analytical chemistry

Article

pubs.acs.org/ac

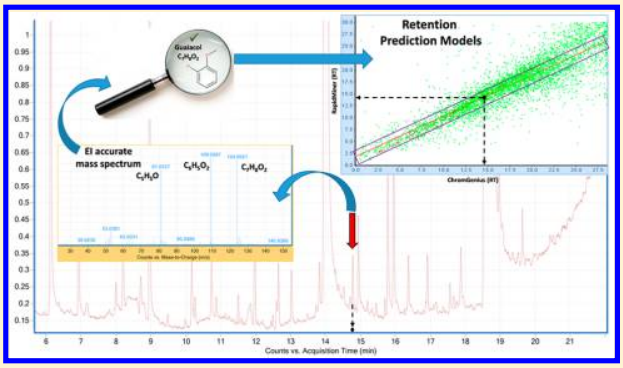
Prediction Models of Retention Indices for Increased Confidence in Structural Elucidation during Complex Matrix Analysis: Application to Gas Chromatography Coupled with High-Resolution Mass Spectrometry

Eric Dossin, Elyette Martin, Pierrick Diana, Antonio Castellon, Aurelien Monge, Pavel Pospisil, Mark Bentley, and Philippe A. Guy*

Philip Morris International R&D, Philip Morris Products S.A., Quai Jeanrenaud 5, CH-2000 Neuchatel, Switzerland

Supporting Information

ABSTRACT: Monitoring of volatile and semivolatile compounds was performed using gas chromatography (GC) coupled to high-resolution electron ionization mass spectrometry, using both headspace and liquid injection modes. A total of 560 reference compounds, including 8 odd *n*-alkanes, were analyzed and experimental linear retention indices (LRI) were determined. These reference compounds were randomly split into training (n = 401) and test (n = 151) sets. LRI for all 552 reference compounds were also calculated based upon computational Quantitative Structure−Property Relationship (QSPR) models, using two independent approaches RapidMiner (coupled to Dragon) and ACD/ChromGenius software. Correlation coefficients for experimental versus predicted LRI values calculated for both training and test set compounds were calculated at 0.966 and 0.949 for RapidMiner and at 0.977 and 0.976 for ACD/ChromGenius, respectively. In addition, the cross-validation correlation was calculated at 0.96 from RapidMiner and the residual standard error value obtained from ACD/ChromGenius was 53.635. These models were then used to predict LRI values for several thousand compounds reported present in tobacco and tobacco-related fractions, plus a range of specific flavor compounds. It was demonstrated that using the mean of the LRI values predicted by RapidMiner and ACD/ChromGenius, in combination with accurate mass data, could enhance the confidence level for compound identification from the analysis of complex matrixes, particularly when the two predicted LRI values for a compound were in close agreement. Application of this LRI modeling approach to matrixes with unknown composition has already enabled the confirmation of 23 postulated compounds, demonstrating its ability to facilitate compound identification in an analytical workflow. The goal is to reduce the list of putative candidates to a reasonable relevant number that can be obtained and measured for confirmation.



Unambiguous chemical characterization still remains a major hurdle for analytical chemists when dealing with nontargeted analyses, despite significant improvements in chromatographic separation techniques and mass spectrometric instrumentation over the past decade. Hence, the final step for compound identification still requires the purchase of putative chemicals and matching of spectra and retention times under identical analytical conditions. However, reference standards are not always commercially available and several standards may need to be analyzed before finding an exact match (e.g., in case of isomeric compounds), which can represent a very costly process. Therefore, a suitable balance between the cost for purchasing chemical standards and the speed for certainty in compound identification is of great importance. Gas chromatography coupled with mass spectrometric detection (GC/MS) is currently the most widely used technique because of its high reproducibility of results across laboratories for both qualitative and quantitative aspects and the availability of extensive and reliable mass spectral libraries. Currently, the Wiley Registry (10th edition) together with National Institute of Standards and Technology version 14 (NIST 14) contains over 960 000 electron ionization (EI) mass spectra representing approximately 750 000 unique compounds.[1] However, such numbers are small in comparison with the PubChem compound database, which contains nearly 200 million compounds,[2,3] or the ChemSpider repository, which currently contains 37 million chemical structures.[4]

Obviously, there is a need to develop additional tools in order to strengthen the confidence level for compound identification, particularly when mass spectral information for chemicals are not available in any existing MS libraries. To overcome this gap, several software packages have been

