Indirect Quantitative Structure-Retention Relationship for Steroid Identification: A chemometric challenge at “Chimométrie 2016”

AUTHORS: Giuseppe Marco Randazzo⁽¹⁾, Evelyne Vigneau⁽²⁾, Philippe Courcoux⁽²⁾, Corentin Harrouet⁽³⁾, Yves Lijour⁽³⁾, Pierre Dardenne⁽⁴⁾, Julien Boccard⁽¹⁾, Serge Rudaz⁽¹⁾

ABSTRACT

A chemometric challenge was proposed during the "Chimimétrie" congress 2016, held in Namur, Belgium, on 17-20 January. The aim of this contest was to challenge the ability of congress participants to build indirect Quantitative Structure-Retention Relationship models (QSRR) using the linear solvent strength (LSS) theory of reversed-phase liquid chromatography. QSRR is a very helpful method for the identification of unknown analytes, including the prediction of chromatographic retention time. Because of the potential presence of various isomeric compounds, accurate retention time prediction is particularly important in the context of steroid identification. In addition, the indirect prediction of retention time using the linear solvent strength (LSS) parameters S and $\log k_w$ provides a great advantage for use in any gradient conditions. In the proposed dataset, the experimental values of S and $\log k_w$ were estimated using Ultra High Pressure Liquid Chromatography separation with two linear gradients (5-95% ACN + 0.1% FA) of 15 and 60 minutes, respectively. The aim of the challenge was the accurate estimation of retention time for a 45 minute gradient by applying the LSS theory based on the predicted S and $\log k_w$ values. Molecular descriptors were calculated from a series of reference steroid compounds using the VolSurf+ software. By these means, a collection of 128 variables related to molecular shape, volume, polarisability, polar surface area, hydrophobic surface area, lipophilicity, molecular diffusion, and solubility was generated automatically. The dataset ($n=95$) included 76 steroid compounds for calibration and 19 for validation. Experimental $\log k_w$, S and retention time values were provided for the calibration set only. The results were evaluated according to the smallest RMSEC obtained for the retention time predictions of the validation set with the 45 minute gradient using the LSS parameters. Moreover, each individual relative error should not exceed 5% of the experimental retention time for both the calibration and validation sets. This paper summarises the approaches proposed by the best three participants and the challenge organiser.

KEYWORDS: QSRR, Chemometrics, Challenge, Linear Solvent Strength theory, Steroids, isotopomers identification

ABBREVIATIONS

$\log k_w$	: LSS $\log k_w$ parameter
S	: LSS S parameter
t_R	: retention time
PCA	: principal component analysis
MLR	: multiple linear regression
PLS	: partial least squares
RF	: random forest
SVR	: support vector regression
ANN	: artificial neural network
RMSEC	: root mean square error of calibration

1. INTRODUCTION

In every year since 2005 [1-5], a challenge was proposed in the context of the annual congress of “Chimiométrie” organised by the Chemometric group of the Société Française de Statistique (SFdS). The 2016 congress was held on 17-20 January in Namur, Belgium. For the first time, a molecular modelling problem was proposed to the participants, who were asked to implement chemometric methods for the development of a Quantitative Structure-Retention Relationship (QSRR) model. Chromatographic retention time results from complex intermolecular interactions between a solute, a stationary and a mobile phase [6]. Among all existing methods, Liquid Chromatography (LC) currently constitutes one of the most widely used analytical techniques for rapid sample analysis. Its combination with high-resolution mass spectrometry detection (HRMS) allows improved sensitivity and resolution for the analysis of complex samples, such as biological fluids. Despite its indisputable advantages, HRMS remains limited for distinguishing isotopomers, which are characterised by their identical mass and molecular formula [6]. More specifically, accurate retention time prediction constitutes an important support in the context of steroid identification because of the many isomeric compounds. In such cases, retention time is a crucial parameter for molecular identification. Starting from the principle that different

