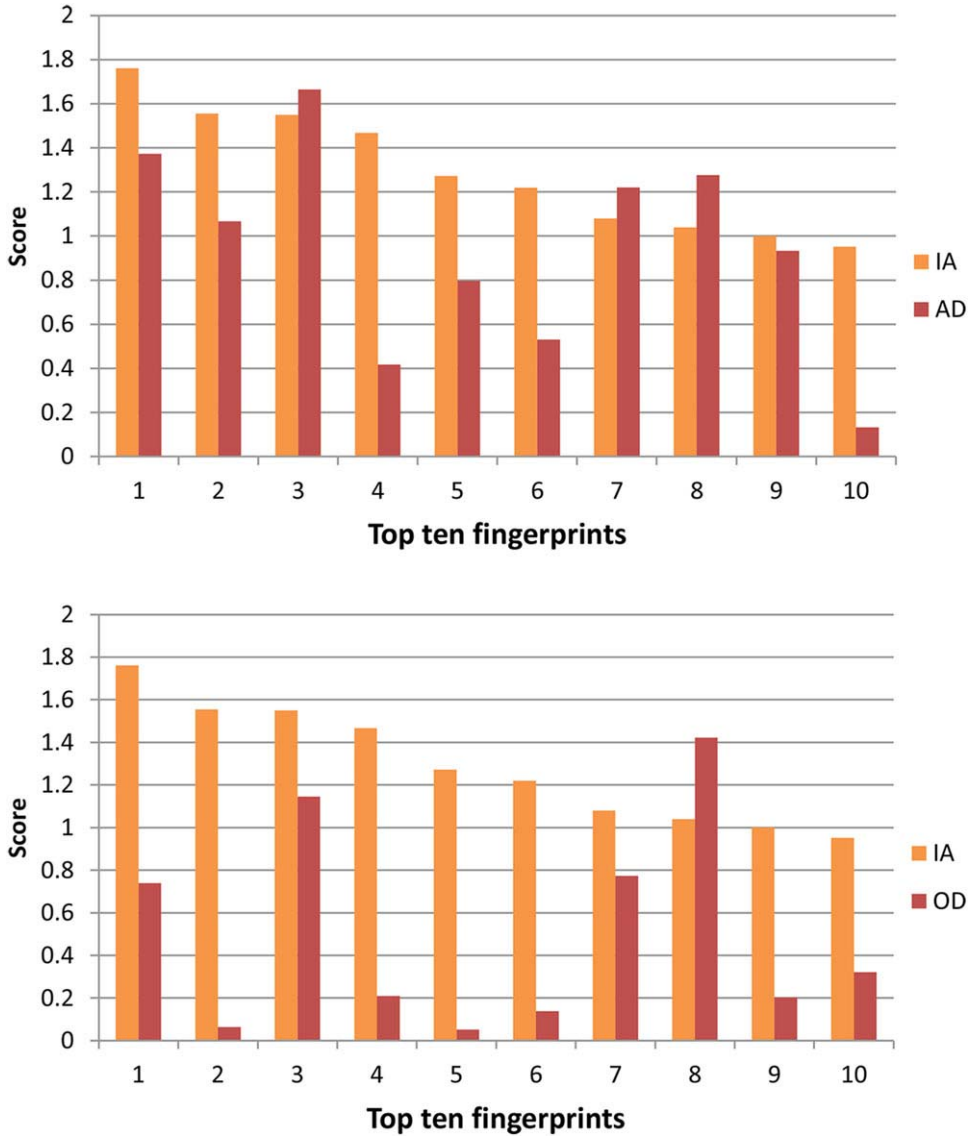


**FIGURE 3** Chiralpak IA fingerprints ranked by increased importance scores in RF model (top ten) and compared to their corresponding distributions in Chiralpak AD and Chiralcel OD models

the most important fingerprints used by the RF classifiers e.g. the fingerprints that were the most frequently selected to split the tree branches.

This concept is exemplified in Figure 3 where we display the histogram of the ten fingerprints that were the most frequently used across all trees of Chiralpak IA model. In this figure, we have added the scores found for the same fingerprints in Chiralpak AD and Chiralcel OD models. We can see that Chiralpak IA and AD have more feature importance in common than Chiralpak IA and Chiralcel OD. This result is somewhat coherent knowing that Chiralpak IA and AD columns are based on the same chiral selector structure.