Received: March 4, 2016
Accepted: July 12, 2016
Published: July 12, 2016

DOI: 10.1021/acs.analchem.6b00868
Anal. Chem. 2016, 88, 7539−7547

developed to predict *in silico* EI fragmentation using either an MS rule-based approach such as Mass Frontier,[5] high-throughput automation of mass frontier (HAMMER),[6] MS interpreter,[7] and ACD/MS Fragmenter[8] or using combinatorial proposals such as Fragment Identificator,[9,10] Fragment Formula Calculator,[11,12] Molecular Structure Correlator,[13] and MetFrag.[14−17] Most of them use chemical structures as input and generate spectra which are then compared against measured ones before ranking the compounds proposal. However, Schymanski et al. highlighted a need for caution when taking the highest matching score value for compound ranking.[18] Indeed, the match value versus assignment quality index revealed some issues regarding the relative abundance of fragment ions proposed by *in silico* prediction approaches and highlighted a limited ability for the consideration of structural rearrangements beyond those related to simple hydrogen movements. Small changes in structure can also lead to the proposal of a significantly different fragmentation mechanism using an MS rule-based fragmentation approach.[19] Although *in silico* fragmentation software brings important added value to compound identification, EI mass spectra of some isomeric compounds will result in the same or very similar fragmentation patterns, where only the intensity of some fragment ions may differ. Even accurate mass measurements are not helpful for distinguishing between isomeric forms in such cases. Another example of compound identification difficulties arising in GC/MS analysis of complex samples is the fact that the same class of compounds, such as terpenes, have identical mass spectra. This is a consequence of similarities either in the molecules themselves or in the fragmentation patterns and rearrangements occurring upon the electron ionization process. Hence, there is clearly a need for implementing additional tools to improve the confidence level for compound identification. Kim et al. described a model-based approach to control the false identification rate of compounds using the distribution of the difference between the first and second highest spectral similarity scores.[20]

In addition to the calculated mass spectral match, Lee retention index values, linear retention index (LRI), boiling point correlation, NIST Kovats retention index values, octanol−water partitioning, and steric energy calculations have been taken into consideration for compound identification.[16,21] Retention indexing approaches have been described using liquid chromatography coupled to MS detector as an additional aid in compound identification.[22−26] Kumari et al. have also reported a similar approach in GC/MS for silylated compounds.[27] Several authors have already reported the combined use of both mass spectral matching score together with retention index (RI) value in order to enhance the identification accuracy.[28,29] Although chromatographic indexing data has recently become popular, with NIST 14 containing Kovats and LRI data for 385 872 and 82 337 compounds, respectively, these measured RI data are not exhaustive in comparison to the millions of known chemical structures and a capability to predict their RI values. NIST values report RI according to three GC column types having either standard polar, nonpolar, or semistandard nonpolar stationary phases. Several authors have attempted to predict chromatographic retention data using mainly specific classes of compounds or contextual databases.[29−31] For example, Garkani-Nejad et al. have studied the applicability of quantitative structure−property relationships (QSPR) for the prediction of gas chromatographic RI using a set of 846 toxicologically relevant organic compounds.[30] In addition, Stein et al. have assessed a simple linear group incremental model for the estimation of Kovats retention indices using 84 different chemical functionalities but concluded that, although it could reliably eliminate false identifications from automated library search systems, the approach was very approximate.[31]

The present work focuses on building a robust LRI system using QSPR-based modeling with the evaluation of two commercially available software packages.[32−35] Experimental LRI values for 552 volatile and semivolatile reference standards, analyzed by gas chromatography coupled to high-resolution mass spectrometry (GC-HR-MS), were randomly split into training and test sets. Once the two modeling approaches were optimized from the training set and assessed on the test set, LRI values were predicted for all tobacco, tobacco-related smoke constituents[36] as well as characteristic flavor compounds,[37] which represents a broad chemical space above 11 000 compounds. In order to enhance the confidence level of compound identification, predicted LRIs were matched with the experimental data together with accurate mass measurements. Usefulness of LRI models to be able to predict LRI for any new compound candidate and a repository database for such information enhances the process for compound identification. The next section discusses the strength of using two modeling approaches and the overall increase in confidence for chemical identification from the analysis of complex samples before confirmation with reference standards.