structures possess specific molecular properties, the aim of the challenge was to develop an indirect retention time prediction model based on the Linear Solvent Strength (LSS) theory [7], which constitutes a linearisation of the retention factor behaviour towards the amount of organic solvent in one of the most commonly used chromatographic approaches, the gradient mode in reversed phase liquid chromatography (RP-LC). This approach is based on the determination of the two coefficients of the linear relation, i.e., the intercept $\log k_w$ and the slope S , and provides the great advantage of being usable in any gradient conditions. Furthermore, the estimation of these two model parameters makes it possible to optimise the separation in the case of coeluting analytes. It is noteworthy that this approach is integrated in most chromatographic software. Data were made available through the conference website four months before the event at <https://chimio2016.sciencesconf.org/page/challenge>. The dataset included 76 steroid compounds for calibration and 19 for validation. A series of molecular descriptors was calculated from the structures using the VolSurf+ software [8]. By these means, a collection of 128 variables was generated automatically. Experimental S and $\log k_w$ were estimated using Ultra High Pressure Liquid Chromatography (UHPLC) separation with two linear gradients (5-95% ACN + 0.1% FA) of 15 and 60 minutes, respectively. The experimental $\log k_w$, S and retention time values were provided for the calibration set only. The aim of the study was the accurate estimation of retention time for a 45 minute gradient (smallest RMSEC) using the predicted S and $\log k_w$ values by applying the LSS theory. Moreover, the additional constraint of limiting each individual relative error below 5% of the experimental retention time for both the calibration and validation sets was proposed. Three finalists were invited to present their solutions orally, and their approaches are summarised in this paper.

2. MATERIALS AND METHODS

2.1. Chemical reagents

Reference steroids were obtained from various suppliers (Steraloids, Sigma, LGC Standards, Sterling). ULC-MS grade methanol (MeOH), acetonitrile (ACN) and formic acid were purchased from Biosolve (Valkenswaard, Netherlands). Ultrapure water (18.2 M Ω cm) was

obtained with a Milli Q Advantage A10 purification system from Millipore (Bedford, MA, USA). Stock solutions of 1 mg/mL of each steroid standard were made in methanol. Working solutions (10 µg/mL) were prepared by dilutions of the stock solution in ACN 0.1% FA/water 0.1% FA (5:95).

2.2. Experimental retention time measurements

Retention times were measured using an Acquity UHPLC-QTOF-MS Xevo™ system from Waters (Mildford, MA, USA). Chromatographic separation was achieved using a Cortecs C18 column (3.0 x 100 mm, 2.7 µm, Waters). Different linear gradients of mobile phase A (0.1 % FA in water) and mobile phase B (0.1 % FA in ACN) at a constant flow rate of 0.5 mL/min were used. Linear gradients varying the organic solvent composition from 5% to 95 % were performed in 15, 45 and 60 minutes. 10 µL of each working solution were injected.

The Xevo QTOF was equipped with an electrospray ionization (ESI) source operating in positive mode. The MS operating conditions were as follows: desolvation gas flow was set at 800 L/h with a temperature of 500°C, source temperature was kept at 120°C, capillary voltage and sampling cone voltage were fixed at 4kV and 30 kV respectively, cone gas flow was defined at 20 L/h. A wide-pass quadrupole mode with low collision energy (5 eV) was used for the acquisition (range m/z 50-1000). Data were collected in centroid mode with a scan time of 0.2 s, using dynamic range enhancement (DRE). Recalibration of the data was made thanks to the infusion of a solution of 200 pg/µL of Leucine-enkephalin (Sigma-Aldrich, Buchs, Switzerland) at 10 µL/min. Peak detection and retention time determination were performed using MassLynx v 4.1. (Waters).

2.3. Dataset, molecular descriptors and LSS parameters

The dataset composed of 95 steroids was split into a calibration (76 molecules) and a validation set (19 compounds) applying the most descriptive compounds algorithm (MDC) [9]. This method selects representative compounds positioned in dense regions of a given chemical space, in that case the retention time scale, by computing pairwise distances. MDC allowed the selection of two representative subsets equally representing diastereoisomers, constitutional isomers and positional isomers. Each molecule was characterised by

molecular descriptors calculated using the VolSurf+ software package [8]. Volsurf+ uses the GRID computational procedure [10] to condense the 3D information originating from Molecular Interaction Fields (MIFs). MIFs reflect the attractive and repulsive forces between a chemical probe and a target molecule encoding the chemical information. This information is then converted into numerical values. Different probes generate different types of chemical information: the water probe OH2 provides information about the molecular shape/volume/moment of interaction/capacity factors/polar surface areas, hydrophobic interactions are obtained through the DRY probe, H-bond donor interactions through the NH probe and H-bond acceptor information using the =O probe.

