JOURNAL OF CHEMICAL INFORMATION AND MODELING

Article

pubs.acs.org/jcim

Study of Chromatographic Retention of Natural Terpenoids by Chemoinformatic Tools

Tiago B. Oliveira,[†,‡] Leonardo Gobbo-Neto,[§] Thomas J. Schmidt,[‡] and Fernando B. Da Costa*[,†]

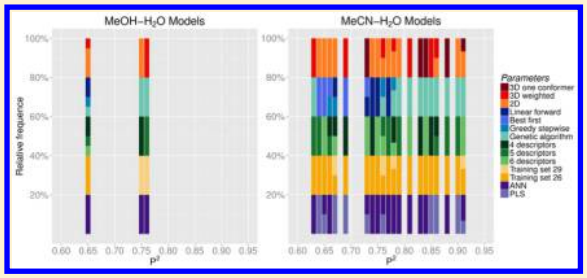
[†]AsterBioChem Research Team, Laboratory of Pharmacognosy, Department of Pharmaceutical Sciences of Ribeirão Preto, University of São Paulo (USP), Av. do Café s/n, 14040-903 Ribeirão Preto, SP, Brazil

[‡]Institute of Pharmaceutical Biology and Phytochemistry (IPBP), University of Münster, Corrensstr. 48, 48159 Münster, Germany

[§]School of Pharmaceutical Sciences of Ribeirão Preto, University of São Paulo (USP), Av. do Café s/n, 14040-903 Ribeirão Preto, SP, Brazil

Supporting Information

ABSTRACT: The study of chromatographic retention of natural products can be used to increase their identification speed in complex biological matrices. In this work, six variables were used to study the retention behavior in reversed phase liquid chromatography of 39 sesquiterpene lactones (SL) from an in-house database using chemoinformatics tools. To evaluate the retention of the SL, retention parameters on an ODS C-18 column in two different solvent systems were experimentally obtained, namely, MeOH−H_2O 55:45 and MeCN−H_2O 35:75. The chemoinformatics approach involved three descriptor type sets (one 2D and two 3D) comprising three groups of each (four, five, and six descriptors), two different training and test sets, four algorithms for variable selection (best first, linear forward, greedy stepwise, and genetic algorithm), and two modeling methods (partial least-squares regression and back-propagation artificial neural network). The influence of the six variables used in this study was assessed in a holistic context, and influences on the best model for each solvent system were analyzed. The best set for MeOH−H_2O showed acceptable correlation statistics with training R^2 = 0.91, cross-validation Q^2 = 0.88, and external validation P^2 = 0.80, and the best MeCN−H_2O model showed much higher correlation statistics with training R^2 = 0.96, cross-validation Q^2 = 0.92, and external validation P^2 = 0.91. Consensus models were built for each chromatographic system, and although all of them showed an improved statistical performance, only one for the MeCN−H_2O system was able to separate isomers as well as to improve the performance. The approach described herein can therefore be used to generate reproducible and robust models for QSRR studies of natural products as well as an aid for dereplication of complex biological matrices using plant metabolomics-based techniques.



INTRODUCTION

Comprehensive studies of low molecular weight analytes in complex samples of biological origin using modern hyphenated chromatographic/spectroscopic methods and aiming to obtain a holistic picture of metabolism in plants, animals, or microorganisms are nowadays subsumed under the term "metabolomics". Metabolomics studies encompass various approaches such as metabolic profiling, metabolic fingerprinting, and metabolic footprinting.[1] As an important prerequisite to metabolite identification, efficient tools for analyte detection with high sensitivity and compound specificity are needed.[2] Furthermore, it is necessary to develop screening procedures with the aim to rapidly and accurately identify compounds that are already known or to establish their degree of novelty. Such a procedure, which is called dereplication, is important, for example, to prioritize compounds for further isolation or to combine samples that contain unknown active components with similar properties in order to avoid reisolation and subsequent structural determination of a natural product that is already known.[3,4]

Quantitative structure−retention relationship studies (QSRR) represent a powerful tool that can be used to increase the speed of compound identification. Moreover, the use of statistical and chemometric methods[5] including chemoinformatic tools combined with in-house natural product libraries[6] are useful to reduce hit multiplicity during dereplication.[7] QSRR is the statistical correlation between chromatographic retention data and theoretical properties of molecules (descriptors). Application of linear or nonlinear regression, analysis of components,[8] or artificial intelligence techniques[9] on such data allows us to explain and predict the behavior of solutes (analytes) in chromatographic systems. The existence of such correlations makes it possible to predict retention times (t_R), retention indices (RI), and retention factors (k) of substances from their chemical structures as well as to assist in the identification of structures from a set of retention indices. These correlations can aid, for example, the dereplication of natural products in metabolomic studies.[2,10]

The theoretical properties of molecules that are the most recommended for studies of the molecular mechanism involved

Received: September 25, 2014
Published: December 17, 2014

DOI: 10.1021/ci500581q
J. Chem. Inf. Model. 2015, 55, 26−38

in reversed-phase liquid chromatography (like HPLC) retention are descriptors related to the reduced linear solvation energy relationship (LSER)-based model of Abraham.[11] Such models employ structural descriptors of analytes from molecular modeling and models correlating retention to the *n*-octanol−water partition coefficient (log *P*).[12]

However, with the fast development of chemoinformatics methods in the last years,[13] the speed and amount of descriptors that can be calculated have rapidly grown. Additionally, the available data-mining tools that can help in descriptor selection, and therefore contribute by adding information still inaccessible by thermodynamics, have also evolved.

With respect to the classification of QSRR studies, the pattern recognition methods most frequently used are linear methods like multiple linear regression,[14−19] partial least-squares regression (PLS),[15,17,20] and nonlinear methods such as artificial neural networks (ANN)[16,17] or kernel partial least-squares (KPLS).[21] All these methods have been proven to have many successful applications in QSRR studies.

In this work, we studied the retention behavior of sesquiterpene lactones (SL) from an in-house database called AsterDB. These compounds are characteristic secondary metabolites of the plant family Asteraceae (sunflower family), although they have also been found in other families such as Canellaceae,[22] Polygonaceae,[23] Schisandraceae,[24] Acanthaceae, Anacardiaceae, Apiaceae, Euphorbiaceae, Lauraceae, Magnoliaceae, Menispermaceae, and Rutaceae as well as Hepatidae (liverworts).[25] SL have an extremely broad range of conspicuous biological activities[26−29] and therefore have been intensely studied in a biomedical context.[30−34] Due to their great structural diversity, with various ring skeletons (around 30 different types are currently known[35,36]) and substitutional patterns, they are also used as chemical markers in various chemosystematic studies.[37−39] Therefore, such compounds are a challenge for chromatographic and chemoinformatics studies.

The performance of several QSRR models obtained with different statistical approaches was compared. Models for two different elution systems (MeOH−H_2O and MeCN−H_2O) were built using linear and nonlinear algorithms based on a variety of descriptors that were selected and calculated in different ways. Our goal was to find models with good accuracy of retention factor prediction by comparing different descriptor selection methods.

Besides understanding the most important molecular properties that would be involved in the retention of our data set, the novelty of this study is that it can be useful for dereplication as well as metabolomic studies using SL-containing plant extracts. Moreover, the results can also be used as a template for developing further models using different classes of natural products. For example, an important and direct application of our approach is the possibility of performing compound characterization in complex mixtures, like plant extracts. Estimation of retention times allied to chemical structure databases can intensely narrow the search and accelerate compound dereplication in such complex mixtures. A common fact that always occurs in the dereplication process in a LC-MS-based approach is when a chromatographic peak of a certain compound that is associated with is molecular weight return dozens of hits after a database search; in this situation, the prediction of the retention times of these hits can drastically decrease the number of compounds that are associated with the chromatographic peak, therefore helping the dereplication process. Moreover, this feature also would be useful if a certain desired compound is literature-new or unknown in a database because the prediction of its retention time can facilitate and speed up its location in a collection of several crude extracts like in metabolomic studies. This procedure can be made either in companies or academia.

MATERIAL AND METHODS

Logarithm of Retention Factor of Sesquiterpene Lactones (SL). The logarithm of the retention factor (log *k*) was calculated for the 39 SL available in our in-house pure compound library. Equation 1 shows the logarithmic form of the retention factor (*k*). According to IUPAC, it expresses how much longer a sample component is retarded by the stationary phase than it would take to travel through the column with the velocity of the mobile phase.[40]

$$\log k = \log \frac{t_R - t_M}{t_M} \quad (1)$$

where t_R and t_M are the total retention time and hold-up time, respectively; *k* is the retention factor; and $(t_R - t_M)$ is the adjusted retention time.

The 39 SL used in this study have been previously isolated from the Asteraceae species by members of our research group, and details about their structure elucidation are published elsewhere.[41−47] This group of compounds comprises structures with six different SL skeletal types (5 germacrolides, 3 melampolides, 8 guaianolides, 10 furanoheliangolides, 12 heliangolides, and 1 eudesmanolide).

The samples were prepared by dissolving 2 mg of each SL in 200 μL of a MeCN−H_2O 35:65 solution that was filtered through a 0.45 μm PTFE membrane; injection volume was 1.0 μL.

The SL were injected and analyzed by HPLC under the following isocratic conditions: MeOH−H_2O (55:45), 1.0 mL min^{-1} (system 1) and MeCN−H_2O (35:65), 1.3 mL min^{-1} (system 2) using an ODS C-18 Shimadzu column (4.6 mm × 250 mm, 5 μm). Compounds were simultaneously detected with a diode array detector (DAD) at 225 and 265 nm. 2,5-Dimethylphenol (DMP) was used as internal standard. All analyses were performed on a Shimadzu SCL-10Avp liquid chromatograph system equipped with a Shimadzu SPD-M10Avp DAD detector, operating with CLASS-VP software version 5.02.

Each retention time (t_R) generated one log *k* in each solvent system (log k_1 and log k_2 for systems 1 and 2, respectively). Data for the 39 SL are reported in Table 1.