EXPERIMENTAL SECTION

Chemicals. Phosphate buffered saline (PBS) solution, *N*,*N*-dimethylformamide (DMF), dichloromethane (DCM), water, methanol (MeOH), as well as odd *n*-alkane chemical retention markers (pentane (C5), heptane (C7), nonane (C9), undecane (C11), tridecane (C13), pentadecane (C15), heptadecane (C17), and nonadecane (C19)) were purchased from Sigma-Aldrich (Buchs, CH). All reference standards have been described in the Supporting Information. The compounds were weighed and solubilized in appropriate solvents to provide stock solutions of concentration around 50 μg/mL per individual compound. After subsequent dilution, these compounds were analyzed by GC-HR-MS, either as a mixture or unique compound in solution, to ensure accurate retention index recording and to register reliable corresponding accurate mass spectra in an in-house Agilent-based Personal Compound Database Library (PCDL; Agilent Technologies, Santa Clara, CA).

Gas Chromatography Conditions. Gas chromatographic separation was realized using an Agilent 7890A instrument equipped with a J&W DB-624 ultra inert (UI) column (30 m × 0.25 mm internal diameter, 1.4 μm film thickness; Agilent, Basel, CH). A conventional static headspace (HS) injection method was used to monitor volatile compounds with a low boiling point. An aliquot (1 mL) of reference standard(s) solution diluted with water (or alternatively with DMF or PBS) was introduced in a 20 mL HS glass vial, incubated for 10 min at 100 °C, and 250 μL of the HS portion was injected via a multimode inlet set at 220 °C, with a split ratio of 5:1. For the liquid injection procedure, an aliquot (1 mL) of diluted standard(s) solubilized in dichloromethane was transferred into a glass autosampler vial and 1 μL was injected via a multimode inlet set at 220 °C in pulsed splitless mode. For both headspace and liquid injections, the column oven was first maintained at 35 °C for 2 min, before being ramped up to 250 °C at a

DOI: 10.1021/acs.analchem.6b00868
Anal. Chem. 2016, 88, 7539−7547

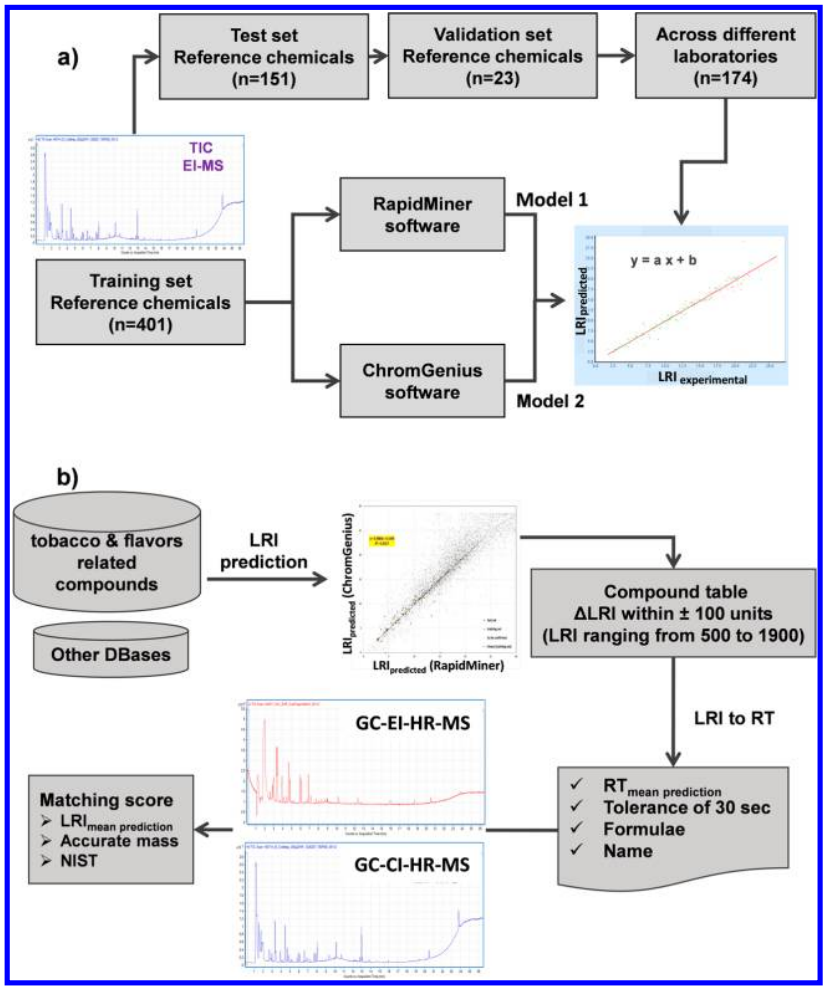


Figure 1. Workflow used to (a) build and validate the retention index prediction models and (b) its application as an additional tool for increasing confidence level in compound identification from complex matrix samples.