For each molecule, 128 molecular descriptors were automatically calculated to build 3D-QSAR/QSPR models. Experimental values of $\log k_w$ and S were extrapolated from experimental retention times based on the python package PyLSS [11]. PyLSS applies the LSS theory developed by Snyder and Dolan through a simplex optimiser. Based on two experimental retention times acquired using linear gradient elution, $\log k_w$ and S were iteratively estimated to minimise the retention time recalculation error. Experimental retention times were also measured for the calibration set using a 45 minute gradient.

3. Results

3.1. Participant 1

As a starting point, descriptive statistics were used to obtain a first insight into the calibration dataset. It turned out that for three molecules, all the values for both LSS parameters, S and $\log k_w$, were equal to zero. These three molecules were therefore removed. Moreover, the HTSflag descriptor was discarded due to zero values in the entire dataset.

A multivariate explorative analysis of the VolSurf+ variables, using PCA, highlighted the steroid RU486 as an atypical molecule with high leverage, strongly influencing the direction of the second Principal Component. RU486 was characterised by high levels of descriptors such as Surf, Vol, POL and MW. Participant 1 therefore decided to exclude this compound from the calibration set. Finally, the input matrix, $\mathbf{X}$, for the calibration of the models consisted of 72 observations and 127 VolSurf+ variables.

Participant 1 was mainly motivated by the investigation of machine learning tools, more specifically regression trees and random forests (RF) approaches [12]. Among the advantages of these approaches, the easy interpretation of the models with recursive dichotomic decision rules and the possibility to handle nonlinear relationships without any distributional hypotheses were underlined. Another key point is that RF not only led to a model of prediction for a quantitative (or qualitative) response but also provided an evaluation of the importance of each variable in this model.

The construction of two models, one for each of the LSS parameters, based on the RF approach was decomposed into different steps to address specific issues. The whole process was repeated separately for each of the LSS parameters.

- (i) Using y to denote the response to be predicted, a general RF was built using all the molecular descriptors, i.e., $p=127$ predictors. The Variable Importance (VI), i.e., the permutation-based Mean Decrease in Accuracy measure introduced by Breiman [12], was assessed for each predictor. All the predictors were ranked according to their importance, the most important variables being the ones for which the permutation procedure had a large impact on model accuracy. More precisely, 50 random forests of 2000 trees were built. The averaged values of the VI over the 50

forests were used to rank the variables in decreasing order. Participant 1 chose to retain a subset of k predictors based on their stability in the list of the most important variables. As proposed by Genuer et al. [13], the standard deviation associated with the VI estimated values was considered. Each of the 50 forests provided an ordered list of predictors. These lists may vary, but if the k top variables are truly predictive, the ordered subsets of the k first variables are expected to be stable. By identifying robust subsets of variables over the 50 forests, the aim was to select truly predictive variables.

- (ii) Once a subset of k predictors was chosen, the RF parameters were thoroughly investigated. In essence, a forest is random for two reasons: first, each tree in the forest involves a bootstrapped set of observations; the observations not selected during the bootstrapping process belong to the Out-Of-Bag (OOB) set. Second, at each node of each tree, only some of the input variables are considered as candidates for the splitting process. The number of these randomly selected variables is usually denoted $mtry$. This parameter is known to be a key meta-parameter for the RF algorithm [13]. Usually, $mtry=p/3$ is suggested for regression trees. In the proposed procedure, the $mtry$ parameter was chosen based on the Root Mean Squared Error of the OOB observations ($RMSE_{OOB}$). Simultaneously, the $nodesize$ parameter, *i.e.*, the minimum size of the terminal nodes of a tree, was also optimised. As expected, this last parameter was less crucial than the $mtry$ parameter (" $nodesize$ " values ranging from 1 to 3 led to similar results). The $RMSE_{OOB}$ criterion was also considered for determining the best solution. In practice, the curves of the stability criterion as a function of k usually showed a small number of flat levels (data not shown). Consequently, only three alternatives were considered (instead of testing a whole range of values for k). These three alternatives were as follows:

- a RF model based on the $p=127$ predictors,
- a model using only a short subset of k_1 predictors (with a high degree of stability),

- and a model based on a slightly larger subset of k_2 predictors (with a moderate degree of stability).
- (iii) Finally, the prediction of the y response was estimated using the aggregated values of predictions obtained over all the trees (here 5000) of the RF defined with the chosen parameters (size, k , of the subset of predictors and values of the *mtry* and *nodesize* parameters).