Computer Hardware and Software. Statistic calculations were carried out in a 2.7 GHz Intel Pentium Dual Core E5400 with 3 GB of RAM using the OpenSuse 12.2 (openSUSE Project) operating system. MarvinSketch 5.12.4 (ChemAxon) was used for drawing the molecular structures.[48] Weka 3.6.6 (Waikato Environment for Knowledge Analysis)[49] and R statistical computing environment (version 2.15.3, Security Blanket)[50] were employed to select the attributes as well as to generate and validate the models using ANN. Descriptor selection with genetic algorithm (GA), conformational analyses, calculation of 3D molecular descriptors, and PLS models were performed with MOE (Molecular Operating Environment, v. 2010.11)[51] on an Intel Core 2 Duo personal computer with 2 GB of RAM running under Microsoft Windows XP.

Molecular Modeling and Geometry Optimization. The 2D chemical structures of the 39 SL (Figure 1) (with stereochemistry, Supporting Information 1) were first preoptimized using the 3D structure generator CORINA (COoRdINAtes).[52] The resulting geometries were further refined by means of a low

DOI: 10.1021/ci500581q
J. Chem. Inf. Model. 2015, 55, 26−38

Table 1. List of Investigated SL and Logarithm of Retention Factor in Systems MeOH−H_2O (55:45), log k_1, and MeCN−H_2O (35:65), log k_2

compound	name	log k_1	log k_2
1	eriofertin	0.502	0.502
2	ovatifolin	0.746	0.883
3	tagitinin A	0.536	0.436
4	viguilenin	0.780	0.660
5	budlein B	0.472	0.373
6	(5R,6R,7R,8R)-3-oxo-4β,10α-dihydroxy-8β-methylpropanoyloxy-guaia-11(13)-en-6α,12-olide	0.668	0.494
7	zaluzanin C	0.975	0.722
8	15-deoxybudlein A	1.134	1.270
9	atripliciolide tiglate	1.105	1.233
10	budlein A	0.682	0.564
11	budlein A tiglate	0.626	0.516
12	erioflorin	0.623	0.636
13	heliangin	0.460	0.778
14	leptocarpin	0.557	0.839
15	tagitinin E	0.642	0.639
16	tagitinin F	0.885	1.216
17	1β,10α-epoxy-3β-hydroxy-8β-(2′R,3′R-epoxyangeloyloxy)-germacra-4,11(13)-dien-6α,12-olide	0.446	0.406
18	1β,10α-epoxy-3β-acetoxy-8β-(2′R,3′R-epoxyangeloyloxy)-germacra-4,11(13)-dien-6α,12-olide	0.669	0.842
19	1β,10α-epoxy-3β-(2-methylbutanoyloxy)-8β-(2′R,3′R-epoxyangeloyloxy)-germacra-4,11(13)-dien-6α,12-olide	1.088	1.423
20	1β,10α-epoxy-3β-isovaleroyloxy-8β-(2′R,3′R-epoxyangeloyloxy)-germacra-4,11(13)-dien-6α,12-olide	1.347	1.574
21	1β,10α-epoxy-3β-(2-methylpropanoyloxy)-8β-(2′R,3′R-epoxyangeloyloxy)-germacra-4,11(13)-dien-6α,12-olide	0.832	1.271
22	1β,3α-dimethoxy-3β,10β-epoxy-8β-methylpropanoyloxy-germacra-4,11(13)-dien-6α,12-olide	1.189	1.258
23	heliangin acetate	1.074	1.267
24	2β-methoxy-3α-hydroxy-3,10β-epoxy-8β-methylpropanoyloxy-germacra-11(13)-en-6,12-olide	1.309	1.233
25	1β,2α-epoxytagitinin C	0.718	0.730
26	acanthospermolide angelate	1.214	1.451
27	2-oxo-8β-methacryloyloxy-guaia-3,10(14),11(13)-trien-6α,12-olide	1.054	0.998
28	2-oxo-8β-methacryloyloxy-guaia-1(10),3,11(13)-trien-6α,12-olide	1.381	1.160
29	2-oxo-8β-methacryloyloxy-10α-hydroxy-guaia-3,11(13)-dien-6α,12-olide	0.770	0.816
30	2-oxo-8β-methacryloyloxy-10β-hydroxy-guaia-3,11(13)-dien-6α,12-olide	0.609	0.427
31	2-oxo-8β-tigloyloxy-guaia-1(10),3,11(13)-trien-6α,12-olide	1.357	1.376
32	2α,3β-dihydroxy-8β-methacryloyloxy-germacra-1(10),4,11(13)-trien-6α,12-olide	0.755	0.435
33	parthenolide	1.076	1.154
34	niveusin A	0.691	0.376
35	tagitinin C	0.855	0.787
36	1α-methacryloyloxy-2β,8α-dihydroxy-3α,4α-epoxy-eudesm-11(13)-en-6α,12-olide	0.558	0.328
37	enhydrin	0.831	1.177
38	uvedalin	1.188	1.437
39	2-oxo-8β-epoxyangeloyloxy-guaia-1(10),3,11(13)-trien-6α,12-olide	0.785	0.903

mode dynamics (LMD) conformational search using the MMFF94x force field and standard settings in MOE.

Generation of Descriptors. 2D Coordinate matrices of all atom positions in the molecules were used for the calculation of the theoretical descriptors. Global descriptors and 2D autocorrelation coefficients using polarizabilities as atom pair properties were calculated using Adriana.Code 2.2.6 (Molecular Networks).[53] Molecular properties, information indices, and constitutional descriptors were calculated using Dragon 5.5 (Talete).[54] Functional group counts, atom type electrotopological states, connectivity indices, fragment complexity, rotatable bonds counts, molecular properties, complexity of a molecule, and 2D matrix-based descriptors were calculated using PaDEL 2.15.[55] This set of descriptors (655) henceforth will be called by "2D-descr" (Supporting Information 2). Descriptors obtained on the basis of Cartesian coordinate matrices of fully optimized 3D molecular structures such as surface area, volume, shape, and conformation-dependent charge descriptors were calculated using MOE. The set of 123 3D-descriptors calculated from the single lowest minimum energy conformer found during the conformational search will be called "3D-1conf" (Supporting Information 3), and the same 123 3D-descriptors calculated from a Boltzmann weighted conformational ensemble will be termed "3D-weigh" (Supporting Information 4).[56] This latter series was calculated using a maximum of 10 different conformers of each molecule, for which all descriptors were calculated and weighted according to their AM1 energy using the Boltzmann equation[57]

$$p_i = \frac{e^{-\Delta E_i/RT}}{\sum_{i=1}^{t} e^{-\Delta E_i/RT}} \quad (2)$$

where the probability p of finding a compound in the particular state i in an equilibrium of t states (conformations) at a given absolute temperature T is related to the energy difference of state i from the global minimum ΔE, and R is the universal gas constant.

Descriptor Selection. As in any QSRR study, the selection of molecular descriptors is the most crucial step that affects the quality of the models.[18] The three descriptor sets previously

DOI: 10.1021/ci500581q
J. Chem. Inf. Model. 2015, 55, 26−38

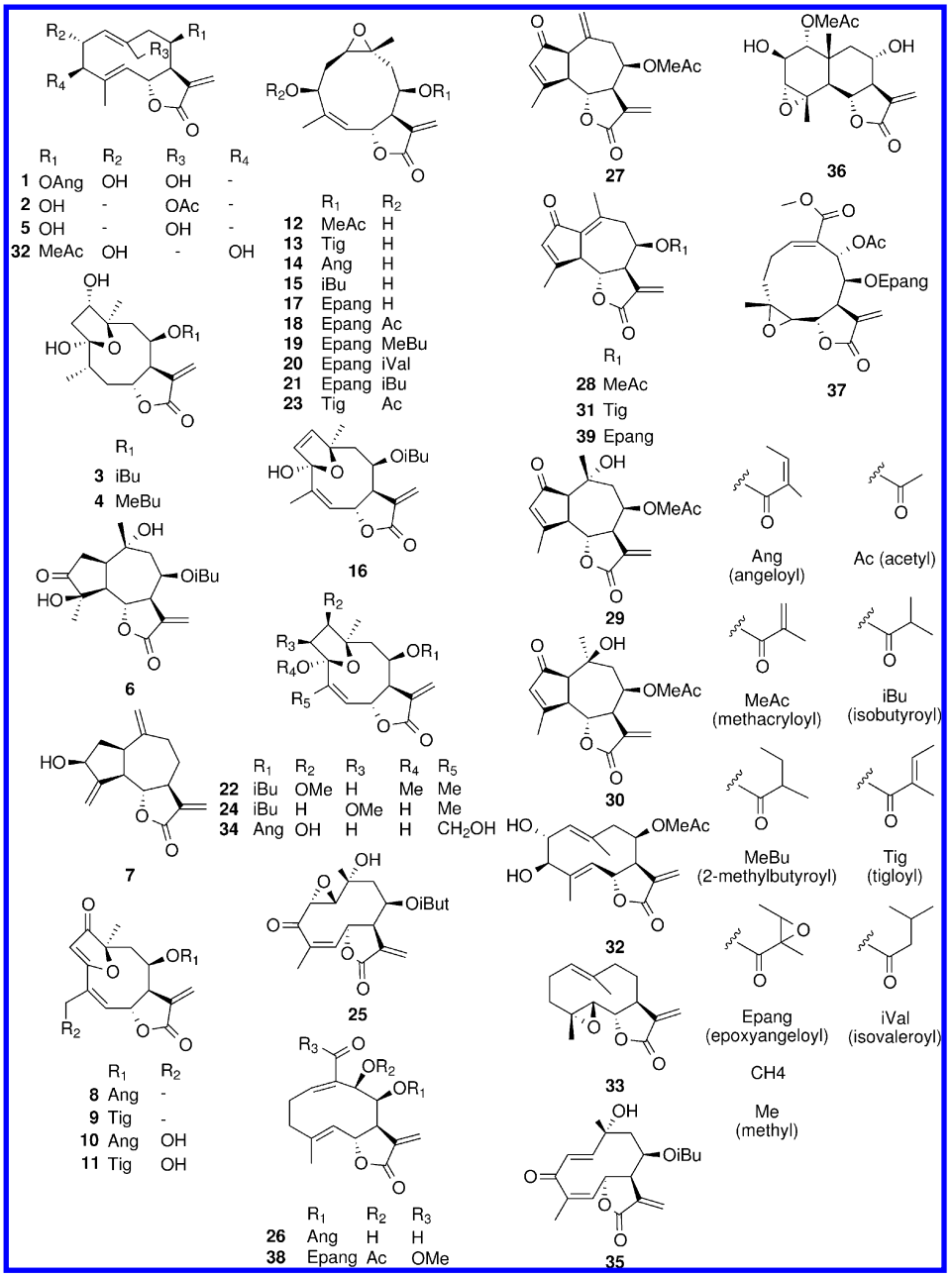


Figure 1. Chemical structures of the 39 sesquiterpene lactones (SL).

described (2D-descr, 3D-1conf, and 3D-weigh) were pretreated with the Caret package 5.15−87[58] using the *nearZeroVar* and *findCorrelation* functions. In the first step, all descriptors with zero variance and variance close to zero were excluded. In the second step, descriptors with a high degree of pairwise correlation that could create defective models were excluded. For theoretical descriptors, it is not uncommon to have many very large correlations among each other.[58] In this work, using these two steps and a threshold of 0.90 with *findCorrelation* functions, 565 descriptors were eliminated from the 2D-descr group, 62 from the 3D-1conf group, and 60 from the 3D-weigh group.