constant rate of 10 °C/min. Finally, the temperature was maintained at 250 °C for an additional 3 min (total run time of 26.5 min). The transfer line was held at 260 °C, and the nitrogen flow rate was kept constant at 1.8 mL/min during the whole analysis.

Mass Spectrometry Conditions. Detection was carried out using a 7200A Q-TOF (quadrupole with time-of-flight) accurate mass spectrometer system (Agilent Technologies, Santa Clara, CA). A solvent delay of 4.8 min was used for the liquid injection experiments in order to extend the filament lifetime. Temperature of the ion source and emission current were set at 230 °C and 35 μA, respectively. Mass spectrometric (MS) data were acquired in full scan mode by scanning m/z values ranging from 22 to 500 using positive electron (+EI), negative chemical (NCI), and positive chemical (PCI) ionization modes. Ammonia and methane were used as reactant gases for NCI and PCI measurements, respectively. Data processing was performed using MassHunter Qualitative software (version B7.00.0, Agilent Technologies, Santa Clara, CA). A background subtracted EI accurate mass spectrum for each standard was exported in "compound exchange file" (.cef) format and used to build an in-house PCDL accurate mass library database, including retention time, molecular structure, and experimental LRI data. This database library could then be used as an additional search parameter.

Retention Index Prediction Modeling. All reference standard structures were drawn using Accelrys Draw 4.1 and each chemical structure was standardized by neutralizing charges, generating canonical tautomers and adding hydrogens using Pipeline Pilot 9.1 (PP) software.[33,38] The compounds were randomly split into training (n = 401, 73%) and test (n = 151, 27%) sets and Dragon software (version 5.5 for Windows) was used to generate two-dimensional molecular descriptors.[34] A Pipeline Pilot protocol with genetic function approximation (GFA) was used with a linear model, a maximum equation length of 10 up to 25 (bin size of 5), population size of 100, and maximum generation of 5 000 (see the Supporting Information). The Pareto algorithm (NSGA-II) was used as a scoring method with adjusted R-square (Supporting Information). Three types of learning algorithms, namely, using k-nearest neighbors (k-NN), multilinear regression (MLR), and support vector regression (SVR), were evaluated within RapidMiner software (version 5)[35] to improve the prediction model. This software was also used to optimize the C parameter of the SVR algorithm and to evaluate the generated models by performing cross-validation with squared correlation (q^2). The best algorithm was chosen according to the correlation coefficients of the training and test sets together with the cross-validation score (q^2 value) according to the number of descriptors selected. In parallel, the same training and test sets were used to

DOI: 10.1021/acs.analchem.6b00868
Anal. Chem. 2016, 88, 7539−7547

Table 1. Results Summary Obtained from RapidMiner and ACD/ChromGenius Optimized Prediction Models

	RapidMiner		ChromGenius	
	training set	test set	training set	test set
no. of ref compounds	n = 401	n = 151	n = 397	n = 149
correlation coefficient	r^2 = 0.966	r^2 = 0.949	r^2 = 0.977	r^2 = 0.976
cross-validation	Q^2 = 0.96		NA	residual standard error: 53.635
accuracy within (85−115%)	n = 392	n = 145	n = 392	n = 147
accuracy outside (85−115%)	n = 9	n = 6	n = 5	n = 2
	Extended Application			
no. of ref compounds	n = 23		n = 23	
accuracy	91.0−109.8%		87.5−112.2%	

optimize the LRI prediction model using ACD/ChromGenius Batch software (version 2014, ACD/Laboratories, Toronto, CA).[32] In this case, the prediction is based upon calculated physicochemical parameters and structural similarity with known retention time contained within a knowledge base, which is created using the training set compounds. The calculated physicochemical parameters used are boiling point (BP), log P, polar surface area (PSA), molecular volume (MV), molecular weight (MW), molar refractivity (MR), number of hydrogen donor (ND) and number of hydrogen acceptor (NA). These parameters are used to create a prediction equation.