These most common fingerprints of Chiralpak IA column are related to the following chemical features (for illustration purpose, fingerprints were converted into Molecular Access System format keys): N-C = 0, tBu, N-A-A-N, NH, Heterocyclic Atom > 1, A-Chain-A-cycle-A-chain-A, N Heterocycle, A-A-aromatic-A-A, N-A-A-N, O > 2 with A = any atom.

By taking a look at the chemical structures of Chiralpak IA data sets, we found that a majority of compounds associated with high enantioselectivity values share at least three of these fingerprints (Figure 4).

## 5 | CONCLUSION

We have successfully categorized into two classes the compound enantioselectivities obtained on thirty six CSPs with AUC values above 0.8 using RF as classifier. In this article, we found that the use of optimally binned and balanced training data sets significantly increased the prediction

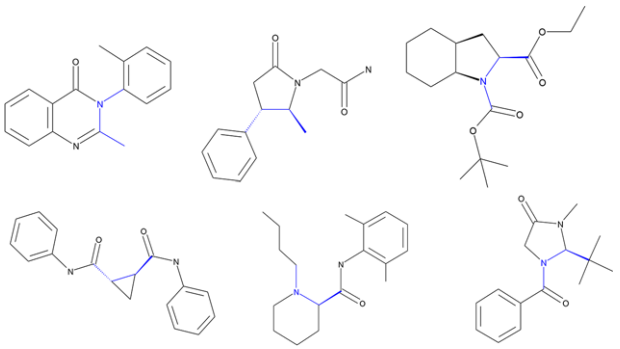

**FIGURE 4** Illustration of the frequent presence of a "A-Chain-A-cycle-A-chain-A", fingerprint among compounds separated with a high enantioselectivity on Chiralpak IA

accuracy in comparison with the full original imbalanced data sets. We have evaluated two approaches for generating a binary classification: one based on a direct binary classification and another based on an aggregation strategy applied after training a multi-class data set. Our results indicate that a significant improvement in classification performance of RF can be achieved when merging the multicategory classification results into a new aggregated 2 × 2 table. Another interesting finding in this study is that RF could efficiently capture the molecular differences between highly and no-to-moderate enantioselectivities of chiral separations. These results demonstrate for the first time that it is feasible to qualitatively estimate the enantioselectivity outcomes for a given compound on a variety of different CSPs.

We expect our new proposed methodology could further encourage the use of computational methods in chiral column screening strategies as it can be a very practical tool to prioritize the CSPs so that the ones with the most chance of being successful are considered first.