After that, four stepwise algorithms were used to select the most relevant descriptor combinations for the QSRR models: (i) best first, (ii) greedy stepwise, (iii) linear forward, and (iv) genetic algorithm (GA). The first three methods (Weka) evaluate the worth of a subset of attributes by considering the individual predictive ability of each feature along with the degree of redundancy among them and with 10-fold cross-validation.[59] The fourth method (GA, MOE) will be briefly discussed below. These four algorithms were also used to reduce the subsets of descriptors to an appropriate size (four, five, or six descriptors) using different selection criteria.

The best first selection (i) uses an algorithm that starts with a single and arbitrary descriptor that it compares with log *k* correlation and then finds a better solution by backward and forward search and adds a single new descriptor. The greedy stepwise selection algorithm (ii) starts from an arbitrary point in the descriptor matrix and performs a greedy forward search through the space of attribute subsets. It stops when the addition of any remaining attribute results in a decrease in log *k* correlation. The linear forward selection (iii) is an extension of the best first method that takes a restricted number of *n* descriptors. A fixed set that selects a fixed number *n* of descriptors was used. All of these three algorithms were evaluated by

DOI: 10.1021/ci500581q
J. Chem. Inf. Model. 2015, 55, 26−38

considering the individual predictive ability of each descriptor along with the degree of redundancy between them by 10-fold cross-validation.

The GA algorithm (MOE script GA.svl as available via the CGC/MOE exchange Web site) directly optimizes a family or population of descriptor combinations by means of their performance in multiple linear regression (MLR). In this procedure, each descriptor within a particular combination represents a "gene" within the model. Individual genes can be randomly exchanged for new ones in a predefined number of models ("mutation"), thus creating new models. Exchange of gene groups (i.e., more than one descriptor) between already existing models represents "crossing over" events. Evolving a population of such models over a large number of "mutation"/"cross over" cycles under consequent re-evaluation after each cycle and elimination of models with poor performance, for the sake of keeping models with better performance, automatically leads to an optimization of the population with respect to the parameter chosen to represent performance. The performance criterion chosen here was a decrease in the lack of fit (LOF) in the MLR of the descriptors versus the target value, log k. The chosen population size for one optimization run was 100 models. The termination criterion was either that a predefined LOF was reached or that a maximum number of 1000 cycles had been performed. Upon termination, the whole population was subjected to leave-one-out cross-validation, and the descriptor combinations with the highest cross-validation coefficient of determination were chosen for further evaluation.

The aim in this step was to select by using each method (i, ii, iii, and iv) at least three different groups comprising four, five, and six descriptors by each selected method.

Model Building. The SL structures were ordered sequentially from the lowest to the largest log k, and two groups were selected for testing and training. The training and test sets were selected in such a way that the overall distribution of the samples was preserved whenever possible. In the first group, 26 structures were selected as the training set and 13 as the test set, and in the second group, 29 structures were used as the training set and 10 as the test set. The test sets were used to validate the models (see next section).

Then, QSRR models were built using the two different training sets. The selected descriptors and the log k were correlated by PLS and a nonlinear model, namely, ANN.

In PLS, a linear method, first the vectors were generated by a Krylov sequence $\{b, Sb, S^2b, ..., S^{A-1}b\}$. After this, a weight matrix $V_A = (v_1, v_2, ..., v_A)$ was constructed by the Gram−Schmidt orthogonalization, where v_i is a column vector of length n, and A, the degree of the PLS fit, is an integer less than or equal to n. Then, to get the A^{th} PLS coefficient vector, we solved for[51,60]

$$a = V_A(V_A^{tr}SV_A)^{-1}V_A^{tr}b \tag{3}$$

where tr means transpose.

An algorithm to determine successive PLS fits will terminate at the A^{th} step when $\|v_{A+1}\| = 0$. One may, however, wish to use an even lower order PLS regression vector to fit a linear model. If the algorithm continues until $A = n$, the weight matrix is square and of full rank. The resulting regression vector is the ordinary least-squares.[51]

In ANN, a multilayer perceptron with the algorithm for determining the weights called back-propagation was used in Weka. In each node (or neuron), the sigmoid function was used, and the equations for this model were defined by[61]

$$f(x) = \frac{1}{1 + e^{-x}} \tag{4}$$

To apply the gradient descent procedure, the error function is

$$E = \frac{1}{2}(y - f(x))^2 \tag{5}$$

where $f(x)$ is the network's prediction obtained from the output node, and y is the log k (normalized between 0 and 1). The weight for each node is calculated for a perceptron with one hidden layer by derivation using the following equation

$$\frac{dE}{dw_{ij}} = (f(x) - y)f'(x)w_jf'(x_i)a_j \tag{6}$$

where $f'(x)$ denotes the derivative of $f(x)$, w_i and w_{ij} are the weights, and a_j is a vector. The weight is calculated for every training instance. The changes are associated with values regulated during the tuning of the neural network as momentum, learning rate, training time, and number of hidden layer. The weight of the previous instance is multiplied by the learning rate, and the outcome is then subtracted from the next value of w_{ij}. Because of this error propagation mechanism, this version of the generic gradient descent strategy is called back-propagation.[61]

Model Validation. In the present study, 10-fold cross-validation was performed to evaluate the robustness and validity of models as well as their internal predictivity. Test sets with structures not used in the QSRR model development (see above) were then used for external validation. Subsequently, the goodness of fit of the models was evaluated using the following statistical parameters: squared correlation coefficient for model fitting (training-set), R^2; squared correlation coefficient for cross validation (training-set), Q^2; squared correlation coefficient for test set, P^2; mean absolute error (MAE); and root mean square error (RMSE).

Consensus Modeling. From the best models in each chromatographic system, consensus models 1 and 2 were built for the MeOH−H_2O system and models 3 and 4 for the MeCN−H_2O system (each pair containing 29 and 26 structures for the training sets, respectively). Consensus model 5 was built for the MeCN−H_2O system using 26 structures for the training set and taking into account the best statistical performances of 2D descriptors as well as the separation of isomers provided by the 3D descriptors.

The consensus models were built as follows: Consensus1 (MeOH models 40 and 140), Consensus2 (MeOH models 15, 31 and 143), Consensus3 (MeCN models 96, 136 and 142), Consensus4 (MeCN models 19, 23, 27, 31, 35, 37, 39, 42, 43, 45, 47, 59, 87, 91, 95, 133, 137 and 143), and Consensus5 (MeCN models 45, 47 and 143). The averages of predictions for each chemical structure were obtained, and R^2, P^2, MAE, and RMSE were calculated for each consensus model. See Supporting Information 5, 6, and 7.

RESULTS AND DISCUSSION

In summary, six parameters were evaluated in this study:

- two different solvents systems (MeOH−H_2O 55:45 and MeCN−H_2O 35:65)
- three descriptor sets (2D-descr, 3D-1conf, and 3D-weigh)
- two different training and test sets (26:13, and 29:10)
- four algorithms for variable selection for one descriptor group (best first, linear forward, and greedy stepwise, GA)
- three different model sizes (four, five, and six descriptors)
- two modeling methods (PLS and ANN)

DOI: 10.1021/ci500581q
J. Chem. Inf. Model. 2015, 55, 26−38

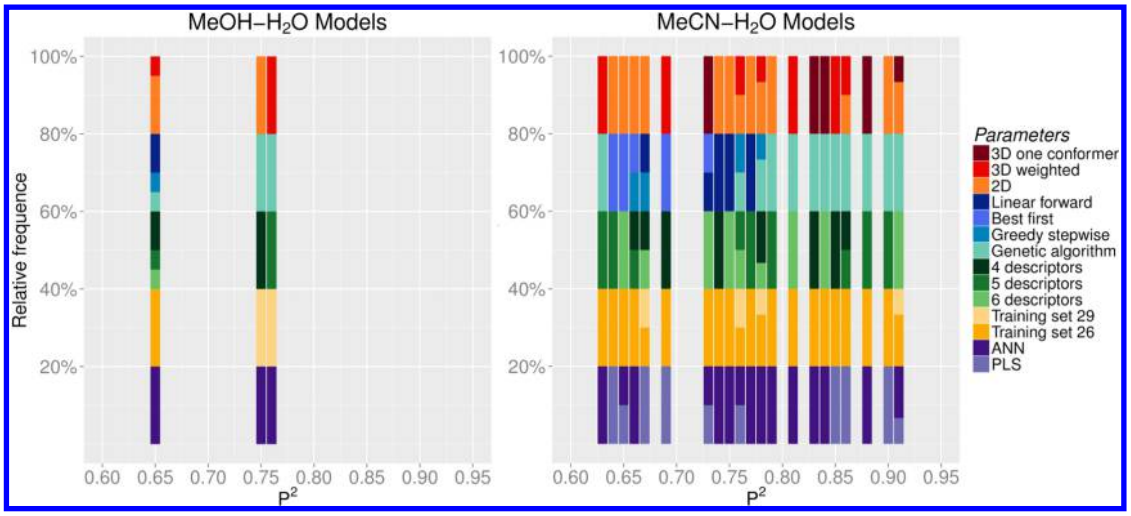


Figure 2. Cumulative charts of share of models for each solvent among the best models. The horizontal axis displays the coefficient of determination by external validation (P^2). The vertical axis displays the share of each parameter (in percentage). Larger areas (colored vertical bars) demonstrate more successful approaches.