Figure 1(a) shows the most important molecular descriptors in the prediction of the LSS parameter S , ordered according to their VI value. The ordered lists of the selected predictors are detailed in the upper part of Table 1. The variables in bold correspond to the reduced subset of the selected predictors ($k_1=6$). The whole list describes the subset of the $k_2=14$ selected predictors. The results of RF models built without predictors pre-selection and with the two different subsets of input variables (with optimised *mtry* and *nodesize* metaparameter values for each condition) are given in the upper part of Table 2. The $RMSE_{OOB}$ criterion highlighted the smallest errors in prediction using a reduced list of 6 molecular descriptors. The model based on the 6 most important descriptors (with *mtry*=1, *nodesize*=1) was then retained for the prediction of the S parameter, obtained using a bagging (bootstrapped aggregating) process. It should be noted that a small number of descriptors seemed to be sufficient for the prediction of S but also that these predictors were correlated: the highest correlation coefficient, in absolute value, was 0.95, between LgBB and CACO2, and the lowest was 0.62, between LOGP.c.Hex and ACACDO. In fact, it is not surprising that CACO2 (P-glycoprotein efflux transport) is correlated with LgBB (blood-brain barrier).

P-glycoprotein is also expressed at the blood-brain barrier as well [14, 15]. It may be concluded that the underlying prediction model for S was a rather simple and parsimonious model.

The same rationale was used for the $\log k_w$ LSS parameter (Figure 1(b) and Table 1). The reduced list consisted of $k_1=5$ descriptors and the extended list of $k_2=21$ descriptors. It turned out (lower part of Table 2) that the best model was obtained using the subset formed by the

21 most important descriptors (with *mtry*=3 and *nodesize*=2). Compared with the prediction models for *S*, lower prediction ability may be achieved regarding the *log k_w* parameter (lower *R*²). Moreover, more descriptors were required. The 21 selected descriptors were clustered into 6 clusters: {Vol, Surf, POL, DIFF, MW, FLEX, DRDRDR, HAS}, {ACACDO, ACACAC}, {D7, CD7, CD5}, {%FU8, %FU9, %FU10, VD}, {LgS7, L0LgS, LgS6} and {IW2}.

Figure 1 Ordered list of molecular descriptors in decreasing order of their Variable Importance regarding the prediction of (a) S , (b) $\log k_w$.

Table 1 List of selected molecular descriptors for both responses, *S* and *log k_w*. The descriptors in bold correspond to a more stringent selection.

Ordered list of the VolSurf+ descriptors	
<i>S</i>	"LgBB" , "LOGP.c.Hex" , "CACO2" , "CW2" , "ACACDO" , "SKIN" , "PSAR", "PHSAR", "CW1", "MetStab", "HSA", "CW4", "CW3", "CW5"
<i>Log k_w</i>	"Vol" , "ACACDO" , "POL" , "ACACAC" , "DIFF" , "CD7", "DRDRDR", "FLEX", "VD", "LgS7", "MW", "CD5", "LgS6", "Surf", "%FU9", "HSA", "%FU8", "IW2", "%FU10", "D7", "L0LgS"

Table 2 Results of RF models for both responses, S and $\log k_W$. For each response, the three conditions with various model's parameters were considered.

	# predictors	$mtry$	$nodesize$	RMSE _{OOB}	R^2
S					
	127	40	2	0.703	0.973
	14	3	2	0.675	0.975
	6	1	1	0.665	0.974
Log k_W					
	127	60	2	0.101	0.953
	21	3	2	0.088	0.964
	5	1	1	0.101	0.947

3.2. Participant 2

Steroids were first assigned to different classes based on their known structures, *i.e.* Sterone, Corticosterone, Pregnanolone, and Androsterone. PCA was then conducted to explore the data, highlight outliers and remove unnecessary variables. As on the work of participant 1, RU486 was considered as an outlier. Nine variables with more than 50% missing values and four others with limited variability, *i.e.* NCC, DRDRDO, DRACDO, DRDODO, ACACAC, ACACDO, ACDODO, DODODO, and HTSflag, were excluded from the dataset. Because two LSS parameters, *i.e.*, $\log k_W$ and S , were given in the calibration set to predict t_R , their correlations were examined and 3 steroids removed because their Y values were equal to zero.