1. Patel, D. C., Wahab, M. F., Armstrong, D. W., Breitbach, Z. S., Advances in high-throughput and high-efficiency chiral liquid chromatographic separations. *J. Chromatogr. A* 2016, *1467*, 2–18.
2. Mattrey, F. T., Makarov, A. A., Regalado, E. L., Bernardoni, F., Figus, M., Hicks, M. B., Zheng, J., Wang, L., Schafer, W., Antonucci, V., Hamilton, S. E., Zawatzky, K., Welch., C. J., Current challenges and future prospects in chromatographic method development for pharmaceutical research. *Trends in Anal. Chem.* 2017, *95*, 36–46.
3. Breiman, L., Random forests. *Machine Learning* 2001, *45*, 5–32.
4. Svetnik, V., Liaw, A., Tong, C., Culberson, J. C., Sheridan, R. P., Feuston, B. P., Random forest: a classification and regression tool for compound classification and QSAR modeling. *J. Chem. Inf. Comput. Sci.* 2003, *43*, 1947–1958.
5. Palmer, D. S., O'Boyle, N. M., Glen, R. C., Mitchell, J. B., Random forest models to predict aqueous solubility. *J. Chem. Inf. Model.* 2007, *47*, 150–158.
6. Polishchuk, P. G., Muratov, E. N., Artemenko, A. G., Kolumbin, O. G., Muratov, N. N., Kuz'min, V. E., Application of random forest approach to QSAR prediction of aquatic toxicity. *J. Chem. Inf. Model.* 2009, *49*, 2481–2488.
7. Cano, G., Garcia-Rodriguez, J., Garcia-Garcia, A., Perez-Sanchez, H., Benediktsson, J. A., Thapa, A., Barr, A., Automatic selection of molecular descriptors using random forest: Application to drug discovery. *Expert Sys. Appl.* 2017, *72*, 151–159.
8. Yu, H., Chen, J., Xu, X., Li, Y., Zhao, H., Fang, Y., Li, X., Zhou, W., Wang, W., Wang, Y., A systematic prediction of multiple drug-target interactions from chemical, genomic, and pharmacological data. *PLoS ONE* 2012, *7*(5): e37608.
9. Qi, Y., Ensemble Machine Learning. in: Zhang, C., Ma, Y. (eds), Random Forest for Bioinformatics, Springer, New York 2012, pp. 307–323.
10. Boulesteix, A. L., Janitza, S., Kruppa, J., König, I. R., Overview of random forest methodology and practical guidance with emphasis on computational biology and bioinformatics. *WIREs: Data Mining Knowl. Discov.* 2012, *2*, 493–507.
11. Zilian, D., Sotriffer, C. A., SFCscore(RF): a random forest-based scoring function for improved affinity prediction of protein-ligand complexes. *J. Chem. Inf. Model.* 2013, *53*, 1923–1933.
12. Sardella, R., Macchiarulo, A., Urbinati, F., Ianni, F., Carotti, A., Kohout, M., Lindner, W., Péter, A., Ilisz, I., Exploring the enantiorecognition mechanism of Cinchona alkaloid-based zwitterionic chiral stationary phases and the basic trans-paroxetine enantiomers. *J. Sep. Sci.* https://doi.org/10.1002/jssc.201701068.
13. Del Rio, A., Gasteiger, J., Simple method for the prediction of the separation of racemates with high-performance liquid chromatography on Whelk-O1 chiral stationary phase. *J. Chromatogr. A* 2008, *1185*, 49–58.
14. Del Rio, A., Exploring enantioselective molecular recognition mechanisms with chemoinformatic techniques. *J. Sep. Sci.* 2009, *32*, 1566–1584.
15. Mitchell, J. B. O., Machine learning methods in chemoinformatics. *WIREs Comput. Mol. Sci.* 2014, *4*, 468–481.
16. Sheridan, R., Schafer, W., Piras, P., Zawatzky, K., Sherer, E. C., Roussel, C., Welch, C. J., Toward structure-based predictive tools for the selection of chiral stationary phases for the chromatographic separation of enantiomers. *J. Chromatogr. A* 2016, *1467*, 206–213.
17. Roussel, C., Piras, P., Chirbase: A molecular database for storage and retrieval of chromatographic chiral separations, *Pure Appl. Chem.* 1993, *65*, 235–244.
18. Sheridan, R., Piras, P., Sherer, E. C., Roussel, C., Pirkle, W. H., Welch, C. J., Mining Chromatographic Enantioseparation Data Using Matched Molecular Pair Analysis. *Molecules* 2016, *21*, 1297.
19. Chen, C., Liaw, A., Breiman L., Using random forest to learn imbalanced data. University of California, Berkeley, 2004.
20. Open-Source Cheminformatics Software, http://www.rdkit.org (last time accessed: January *12*, 2018).

21. Brown, C. D., Davis, H. T., Receiver operating characteristics curves and related decision measures: A tutorial. *Chemometr. Intell. Lab.* 2006, *80*, 24–38.

22. Berthold, M. R., Cebron, N., Dill, F., Gabriel, T. R., Kötter, T., Meinl, T., Ohl, P., Sieb, C., Thiel, K., Wiswedel, B., KNIME: The Konstanz Information Miner. In: Preisach, C., Burkhardt, H., Schmidt-Thieme, L., Decker R. (Eds) Data Analysis, Machine Learning and Applications. Studies in Classification, Data Analysis, and Knowledge Organization. Springer, Berlin, Heidelberg 2008.

23. Stalcup, A. M., Chang, S. C., Armstrong, D. W., Effect of the configuration of the substituents of derivatized β-cyclodextrin bonded phases on enantioselectivity in normal-phase liquid chromatography. *J. Chromatogr. A* 1991, *540*, 113–128.