Given all possible combinations using these six parameters, 288 different models were generated and compared among each other. The 40 top performing models were selected, six from the MeOH−H2O models and 36 from MeCN−H_2O. The selection was performed based on a P^2 value higher than 0.60, and $R^2 > Q^2 > P^2$ (Supporting Information 5, 6, and 7).

Figure 2 shows for each solvent system the relative frequency of the various investigated parameters (%, *y* axis) plotted against the resulting models' predictive quality (external validation P^2, *x* axis). Among the used methods, the best performance for the MeOH−H_2O system ($P^2 = 0.76$) was achieved with ANN, while three models with $P^2 = 0.91$ were obtained for MeCN−H_2O, two with ANN, and one with PLS.

GA was by far the best method for variable selection for our set of structures and in both solvent systems. This can be explained by two main reasons. GA used, as selection criterion, sets of variable numbers (four, five, and six descriptors), that is, LOF in MLR with a set of descriptors, while the other selection methods used the individual capacity of each descriptor. This in fact reduces machine time for selection when used with large data.

In Figure 2, by considering the three sets of descriptors (2D-descr, 3D-1conf, and 3D-weigh), it is possible again to observe no difference between the two solvent systems. For MeOH−H_2O and MeCN−H_2O, the use of descriptors with 2D coordinates generated more viable models. This would appear a relative advantage because there is no need to generate the 3D coordinates; therefore, it reduces one laborious step in building models. However, 2D-descr is not capable of separating epimers (e.g., structure pair **29** and **30**) and geometric isomers (e.g., structure pairs **8** and **9**, **10** and **11**, and **13** and **14**, all of them showing isomerism in their side chain esters, Figure 1), but it may explain the more simple interactions among the analytes.

In two MeOH−H_2O models (Figure 3), furanoheliangolide **24** is one of the worst predicted, with absolute errors of 0.426 and 0.390 in models MeOH140 and MeOH40, respectively. Model MeOH40 (2D) was able to predict correctly the log *k* of all melampolides. Among the heliangolides, 23% (structures **11**, **12**, **20**, **21**, and **24**) of the absolute value errors were higher than 0.13.

It is noteworthy to observe that there are far more MeCN−H_2O models (Figure 4) with good correlation results than MeOH−H_2O ones (Figure 3). It is therefore an interesting question as to why the retention data obtained in MeOH−H_2O are much more difficult to describe in a QSRR model than those obtained in MeCN−H_2O. The answer could lie in the fact that methanol (a protic solvent) and water interact among each other by hydrogen bonding to form species that are different from those when only pure water or methanol are taken into account. Therefore, the system could in fact be considered a ternary solvent (methanol, water, and methanol−water) as suggested by Bosch et al.[62] This peculiar condition in fact generates a much more complex interaction among the analytes, the two solvents, and the stationary phase, in comparison to when MeCN (aprotic solvent) and water are used as the mobile phase.[63]

For comparative purposes, all descriptors selected by different selection algorithms were analyzed in radar graphs (Figure 5). It is important to point out that the descriptors with the greatest intensity in the graphs are present in the best models. In this graph, it is possible to evaluate the descriptors selected by solvent system and type of descriptor.

Independent of the dimensionality of the descriptors, the factors with great importance for correlation with log *k* were descriptors of hydrophobicity and hydrophilicity and the presence of polar groups in the structures along with their intra- and extramolecular interactions.

It can be observed that the 2D descriptors (Figure 5) that were selected most frequently for the solvent system MeOH−H_2O were ALOGP2 (squared Ghose−Crippen *n*-octanol−water partition coefficient), SIC3 (structural information content, neighborhood symmetry of 3-order[64]), maxsssCH (maximum atom-type e-state: >CH−[65,66]), maxHBint7 (maximum e-state descriptors of strength for potential hydrogen bonds of path length 7[65,66]), and BIC4 (bonding information content, neighborhood symmetry of 4-order[67]). The 2D descriptors that were selected most frequently for the solvent system MeCN−H_2O were also ALOGP2, maxHBint7, and BIC4, while the RBF descriptor (rotatable bond fraction) also appeared.

As shown in Figure 5, the 3D-1conf descriptors, vsurf ID7 (distance between the center of mass of a molecule and the center of the hydrophobic regions around it at −1,4 *e*[68,69]), vsurf D1 (hydrophobic region of the molecules and is defined when a dry probe is interacting with a target molecule at −1,2 *e*[68,69]), npr1 (normalized principal moment of inertia ratio, *pmi*1/*pmi*2[51]), glob

DOI: 10.1021/ci500581q
J. Chem. Inf. Model. 2015, 55, 26−38

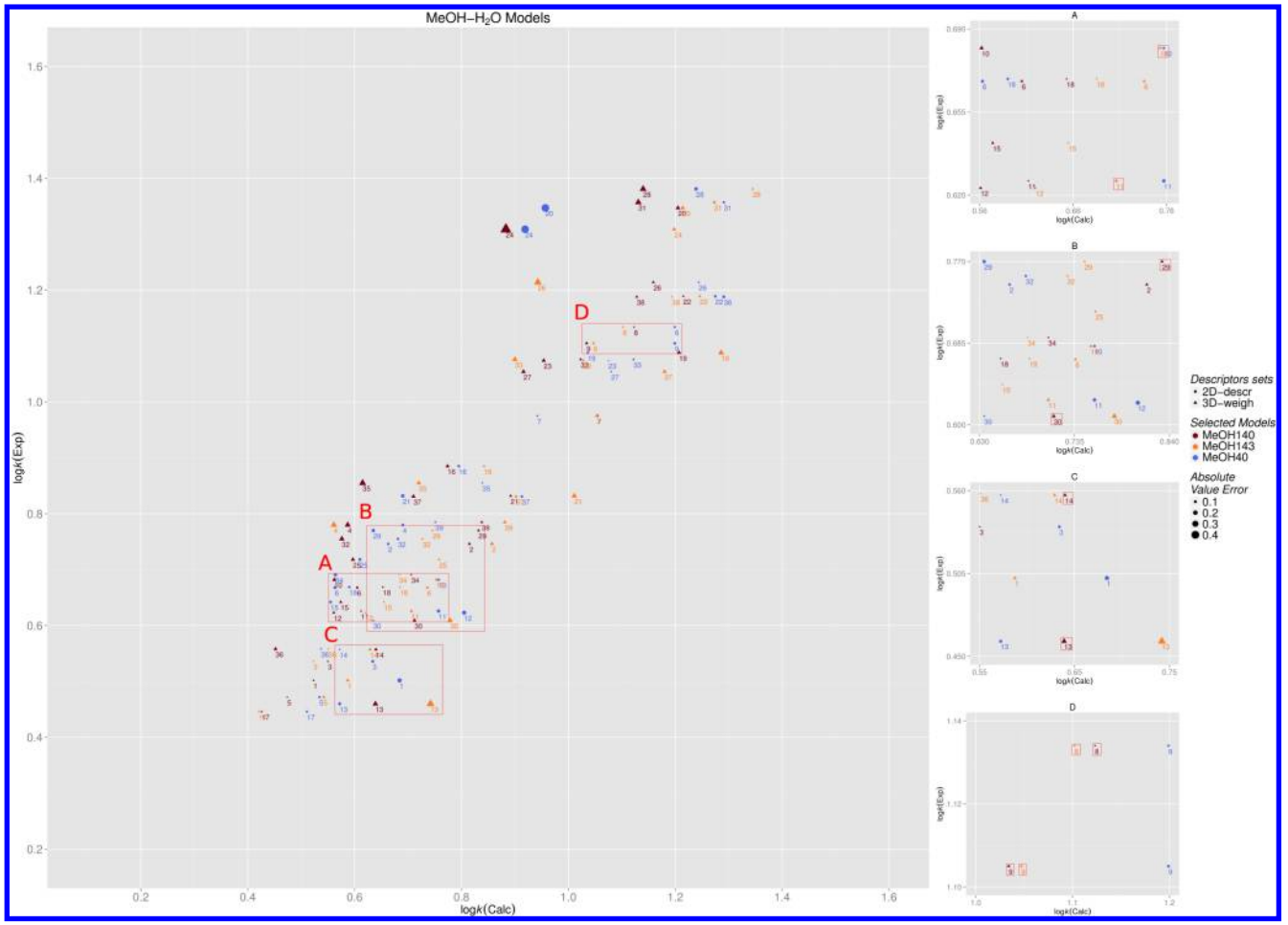


Figure 3. Experimental vs calculated log *k* values from MeOH140, MeOH143, and MeOH40 (Table 2). Models for the solvent system MeOH–H_2O and the 39 SLs combined from the training and external test sets. In the expanded plots (regions A–D marked by red retangles), it is possible to observe the models' ability to discriminate between epimers and geometric isomers (structures pairs A, **10** and **11**; B, **29** and **30**; C, **13** and **14**; and D, **8** and **9**).

(globularity), and ASAP7 (fractional accessible surface area due to atoms in partial charge interval >0.3 e[70]) were selected for the MeOH–H_2O solvent system, and the 3D-1conf descriptors, vsurf D1, vsurf CW3, CW2, and CW1 (capacity factor of the molecules and they are calculated at the energy level of −1.0, −0.5, and −0.2 kcal/mol, respectively[68,69]), and npr1 were selected for the MeCN–H_2O solvent system.

Finally, the 3D-weigh descriptors selected were vsurf Wp6 (polar volume of the molecule and calculated at −4.0 kcal/mol energy level[68,69]), vurf DD12 (contact distances of vsurf DDmin), vsurf D1, vsurf CW2, ASAP1 (fractional accessible surface area due to atoms in partial charge interval 0 to 0.05 e[70]), and ASAN2 (fractional accessible surface area due to atoms in partial charge interval −0.05 to −0.1 e[70]) for the MeOH–H_2O solvent system, and vsurf R (measure of molecular wrinkled surface, rugosity[69]), vsurf CW2 and CW1, glob, and ASAP5 (fractional accessible surface area due to atoms in partial charge interval 0.2 to 0.25 e[70]) were selected for the solvent system MeCN–H_2O

The classification of the best models is given by external validation, that is, the ability of the model in predicting the log *k* of a group that has not been part of the construction of the model. In other words, the best models were selected on the basis of the highest P^2, least mean absolute error (MAE), and root mean square error (RMSE).