Starting from the proven relationships between both responses, a simple PLS2 model was calculated and compared with PLS1. As t_R was the only response required for the challenge, the performance of the PLS2 and PLS1 models was estimated based on the Root Mean Square Error in Cross-Validation (RMSECV) calculated for the prediction of t_R . As their errors of prediction would propagate to the t_R prediction model, $\log k_W$ and S were not further considered, although the advantage of predicting retention times for any gradient condition is thus lost. The usual PLS validation steps were conducted, including variable selection, outlier

detection, and examination of linearity. The final best model was a PLS2 model with 6 latent variables, 80 variables out of the 128 initial ones, and 65 included observations, as shown in Figure 2.

Figure 2 Observed t_R vs. Predicted t_R for a 6-component PLS2 model

Interestingly, the four classes of steroids were highlighted by the first two PLS components (Figure 3). Better predictions could be expected from class models, especially for lower t_R , but more individuals per group would be necessary.

Figure 3 Steroid groups on the first two PLS components for the calibration set

Regarding prediction, the test set was projected onto the 6-component PLS2 model. The Hotelling T^2 and distances to the model for the test objects fell within the class limits. Moreover, each observation in the test set was projected into the corresponding group identified in the calibration set from its chemical nomenclature. No outlier was found in the test set, which was thereby proved to be very similar to the calibration set. The PLS2 model was applied to obtain the t_R values on the test set, with reasonable results. However, due to the uncertainties of standard deviations and particularly the calculated RMSEC with 19 degrees of freedom, the RMSEC winner was expected to be at least 1.5 lower than the worst RMSEC at the 5% level.

3.3. Participant 3

Because of his expertise in spectroscopy, Participant 3 chose a similar approach to handle the QSAR data. First, Participant 3 carefully examined the structure of the data in the Y and X spaces. The first scatter plot between $\log k_W$ and S allowed 3 objects with missing values to be removed. $\log k_W$ and S were negatively correlated, with $r=-0.41$. Among the X values, several dozen scatter biplots were drawn between the X variables including the calibration and test sets. Figure 4 gives an example for the first 4 X variables.

Figure 4 Biplot of the first 4 VolSurf+ variables.

A performed PCA on the studentised X variables (Figure 5) showed that several objects were quite far from the ones to be predicted. The strategy was then to remove the objects in the calibration set that were outside the range of the test set in the X space.

Figure 5 PCA scatter plot of X matrix PC1 vs. PC3. The blue ellipse highlights the compounds out of the prediction space.

The objects inside the ellipse were removed, assuming that, being quite far from the ones to be predicted, the information they carry would not be useful for the modelling or to predict the red objects of the test set. Based on the 50 remaining objects, MLR models were developed using a manual step up procedure (Foss, Winisi package). The selection criteria for the final models were the maximum values of the regression coefficient F -tests and the minimum Mahalanobis distances of the test objects vs. each model. Table 3 reports the last 2 models used to predict the 19 test objects.

Table 3 MLR calibration results

MLR MODEL FOR LOGK R2=0.82 SEC=0.07					MLR MODEL FOR S R2=0.93 SEC=0.39				
	Coefficients	Ftest	Data Point	Var		Coefficients	Ftest	Data Point	Var
	-0.7777					7.4249			
1	-4.6159	50.6	40	ID2	1	-0.3591	28.8	69	LgD8
2	12.0909	60.4	39	ID1	2	-1.6329	254.4	97	SKIN
3	2.1894	56.4	42	CD1	3	0.1653	16.8	107	DD3
4	0.0796	139.1	54	POL	4	-0.0207	32.7	55	MW
5	-0.0469	17.5	60	LOGP c-Hex	5	94.0455	44.9	46	CD5
6	6.6006	3.4	38	CW8					

3.4. Challenge organiser

The organisers' approach was based on the comparison of four different regression algorithms to predict $\log k_W$ and S , selecting the smallest prediction error. The dataset was first analysed by PCA using the 128 VolSurf+ descriptors. RU486 was also highlighted as an outlier compound because it was located outside the Hotelling confidence ellipse at 95%. This exogenous steroid is characterised by a structure clearly different from the rest of the dataset. Additionally, by closer examination of the $\log k_W$ and S parameters, three compounds were found to have null values, as judiciously observed by some participants. These compounds were removed from the dataset to generate a calibration set composed of 72 molecules.