The best models for the MeOH–H_2O solvent system given in Table 2 show the descriptors provided in Figure 5. However, among the selected models, it was observed that the descriptors Eta Epsilon 5 (a measure of electronegative atom count[65,66]) and nR03 (number of rings of 3-membered) were selected by GA for the model MeOH40 and std dim2 (standard dimension 2, square root of the second largest eigenvalue of the covariance matrix of the atomic coordinates; a standard dimension is equivalent to the stardard deviation along a principal component axis) and vsurf CW5 and CW6 (capacity factor of the molecules, calculated at the energy level of −3.0 and −4.0 kcal/mol, respectively[68,69]) were selected by GA for the model MeOH140.

In Table 3, MeCN–H_2O gives the descriptors provided in Figure 5, the descriptors IC2 (information content index, neighborhood symmetry of 2 order[64]), SHBint2, SHBint7, and SHBint8 (product of the atom-level e-state for the acceptor atom and the atom-level hydrogen He-state of the hydrogen atom on the donor separated by 2-skeletal, 7-skeletal, and 8-skeletal bonds[71]), and minsCH3 (minimum value of the −CH3 atom e-state (electrotopological state) descriptor) in models MeCN47 and MeCN45. The model with the highest P^2 and 3D-1conf (MeCN96) added ASAN7, ASAP1, and ASAP4 descriptors (fractional accessible surface area due to atoms in partial charge interval < −0.30, 0 to 0.05, and 0.15 to 0.2 e, respectively[70]), pmi1 (first diagonal element of diagonalized moment of inertia tensor[51]), and vsurf ID8 (distance between the center of mass of

DOI: 10.1021/ci500581q
J. Chem. Inf. Model. 2015, 55, 26−38

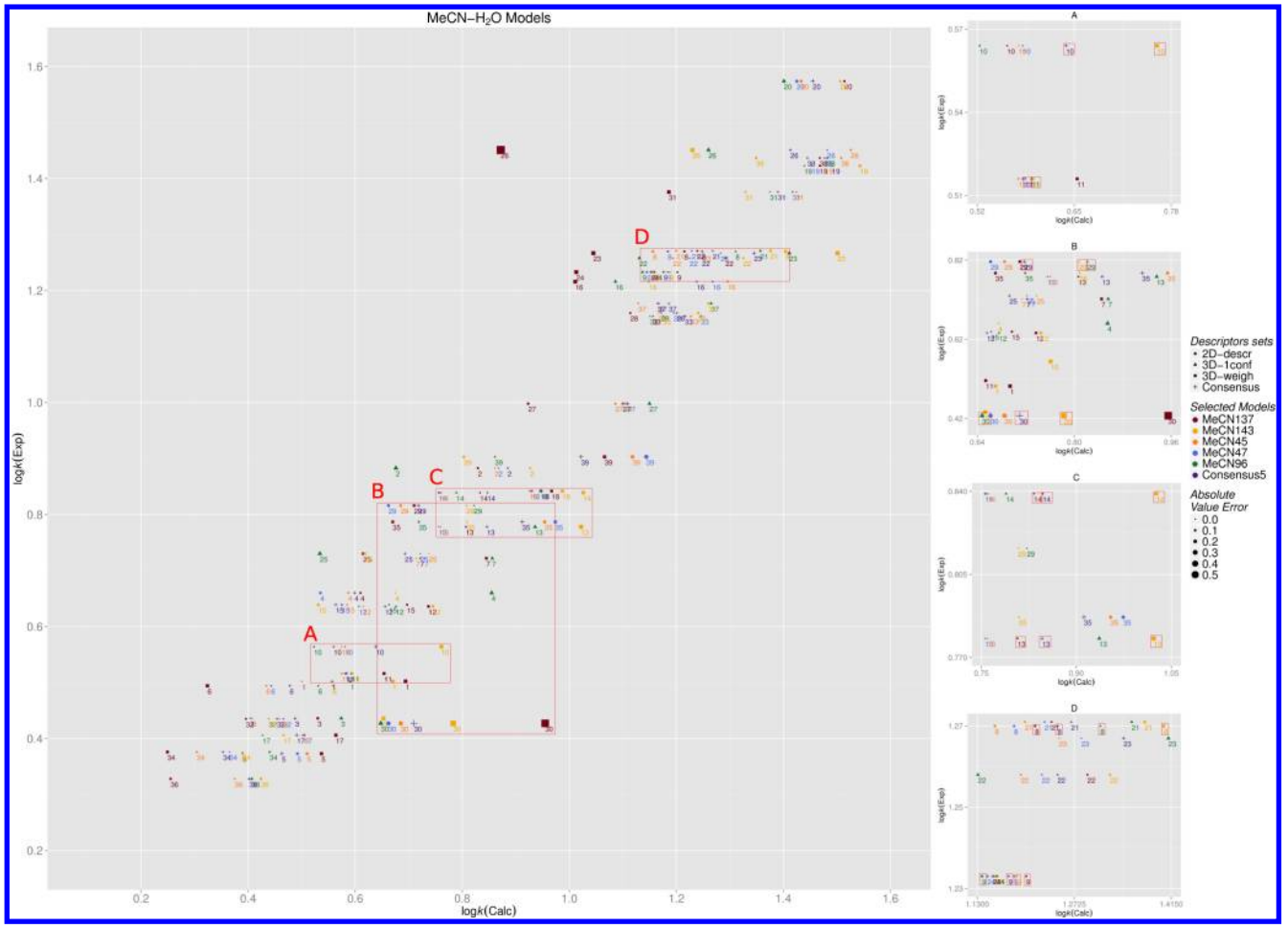


Figure 4. Experimental vs calculated log *k* values from MeCN137, MeCN143, MeCN45, MeCN47, MeCN96, and Consensus5 (Table 3). Models for the solvent system MeCN–H_2O and the 39 SLs combined from the training and external test sets. In the expanded plots (regions A–D marked by red retangles), it is possible to observe the models' ability to discriminate between epimers and geometric isomers (structures pairs A, **10** and **11**; B, **29** and **30**; C, **13** and **14**; and D, **8** and **9**).

a molecule and the center of the hydrophobic regions around it at −1,6 *e*[68,69]). The model with highest P^2 and 3D-weigh (MeCN137) added ASAN6 (fractional accessible surface area due to atoms in partial charge interval −0.25 to −0.3 *e*[70]), ASAP5, pmi1, vsurf HB1 (describes the H-bonding capacity of a molecular target, as obtained with a polar probe[68,69]), and vsurf ID8.

The MeOH–H_2O models (Table 2) showed minimally acceptable correlation statistics with training R^2 = 0.94–0.74, cross-validation Q^2 = 0.91–0.69, and external validation P^2 = 0.76–0.65. The root mean square error (RMSE) for the MeOH140 model was 0.10 for the training set, 0.13 for the cross-validation, and 0.19 for the external validation.

The MeCN–H_2O models (Table 3) showed much higher correlation statistics with training R^2 = 0.96–0.92, cross-validation Q^2 = 0.93–0.86, and external validation P^2 = 0.91–0.81. The RMSE for the ANN, GA, and 2D-descr was 0.08 for the training set, 0.10 for the cross-validation, and 0.11 for the external validation.

The 3D descriptors described above are related to the influence of tridimensionality on log *k* for the models. Additionally, the interactions depend on the molecular hydrophilic area and concentration of exposed polar groups. For example, as expected, neither model that used 2D-descr was able to correctly differentiate the geometric isomers with the angelate and tiglate moieties by log *k*. The model MeCN143 (Table 3), which was built with 3D-weigh descriptors, was the only one that was able to predict the correct sequence of the epimers (structure pair **29** and **30**) and geometric isomers structures pairs **8** and **9**, **10** and **11**, and **13** and **14**.

Although the statistical performances of each of the individual best models (Tables 2 and 3) are comparable among each other, it is difficult to judge which model is better than others and which one should be chosen to predict the log *k* of new SLs. Besides statistical performance, in this study, the separation and the prediction of the correct order of epimers and geometric isomers is another important feature to chose the best model(s). In QSAR studies, many authors recommend the use of consensual models with the goal of increasing stability and generalizability of the predictions, and these consensus models normally had higher external prediction power as compared to any individual model used in the consensus prediction because the errors cancel each other. For this reason, following a strategy that was proven to be successful in previous studies,[72−75] in this study, several consensus models were developed by averaging the individual models considering two different training and test sets and focusing on features that differentiate among stereoisomers. Although the consensus models for the MeOH–H_2O system showed an improved statistical performance (Table 2), only one model

DOI: 10.1021/ci500581q
J. Chem. Inf. Model. 2015, 55, 26−38

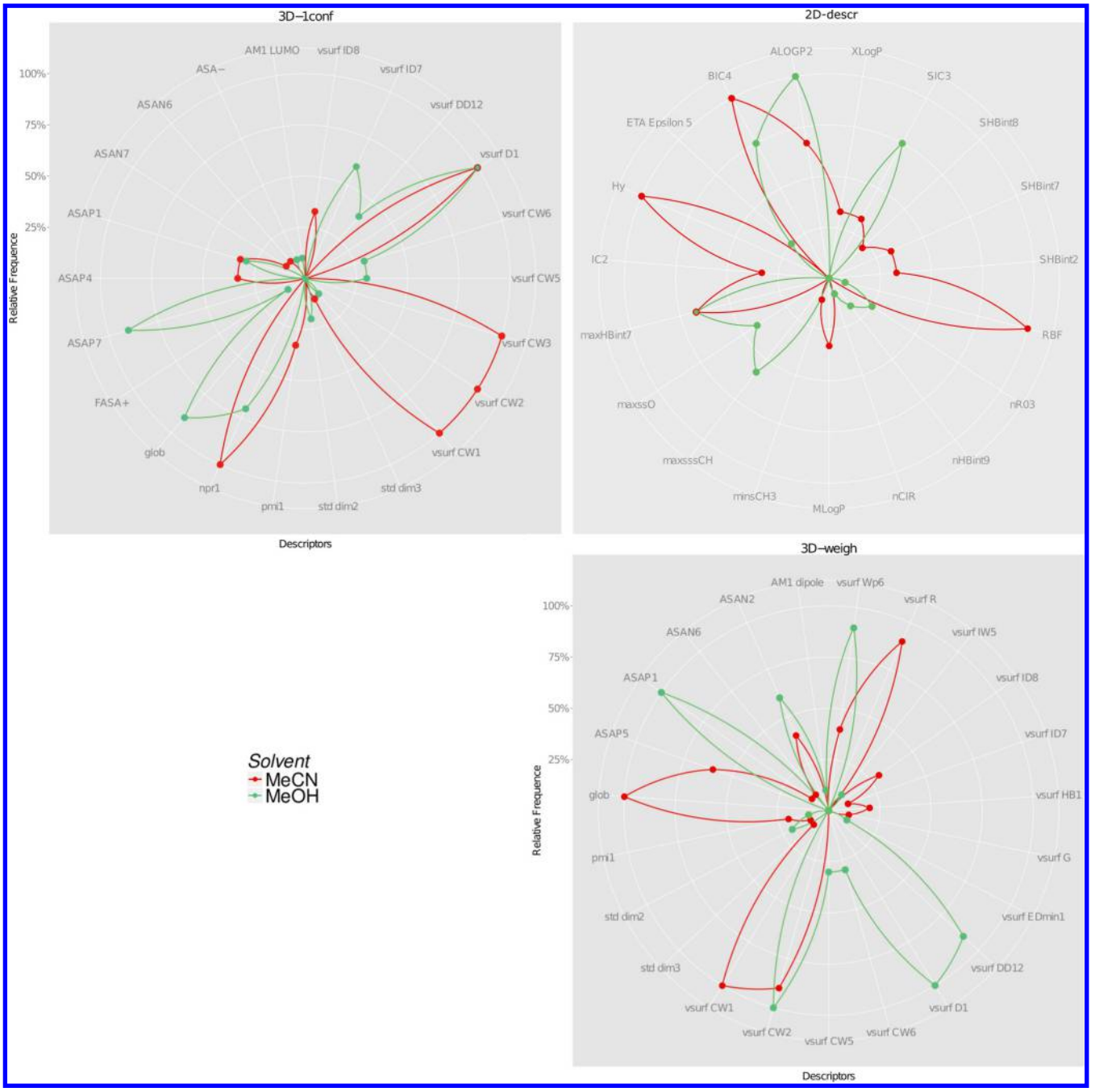


Figure 5. Descriptors selection plot. The distance of a descriptor from the center of the chart correlates with its importance (frequency of selection by the diverse techniques used) and reflects its degree of correlation with log *k*.

for the MeCN−H_2O system (Consensus5) was able to separate isomers (it contains 3D-weigh descriptors) and also improve performance (Table 3).

In this work, it was possible to build good models that are able to predict log *k* of complex terpenoid structures. All models built show information about charge and volume or charge regardless of dimensionality (2D, 3D-1conf, or 3D-weigh). However, it is emphasized that the use of Q-based descriptors fractional accessible surface area descriptors (Q_frASAs) and 3D-descriptors calculated from a Boltzmann weighted conformal ensemble are being reported for the first time in QSRR studies.

We would recommend our best results with Consensus5 in the MeCN−H_2O system for further QSRR studies with SLs provided that a larger or different data set includes structures with reasonable degree of complexity or from different skeletal types. Therefore, this information can be used to assist dereplication of plant extracts in metabolomic studies in at least two concrete situations: (i) to narrow a database search and accelerate the identification of unknowns in plant extracts by comparing their molecular masses and log *k* obtained by LC/MS analyses and (ii) to locate known/desirable compounds in plant extracts after the prediction of their log *k* from their structures and database search combined with molecular masses. The limitation of our approach is that it can be used only for SLs. However, it can deal with isomers and can be applied for an enhanced modeling set provided it contains related chemical structures.

DOI: 10.1021/ci500581q
J. Chem. Inf. Model. 2015, 55, 26−38

Table 2. QSRR Best Models for Each Variable Obtained Using MeOH−H_2O Solvent System[a]

model	descriptors	type of descriptors	selection	modeling	training set	R^2	MAE	RMSE	Q^2	MAE	RMSE	P^2	MAE	RMSE
MeOH140	ASAN2, ASAP1, std dim2, vsurf CW5, vsurf CW6	3D-weigh	GA	ANN	29	0.88	0.08	0.10	0.77	0.11	0.13	0.76	0.15	0.19
MeOH40	ALOGP[2], Eta Epsilon 5, maxsssCH, nR03	2D-descr	GA	ANN	29	0.94	0.06	0.07	0.91	0.07	0.08	0.75	0.11	0.16
MeOH143	AM1 dipole, pmi1, vsurf CW5, vsurf CW6, vsurf EDmin1, vsurf IW5	3D-weigh	GA	ANN	26	0.92	0.06	0.08	0.79	0.10	0.13	0.65	0.13	0.16
MeOH15	ALOGP[2], SIC3, BIC4, maxHBint7	2D-descr	GS/LF	ANN	26	0.74	0.09	0.12	0.71	0.10	0.13	0.65	0.13	0.16
MeOH31	ALOGP[2], SIC3, BIC4, maxHBint7, maxssO	2D-descr	LF	ANN	26	0.74	0.09	0.12	0.69	0.11	0.14	0.65	0.13	0.16
Consensus1	–	–	–	–	29	0.92	0.06	0.08	–	–	–	0.77	0.13	0.17
Consensus2	–	–	–	–	26	0.90	0.07	0.09	–	–	–	0.74	0.12	0.14

[a] R^2, coefficient of determination (set = 26); Q^2, coefficient of determination for cross validation 10-folds; P^2, coefficient of determination for external validation (set = 13); MAE, mean absolut error; and RMSE, root mean square error.

Table 3. QSRR Best Models for Each Variable Obtained Using MeCN−H_2O solvent system[a]

model	descriptors	type of descriptors	selection	modeling	training set	R^2	MAE	RMSE	Q^2	MAE	RMSE	P^2	MAE	RMSE
MeCN47	ALOGP[2], IC2, SHBint2, SHBint7, SHBint8, minsCH3	2D-descr	GA	ANN	26	0.96	0.06	0.08	0.93	0.09	0.10	0.91	0.08	0.11
MeCN45	ALOGP[2], IC2, SHBint2, SHBint7, SHBint8, minsCH3	2D-descr	GA	PLS	26	0.96	0.07	0.08	0.92	0.09	0.10	0.91	0.08	0.11
MeCN96	ASAN7, ASAP1, ASAP4, pmi1, vsurf CW3, vsurf ID8	3D-1conf	GA	ANN	29	0.92	0.08	0.11	0.91	0.08	0.11	0.91	0.10	0.12
MeCN41	ALOGP[2], IC2, SHBint2, SHBint7, SHBint8	2D-descr	GA	PLS	26	0.94	0.07	0.09	0.90	0.10	0.12	0.90	0.09	0.12
MeCN91	ASAN6, ASAP4, pmi1, vsurf CW3, vsurf ID8	3D-1conf	GA	ANN	26	0.92	0.08	0.10	0.89	0.10	0.12	0.88	0.10	0.13
MeCN137	ASAN6, ASAP5, pmi1, vsurf HB1, vsurf ID8	3D-weigh	GA	PLS	26	0.93	0.08	0.10	0.89	0.10	0.12	0.86	0.18	0.25
MeCN37	ALOGP[2], IC2, SHBint2, SHBint7	2D-descr	GA	PLS	26	0.92	0.09	0.10	0.86	0.12	0.14	0.86	0.10	0.14
MeCN133	ASAP5, pmi1, vsurf HB1, vsurf ID8	3D-weigh	GA	PLS	26	0.92	0.09	0.10	0.90	0.10	0.12	0.85	0.18	0.26
MeCN95	ASAN7, ASAP1, ASAP4, pmi1, vsurf CW3, vsurf ID8	3D-1conf	GA	ANN	26	0.93	0.08	0.10	0.88	0.11	0.13	0.84	0.11	0.15
MeCN87	ASAP4, pmi1, vsurf CW3, vsurf ID8	3D-1conf	GA	ANN	26	0.92	0.09	0.11	0.86	0.12	0.14	0.83	0.11	0.16
MeCN143	glob, vsurf CW1, vsurf G, vsurf ID7, vsurf ID8, vsurf Wp6	3D-weigh	GA	ANN	26	0.95	0.08	0.10	0.90	0.10	0.12	0.81	0.14	0.17
Consensus3	–	–	–	–	29	0.94	0.09	0.12	–	–	–	0.96	0.08	0.09
Consensus4	–	–	–	–	26	0.97	0.06	0.07	–	–	–	0.87	0.10	0.13
Consensus5	–	–	–	–	26	0.97	0.05	0.06	–	–	–	0.94	0.07	0.10

[a] R^2, coefficient of determination (set = 26); Q^2, coefficient of determination for cross validation 10-folds; P^2, coefficient of determination for external validation (set = 13); MAE, mean absolut error; and RMSE, root mean square error.

CONCLUSIONS

In this work, the combination of six parameters, including different solvent systems, descriptor sets, test and training sets, algorithms for variable selection, and modeling methods allowed us to propose good QSRR models for SL in reversed-phase liquid chromatography. These combined parameters led to the identification of relevant descriptors before obtaining the models, especially using GA, which was definitely better than other algorithms used in this study. Both modeling methods (ANN and PLS) were able to generate good QSRR models, despite the high complexity of the SL structures and their uneven retention factors. Models from the MeCN−H_2O solvent system outperformed those from MeOH−H_2O, therefore suggesting the use of the former in such studies. One of the novelties of the method described herein is its applicability to real life situations like dereplication purposes of natural products in biological matrices using plant metabolomics-based techniques. Semi-automated methods for the prediction of log k are capable of producing good models, but they still need human supervision for the selection of templates that best reflect reality. In this study, the best models with respect to overall statistical quality were obtained with 2D descriptors, but these per se were not able to differentiate between stereoisomers. Even though much more laborious and time-consuming to calculate, the Boltzmann-weighted 3D descriptors (3D-weight) were the only ones that provided distinction among isomers. Therefore, we would recommend the construction of consensual models that are capable of joining the overall statistical performance of 2D descriptors and the ability to separate isomers of 3D-weight descriptors for further QSRR studies with SLs. To the best of our

DOI: 10.1021/ci500581q
J. Chem. Inf. Model. 2015, 55, 26−38

knowledge, this is the first time that consensus modeling, Botzmann weight, and Q_frASAs descriptors are used in QSRR studies.