As steroids are molecules with acid/basic centres characterised by high pKa values, variable analysis based on chemical knowledge was conducted before computing the regression models. Because the solvent pH was estimated to be approximately 2.5, the molecules were mostly in their neutral form in the retained chromatographic conditions, and molecular

descriptors related to pH were removed. ADME descriptors were also excluded because of their lack of chemico-physical relevance in the retention process. Finally, 91 molecular descriptors were retained for further analysis. Artificial Neural Network (ANN), Random Forest (RF), Support Vector Regression (SVR) and Partial Least Squares (PLS) regression algorithms were compared for their ability to properly estimate the LSS parameters. Because these algorithms include parameters that need to be appropriately tuned, optimisation was conducted with a specific strategy for each case. ANNs were computed using the Levenberg-Marquardt back-propagation algorithm, and the hidden layer size was optimised using a grid search. For that purpose, the calibration set was further divided into a training (70%), a validation (15%) and a test subset (15%). ANN predictions with 2 to 20 neurons in the hidden layer were compared, and the best model was obtained with 10 neurons in the hidden layer for both the $\log k_w$ and S parameters. RF was optimised using nested cross validation, and the number of trees was varied from 100 to 2000 with a step of 100. The best model was obtained with 1000 trees for both the $\log k_w$ and S parameters. SVR was computed using a kernel radial basis function. The penalty parameter of the error term C and the epsilon-tube within which no penalty is associated in the training loss function ξ were optimised using the Nelder–Mead simplex algorithm. Finally, PLS was computed using 3 latent variables for both LSS parameters, as estimated by bootstrap 5-fold cross validation. Finally, the prediction ability of each optimised model was estimated by leave one-out cross validation. The best performance was provided by SVR, achieving an average prediction error of the retention times of 6.6% for the calibration set.

The participants proposed various methodologies based on different learning principles. Random forest is decision trees ensemble strategy, which build a consensus model from the aggregation of multiple decision trees. In that case, a divide-and-conquer strategy is used to model the dataset according to a hierarchy of tests. The choice of the variable to test is based on the ability to divide the remaining data subset. PLS regression is based on latent variables estimated as linear combinations of the measured variables and defining a low-dimensional subspace. The PLS model makes use of all variables by maximising the covariance between X and Y to capture the Y -related variation in X . Multiple linear regression associated with manual stepwise variable selection requires human intervention and expert

knowledge to get reliable results. SVR takes its origin from the statistical learning theory framework for building a linear model in a feature space by applying a kernel function (usually non-linear). For that purpose, a limited number of critical observations is selected, i.e. the support vectors. SVR has a great ability for generalization but direct interpretation is made difficult because the relation between the regression model and the original input space is not explicitly evaluated.

4. Conclusion

For the first time during the “Chimiometrie” congress organized by Chemometric group of the SFdS, a QSPR competition with the aim to predict reversed-phase retention time was proposed. The retention time constitutes a very helpful parameter for identifying unknown analytes when analysing complex samples analysis by LC coupled with HRMS. Moreover, the difficulty to distinguish compounds with the same molecular formula constitute a major bottleneck when investigating steroids. In that context, three-dimensional molecular descriptors were used to predict LSS chromatographic parameters. To cope with this problem, the four different solutions presented during the congress, in addition to being very different approaches, illustrated some of the difficulties currently encountered in QSRR. Table 4 shows the final results of the three finalists and the organiser. The prediction performance was evaluated based on the prediction error of the validation set consisting of 19 steroids. All final competitors obtained excellent prediction results, with the error in prediction below 10%. The best retention time prediction error for the external validation set was obtained by Participant 3, at 6.5%, which was more accurate than the challenge’s organiser, partially advantaged by his previous knowledge of the context of steroid analysis. These results illustrate the fact that linear model combined with clever variable selection can lead to very accurate prediction. Because the initial aim of an individual relative error below 5% in the experimental retention time could not be met by any of the participants, we believe that more specific descriptors that can integrate topological and conformational information are needed and may constitute the next step forward to improve QSPR models in RPLC.

Table 4 Summary of results. Each model was computed with different object and variable sizes. Participant 1; Participant 2; Participant 3: Organiser have developed both models for log k_w and S with 91 molecular descriptors and 72 steroids objects.