■ ASSOCIATED CONTENT

Supporting Information

The 2D chemical structures of the 39 SL (with stereochemistry), a full list of all descriptors used in this study, the statistical values of all models and the predicted values of the best models. This material is available free of charge via the Internet at http://pubs.acs.org.

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: febcosta@fcfrp.usp.br.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

The authors thank Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES, http://capes.gov.br/), Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq, http://www.cnpq.br/, and São Paulo Research Foundation (FAPESP, www.fapesp.br/en) for financial support (2008/05754-5, 2010/51454-3, 2011/17860-7, and 2013/14239-5). This work is an activity within the Research Network Natural Products against Neglected Diseases, ResNet NPND (http://www.Resnetnpnd.org).

■ REFERENCES

(1) Theodoridis, G. A.; Gika, H. G.; Want, E. J.; Wilson, I. D. Liquid chromatography-mass spectrometry based global metabolite profiling: A review. *Anal. Chim. Acta* **2012**, *711*, 7−16.

(2) Kind, T.; Wohlgemuth, G.; Lee, D. Y.; Lu, Y.; Palazoglu, M.; Shahbaz, S.; Fiehn, O. FiehnLib: Mass spectral and retention index libraries for metabolomics based on quadrupole and time-of-flight gas chromatography/mass spectrometry. *Anal. Chem. Anal. Chem.* **2009**, *81*, 10038−48.

(3) Yang, J. Y.; et al. Molecular networking as a dereplication strategy. *J. Nat. Prod.* **2013**, *76*, 1686−99.

(4) Michel, T.; Halabalaki, M.; Skaltsounis, A. L. New concepts, experimental approaches, and dereplication strategies for the discovery of novel phytoestrogens from natural sources. *Planta Med.* **2013**, *79*, 514−32.

(5) Pilon, A. C.; Carneiro, R. L.; Carnevale Neto, F.; Bolzani, V. S.; Castro-Gamboa, I. Interval multivariate curve resolution in the dereplication of HPLC-DAD data from *Jatropha gossypifolia*. *Phytochem. Anal.* **2013**, *24*, 401−6.

(6) Halabalaki, M.; Vougogiannopoulou, K.; Mikros, E.; Skaltsounis, A. L. Recent advances and new strategies in the NMR-based identification of natural products. *Curr. Opin. Biotechnol.* **2014**, *25*, 1−7.

(7) Nakabayashi, R.; Saito, K. Metabolomics for unknown plant metabolites. *Anal. Bioanal. Chem.* **2013**, *405*, 5005−11.

(8) Nikitas, P.; Pappa-Louisi, A.; Tsoumachides, S.; Jouyban, A. A principal component analysis approach for developing retention models in liquid chromatography. *J. Chromatogr. A* **2012**, *1251*, 134−40.

(9) Noorizadeh, H.; Noorizadeh, M. QSRR-based estimation of the retention time of opiate and sedative drugs by comprehensive two-dimensional gas chromatography. *Med. Chem. Res.* **2011**, *21*, 1997−2005.

(10) Kumari, S.; Stevens, D.; Kind, T.; Denkert, C.; Fiehn, O. Applying *in-silico* retention index and mass spectra matching for identification of unknown metabolites in accurate mass GC-TOF mass spectrometry. *Anal. Chem.* **2011**, *83*, 5895−902.

(11) Abraham, M. H. Scales of solute hydrogen-bonding: Their construction and application to physicochemical and biochemical processes. *Chem. Soc. Rev.* **1993**, *22*, 73−83.

(12) Aschi, M.; D'Archivio, A. A.; Maggi, M. A.; Mazzeo, P.; Ruggieri, F. Quantitative structure-retention relationships of pesticides in reversed-phase high-performance liquid chromatography. *Anal. Chim. Acta* **2007**, *582*, 235−242.

(13) Gasteiger, J.; Funatsu, K. Chemoinformatics − An important scientific discipline. *J. Comput. Chem., Jpn.* **2006**, *5*, 53−58.

(14) Akbar, J.; Iqbal, S.; Batool, F.; Karim, A.; Chan, K. W. Predicting retention times of naturally occurring phenolic compounds in reversed-phase liquid chromatography: a quantitative structure-retention relationship (QSRR) approach. *Int. J. Mol. Sci.* **2012**, *13*, 15387−400.

(15) Petric, M.; Crisan, L.; Crisan, M. Synthesis and QSRR study for a series of phosphoramidic acid derivatives. *Heteroat. Chem.* **2013**, *24*, 138−145.

(16) Khodadoust, S. A QSRR study of liquid chromatography retention time of pesticides using linear and nonlinear chemometric models. *J. Chromatogr. Sep. Technol.* **2012**, *03*, 149.

(17) Cirera-Domènech, E.; Estrada-Tejedor, R.; Broto-Puig, F.; Teixidó, J.; Gassiot-Matas, M.; Comellas, L.; Lliberia, J. L.; Méndez, A.; Paz-Estivill, S.; Delgado-Ortiz, M. R. Quantitative structure-retention relationships applied to liquid chromatography gradient elution method for the determination of carbonyl-2,4-dinitrophenylhydrazone compounds. *J. Chromatogr. A* **2013**, *1276*, 65−77.

(18) Yan, J.; Cao, D. S.; Guo, F. Q.; Zhang, L. X.; He, M.; Huang, J. H.; Xu, Q. S.; Liang, Y. Z. Comparison of quantitative structure-retention relationship models on four stationary phases with different polarity for a diverse set of flavor compounds. *J. Chromatogr. A* **2012**, *1223*, 118−125.

(19) Ghavami, R.; Sepehri, B. Investigation of retention behavior of polychlorinated biphenyl congeners on 18 different HRGC columns using molecular surface average local ionization energy descriptors. *J. Chromatogr. A* **2012**, *1233*, 116−25.

(20) Noorizadeh, H.; Farmany, A. Quantitative structure-retention relationship for retention behavior of organic pollutants in textile wastewaters and landfill leachate in LC-APCI-MS. *Environ. Sci. Pollut. Res.* **2012**, *19*, 1252−9.

(21) Noorizadeh, H. Investigation of retention behaviors of essential oils by using QSRR. *J. Chin. Chem. Soc.* **2010**, *57*, 982−991.

(22) Wube, A. A.; Bucar, F.; Gibbons, S.; Asres, K. Sesquiterpenes from *Warburgia ugandensis* and their antimycobacterial activity. *Phytochemistry* **2005**, *66*, 2309−15.

(23) Sultana, R.; Hossain, R.; Adhikari, A.; Ali, Z.; Yousuf, S.; Choudhary, M. I.; Ali, M. Y.; Zaman, M. S. Drimane-type sesquiterpenes from *Polygonum hydropiper*. *Planta Med.* **2011**, *77*, 1848−51.

(24) Shen, Y.; van Beek, T. a.; Claassen, F. W.; Zuilhof, H.; Chen, B.; Nielen, M. W. F. Rapid control of chinese star anise fruits and teas for neurotoxic anisatin by direct analysis in real time high resolution mass spectrometry. *J. Chromatogr. A* **2012**, *1259*, 179−86.

(25) Bedoya, L.; Abad, M.; Bermejo, P. The role of parthenolide in intracellular signalling processes: review of current knowledge. *Curr. Signal Transduction Ther.* **2008**, *3*, 82−87.

(26) Schmidt, T. J. Toxic activities of sesquiterpene lactones: Structural and biochemical aspects. *Curr. Org. Chem.* **1999**, *3*, 577−608.

(27) Maas, M.; Hensel, A.; Da Costa, F. B.; Brun, R.; Kaiser, M.; Schmidt, T. J. An unusual dimeric guaianolide with antiprotozoal activity and further sesquiterpene lactones from *Eupatorium perfoliatum*. *Phytochemistry* **2011**, *72*, 635−44.

(28) Schomburg, C.; Schuehly, W.; Da Costa, F. B.; Klempnauer, K. H.; Schmidt, T. J. Natural sesquiterpene lactones as inhibitors of Myb-dependent gene expression: Structure−activity relationships. *Eur. J. Med. Chem.* **2013**, *63*, 313−20.

(29) Oliveira, R. B.; Chagas-Paula, D. A.; Secatto, A.; Gasparoto, T. H.; Faccioli, L. H.; Campanelli, A. P.; Da Costa, F. B. Topical anti-inflammatory activity of yacon leaf extracts. *Rev. Bras. Farmacogn.* **2013**, *23*, 497−505.

(30) Siedle, B.; Garca-Piñeres, A. J.; Murillo, R.; Schulte-Mönting, J.; Castro, V.; Rüngeler, P.; Klaas, C. A.; Da Costa, F. B.; Kisiel, W.;

DOI: 10.1021/ci500581q
J. Chem. Inf. Model. 2015, 55, 26−38

Merfort, I. Quantitative structure−activity relationship of sesquiterpene lactones as inhibitors of the transcription factor NF-kappaB. *J. Med. Chem.* **2004**, *47*, 6042−54.
(31) Ghantous, A.; Sinjab, A.; Herceg, Z.; Darwiche, N. Parthenolide: From plant shoots to cancer roots. *Drug Discovery Today* **2013**, *18*, 894−905.
(32) Schmidt, T. J.; et al. The potential of secondary metabolites from plants as drugs or leads against protozoan neglected diseases − Part I. *Curr. Med. Chem.* **2012**, *19*, 2128−75.
(33) Ghantous, A.; Gali-Muhtasib, H.; Vuorela, H.; Saliba, N. A.; Darwiche, N. What made sesquiterpene lactones reach cancer clinical trials? *Drug Discovery Today* **2010**, *15*, 668−78.
(34) Chadwick, M.; Trewin, H.; Gawthrop, F.; Wagstaff, C. Sesquiterpenoids lactones: Benefits to plants and people. *Int. J. Mol. Sci.* **2013**, *14*, 12780−805.
(35) Seaman, F. C. Sesquiterpene lactones as taxonomic characters in the Asteraceae. *Bot. Rev.* **1982**, *48*, 121−592.
(36) Emerenciano, V. P.; Ferreira, M.; Branco, M.; Dubois, J. The application of Bayes' theorem in natural products as a guide for skeletons identification. *Chemom. Intell. Lab. Syst.* **1998**, *40*, 83−92.
(37) Da Costa, F. B.; Terfloth, L.; Gasteiger, J. Sesquiterpene lactone-based classification of three Asteraceae tribes: A study based on self-organizing neural networks applied to chemosystematics. *Phytochemistry* **2005**, *66*, 345−53.
(38) Chagas-Paula, D. A.; Oliveira, R. B.; Rocha, B. A.; Da Costa, F. B. Ethnobotany, chemistry, and biological activities of the genus *Tithonia* (Asteraceae). *Chem. Biodiversity* **2012**, *9*, 210−35.
(39) Hristozov, D.; Da Costa, F. B.; Gasteiger, J. Sesquiterpene lactones-based classification of the family Asteraceae using neural networks and k-nearest neighbors. *J. Chem. Inf. Model.* **2007**, *47*, 9−19.
(40) *IUPAC Compendium of Chemical Terminology (The Gold Book)*, 2nd ed.; Nič, M., Jirát, J., Košata, B., Jenkins, A., McNaught, A., Eds.; IUPAC: Research Triagle Park, NC, 2009.
(41) Da Costa, F. B.; Schorr, K.; Arakawa, N. S.; Schilling, E. E.; Spring, O. Infraspecific variation in the chemistry of glandular trichomes of two brazilian *Viguiera* species (Heliantheae; Asteraceae). *J. Braz. Chem. Soc.* **2001**, *12*, 403−407.
(42) Schorr, K.; Garca-Piñeres, A. J.; Siedle, B.; Merfort, I.; Da Costa, F. B. Guaianolides from *Viguiera gardneri* inhibit the transcription factor NF-kappaB. *Phytochemistry* **2002**, *60*, 733−40.
(43) Spring, O.; Zipper, R.; Conrad, J.; Vogler, B.; Klaiber, I.; Da Costa, F. B. Sesquiterpene lactones from glandular trichomes of *Viguiera radula* (Heliantheae; Asteraceae). *Phytochemistry* **2003**, *62*, 1185−1189.
(44) Stefani, R.; Eberlin, M. N.; Tomazela, D. M.; Da Costa, F. B. Eudesmanolides from *Dimerostemma vestitum*. *J. Nat. Prod.* **2003**, *66*, 401−3.
(45) Schorr, K.; Merfort, I.; Da Costa, F. B. A novel dimeric melampolide and further terpenoids from *Smallanthus sonchifolius* (Asteraceae) and the inhibition of the transcription factor NF-kB. *Nat. Prod. Commun.* **2007**, *2*, 367−374.
(46) Ambrósio, S. R.; Oki, Y.; Heleno, V. C. G.; Chaves, J. S.; Nascimento, P. G. B. D.; Lichston, J. E.; Constantino, M. G.; Varanda, E. M.; Da Costa, F. B. Constituents of glandular trichomes of *Tithonia diversifolia*: relationships to herbivory and antifeedant activity. *Phytochemistry* **2008**, *69*, 2052−2060.
(47) Chaves, J. S.; Da Costa, F. B. A proposal for the quality control of *Tanacetum parthenium* (feverfew) and its hydroalcoholic extract. *Rev. Bras. Farmacogn.* **2008**, *18*, 360−366.
(48) Marvin. ChemAxon, Ltd., 2012. http://www.chemaxon.com/ (accessed December 2014).
(49) Hall, M.; Frank, E.; Holmes, G.; Bernhard, P.; Reutemann, P.; Witten, I. H.; Pfahringer, B. The WEKA data mining software: An update. *ACM SIGKDD* **2009**, *11*, 10−18.
(50) R Core Team. R: A Language and Environment for Statistical Computing, 2013. http://www.r-project.org/ (accessed December 2014).
(51) MOE: Molecular Operating Environment. Chemical Computing Group, Inc., 2010. http://www.chemcomp.com/index.htm (accessed December 2014).
(52) CORINA (COoRdINAtes). Molecular Networks GmbH, 2011. http://www.molecular-networks.com (accessed December 2014).
(53) Adriana.Code − Calculation of Molecular Descriptors. Molecular Networks GmbH, 2011. http://www.molecular-networks.com/ (accessed December 2014).
(54) Dragon. Talete SRL, 2007. http://www.talete.mi.it/index.htm (accessed December 2014).
(55) Yap, C. W. PaDEL-descriptor: An open source software to calculate molecular descriptors and fingerprints. *J. Comput. Chem.* **2011**, *32*, 1466−74.
(56) Chen, I. J.; Foloppe, N. Conformational sampling of druglike molecules with MOE and catalyst: Implications for pharmacophore modeling and virtual screening. *J. Chem. Inf. Model.* **2008**, *48*, 1773−1791.
(57) Widom, B. *Statistical Mechanics: A Concise Introduction for Chemists*, 1st ed.; Cambridge University Press: Cambridge, 2002; Chapter 1, p 182.
(58) Kuhn, M. Building predictive models in R using the Caret package. *J. Stat. Softw.* **2008**, *28*, 1−26.
(59) Hall, M. A.; Smith, L. A. Feature Subset Selection: A Correlation Based Filter Approach. In *Progress in Connectionist-Based Information Systems*, Volumes 1 and 2; Kasabov, N., Kozma, R., Ko, K., O'Shea, R., Coghill, G., Gedeon, T., Eds.; Springer: New York, 1998; pp 855−858.
(60) Helland, I. S. On the structure of partial least squares regression. *Commun. Stat.: Simul. C* **1988**, *17*, 581−607.
(61) Witten, I. H.; Frank, E.; Hall, M. A. *Vasa*, 3rd ed.; Elsevier: Amsterdam, 2008; p 629.
(62) Bosch, E.; Bou, P.; Allemann, H.; Rosés, M. Retention of ionizable compounds on HPLC. pH scale in methanol-water and the pK and pH values of buffers. *Anal. Chem.* **1996**, *68*, 3651−3657.
(63) Snyder, L.; Kirkland, J.; Dolan, J. *Introduction to Modern Liquid Chromatography*, 3rd ed.; John Wiley & Sons, Inc.: Hoboken, NJ, 2009.
(64) Basak, S. C.; Harriss, D. K.; Magnuson, V. R. Comparative study of lipophilicityversus topological molecular descriptors in biological correlations. *J. Pharm. Sci.* **1984**, *73*, 429−437.
(65) Roy, K.; Ghosh, G. Introduction of Extended Topochemical Atom (ETA) indices in the Valence ElectronMobile (VEM) environment as tools for QSAR/QSPR studies. *Internet Electron. J. Mol. Des.* **2003**, *2*, 599−620.
(66) Roy, K.; Kabir, H. QSPR with extended topochemical atom (ETA) indices: Modeling of critical micelle concentration of non-ionic surfactants. *Chem. Eng. Sci.* **2012**, *73*, 86−98.
(67) Todeschini, R.; Consonni, V. In *Handbook of Molecular Descriptors, Methods and Principles in Medicinal Chemistry*, Vol. *11*; Mannhold, R., Kubinyi, H., Timmerman, H., Eds.; Wiley-VCH Verlag GmbH: Weinheim, 2000; p 667.
(68) Moorthy, N. S. H. N.; Ramos, M. J.; Fernandes, P. A. Analysis of van der Waals surface area properties for human ether-a-go-go-related gene blocking activity: computational study on structurally diverse compounds. *SAR QSAR Environ. Res.* **2012**, *23*, 521−36.
(69) Cruciani, G.; Crivori, P.; Carrupt, P. A.; Testa, B. Molecular fields in quantitative structure-permeation relationships: the VolSurf approach. *J. Mol. Struct.: THEOCHEM* **2000**, *503*, 17−30.
(70) Schmidt, T. J.; Heilmann, J. Quantitative structure-cytotoxicity relationships of sesquiterpene lactones derived from partial charge (Q)-based fractional Accessible Surface Area Descriptors (Q_frASAs). *Quant. Struct.-Act. Relat.* **2002**, *21*, 276−287.
(71) Zhivkova, Z.; Doytchinova, I. Quantitative structure-plasma protein binding relationships of acidic drugs. *J. Pharm. Sci.* **2012**, *101*, 4627−41.
(72) Gramatica, P.; Giani, E.; Papa, E. Statistical external validation and consensus modeling: A QSPR case study for Koc prediction. *J. Mol. Graphics Modell.* **2007**, *25*, 755−66.
(73) Zhu, H.; Tropsha, A.; Fourches, D.; Varnek, A.; Papa, E.; Gramatica, P.; Oberg, T.; Dao, P.; Cherkasov, A.; Tetko, I. V. Combinatorial QSAR modeling of chemical toxicants tested against *Tetrahymena pyriformis*. *J. Chem. Inf. Model.* **2008**, *48*, 766−84.

DOI: 10.1021/ci500581q
J. Chem. Inf. Model. 2015, 55, 26−38

(74) Zhu, H.; Martin, T. M.; Ye, L.; Sedykh, A.; Young, D. M.; Tropsha, A. Quantitative structure−activity relationship modeling of rat acute toxicity by oral exposure. *Chem. Res. Toxicol.* **2009**, *22*, 1913−21.
(75) Héberger, K.; Skrbić, B. Ranking and similarity for quantitative structure-retention relationship models in predicting Lee retention indices of polycyclic aromatic hydrocarbons. *Anal. Chim. Acta* **2012**, *716*, 92−100.

DOI: 10.1021/ci500581q
J. Chem. Inf. Model. 2015, 55, 26−38