Model summary					Calibration set			Validation set		
Participant	Algorithm	Object	Variables		R ²	Error	RMSE	R ²	Error	RMSE
s	m	s	logk _w	S	t _R	t _R	C	t _R	t _R	C
1	RF	73	5	6	0.98	2.8%	0.61	0.89	8.6%	1.43
2	PLS	65	80	80	0.75	8.7%	2.81	0.88	7.9%	1.45
3	Stepwise MLR	50	6	5	0.98	2.9%	0.77	0.91	6.5%	1.12
Organisers	SVR	71	91	91	0.85	6.6%	1.73	0.92	7.0%	1.29

Acknowledgments

We would like to thank and congratulate the three finalists who spent time to develop models and prepare presentations for the challenge:

Pierre Dardenne (Walloon Agricultural Research Centre CRA-W) Gembloux, Belgium

Prof. Evelyne Vigneau and Dr. Philippe Courcoux (Oniris) Nantes

Corentin Harrouet, his colleagues (Master OPEX) as well as and Dr. Yves Lijour (UBO), Brest

Authors also would like to thanks Molecular Discovery for the concession of the VolSurf+ package to calculate molecular descriptors.

5. Reference

- [1] P. Dardenne, J.A. Fernández Pierna, A NIR data set is the object of a chemometric contest at 'Chimiométrie 2004', Chem. Intel. Lab Sys, 80 (2006) 236-242.
- [2] J.A. Fernández Pierna, F. Chauchard, S. Preys, J.M. Roger, O. Galtier, V. Baeten, P. Dardenne, How to build a robust model against perturbation factors with only a few reference values: A chemometric challenge at 'Chimiométrie 2007', Chem. Intel. Lab. Sys., 106 (2011) 152-159.
- [3] J.A. Fernández Pierna, P. Dardenne, Soil parameter quantification by NIRS as a Chemometric challenge at 'Chimiométrie 2006', Chem. Intel. Lab. Sys., 91 (2008) 94-98.
- [4] J.A. Fernández Pierna, L. Duponchel, C. Ruckebusch, D. Bertrand, V. Baeten, P. Dardenne, Trappist beer identification by vibrational spectroscopy: A chemometric challenge posed at the 'Chimiométrie 2010' congress, Chem. Intel. Lab. Sys., 113 (2012) 2-9.
- [5] J.A. Fernández Pierna, H. Duval, P. Valderrama, D.N. Rutledge, V. Baeten, P. Dardenne, A case study of extrapolation in NIR modelling — A chemometric challenge at 'Chimiométrie 2009', Chem. Intel. Lab. Sys., 106 (2011) 205-209.
- [6] G.M. Randazzo, D. Tonoli, S. Hambye, D. Guillarme, F. Jeanneret, A. Nurisso, L. Goracci, J. Boccard, S. Rudaz, Prediction of retention time in reversed-phase liquid chromatography as a tool for steroid identification, Anal. Chim. Acta
- [7] L.R.S.J.W. Dolan, High-Performance Gradient Elution: The Practical Application of the Linear-Solvent-Strength Model, (2007).
- [8] VolSurf+ <http://www.moldiscovery.com>.
- [9] B.D. Hudson, R.M. Hyde, E. Rahr, J. Wood, Parameter based methods for compound selection from chemical databases, QSAR, 15 (1996) 285-289.
- [10] P.J. Goodford, A computational procedure for determining energetically favorable binding sites on biologically important macromolecules, J. Med. Chem., 28 (1985) 849-857.
- [11] PyLSS <https://github.com/gmrandazzo/PyLSS>.
- [12] L. Breiman, Random Forests, Machine Learning, 45 (2001) 5-32.
- [13] R. Genuer, J.-M. Poggi, C. Tuleau-Malot, Variable selection using random forests, Pattern Recognit. Lett., 31 (2010) 2225-2236.
- [14] J. Jodoin, M. Demeule, L. Fenart, R. Cecchelli, S. Farmer, K.J. Linton, C.F. Higgins, R. Beliveau, P-glycoprotein in blood-brain barrier endothelial cells: interaction and oligomerization with caveolins, J. Neurochem., 87 (2003) 1010-1023.
- [15] F. Faassen, G. Vogel, H. Spanings, H. Vromans, Caco-2 permeability, P-glycoprotein transport ratios and brain penetration of heterocyclic drugs, Int. J. Pharm., 263 (2003) 113-122.