RESULTS AND DISCUSSION

Analysis of Chemical Markers and Reference Standard Molecules Using Headspace and Liquid Injection Modes. In total, 552 reference standards plus 8 odd n-alkanes (chemical markers) were purchased, solubilized, and analyzed by GC-HR-MS, generating unique background subtracted EI accurate mass spectra. The selection of these standards was based upon compounds already identified and/or reported to be present in tobacco or tobacco smoke,[36] plus additional relevant flavor compounds.[37] The chemical composition of tobacco and tobacco aerosol fractions is very complex, with over 7 000 constituents already reported.[36] These comprise highly hydrophilic to hydrophobic compounds as well as a broad range of heteroatom functionalities, the majority having a molecular weight below 1 000 Da (see Supporting Information).[36,39,40] The Agilent J&W DB-624 UI was selected as an appropriate GC column for the analysis of volatile and semivolatile organic compounds. Although the temperature gradient conditions could have been optimized to enable a better chromatographic separation of specific key compounds, the core application for the method is untargeted screening analysis. Therefore, there was a need to balance between an adequate chromatographic separation, while reducing the run time in order to minimize instrumental drift. Considering this, and in order to reduce the complexity for the computational chemistry tool to build prediction models for retention index, the column temperature was ramped using a linear gradient. Reference compounds having a BP below 90 °C (which corresponded to a LRI below 627) were injected in headspace mode, whereas those having a higher BP were injected in liquid mode. The LRI cutoff value of 627 corresponds to the solvent delay of 4.8 min applied for liquid injection mode, during which time the detector filament is switched off in order to increase its lifetime. Several reference compounds were analyzed using both injection modes. A set of odd n-alkane chemical markers (from pentane up to nonadecane) was used to bracket these reference standards for the calculation of LRIs, which were used to build the prediction models.

Building a LRI Prediction Model Using Reference Standards. Prior to analysis, the compounds were randomly split as training (n = 401) and test (n = 151) sets following standard default recommendations.[41] Figure 1a summarizes the workflows used to calculate and assess LRI prediction models. The required parameters for ACD/ChromGenius software and the RapidMiner approach were optimized using the same training set of compounds.

RapidMiner Approach. A total of 2 489 two-dimensional descriptors were calculated using Dragon Plus software (version 5.5) for the training set. The descriptors having only a minor contribution were removed from the combination of constant variables, near-constant variables, and variables with pair correlation values above 0.98. According to these criteria, a reduced list of 466 descriptors was kept for this evaluation and the best model was chosen from the algorithm showing the highest cross-validation (q^2) using leave-many-out methodology. In order to minimize the number of descriptors to the most relevant ones, the performance of the model was evaluated using either 10, 15, 20, 25, or all (n = 466) descriptors for each of the three algorithms tested (MLR, k-NN, and SVR). Hence, MLR algorithm using only 20 descriptors demonstrated the optimal prediction models (see optimization results in the Supporting Information). The model relevant descriptors belonged to seven major categories comprising constitutional descriptors (n = 4), topological descriptors (n = 4), connectivity indices (n = 2), functional group counts (n = 3), atom-centered fragments (n = 3), molecular properties (n = 3), and a single 2D-binary fingerprinting descriptor (Supporting Information). Once the prediction model was optimized, the correlation coefficients for the training and test sets were calculated to be 0.966 and 0.949, respectively, with a cross-validation q^2 value of 0.960 (Table 1).

ACD/ChromGenius. *trans*-Nicotine-1′-oxide, 1-dodecene-1,2-$^{13}C_2$, and 2,5-dimethyl-$^{13}C_2$-furan in the training set, and vinyl acetate-$^{13}C_2$ and benz-$^{13}C_6$-aldehyde in the test set were not recognized by ACD/ChromGenius. In the case for *trans*-nicotine-1′-oxide, this was due to the presence nitrogen-oxide quaternary bond. For the stable labeled internal standards used to provide semiquantitative data, the software did not recognize compounds containing a carbon 13 isotope. Experimental LRI values were uploaded into ACD/ChromGenius Batch 2014 for the remaining 398 compounds used for the training set (n = 149 remained in the test set). In this approach, LRI prediction was based upon physicochemical parameters calculated within the ACD/ChromGenius software, as well as structural similarities registered from the training set. Several equation and parameter options available within the software were

DOI: 10.1021/acs.analchem.6b00868
Anal. Chem. 2016, 88, 7539−7547

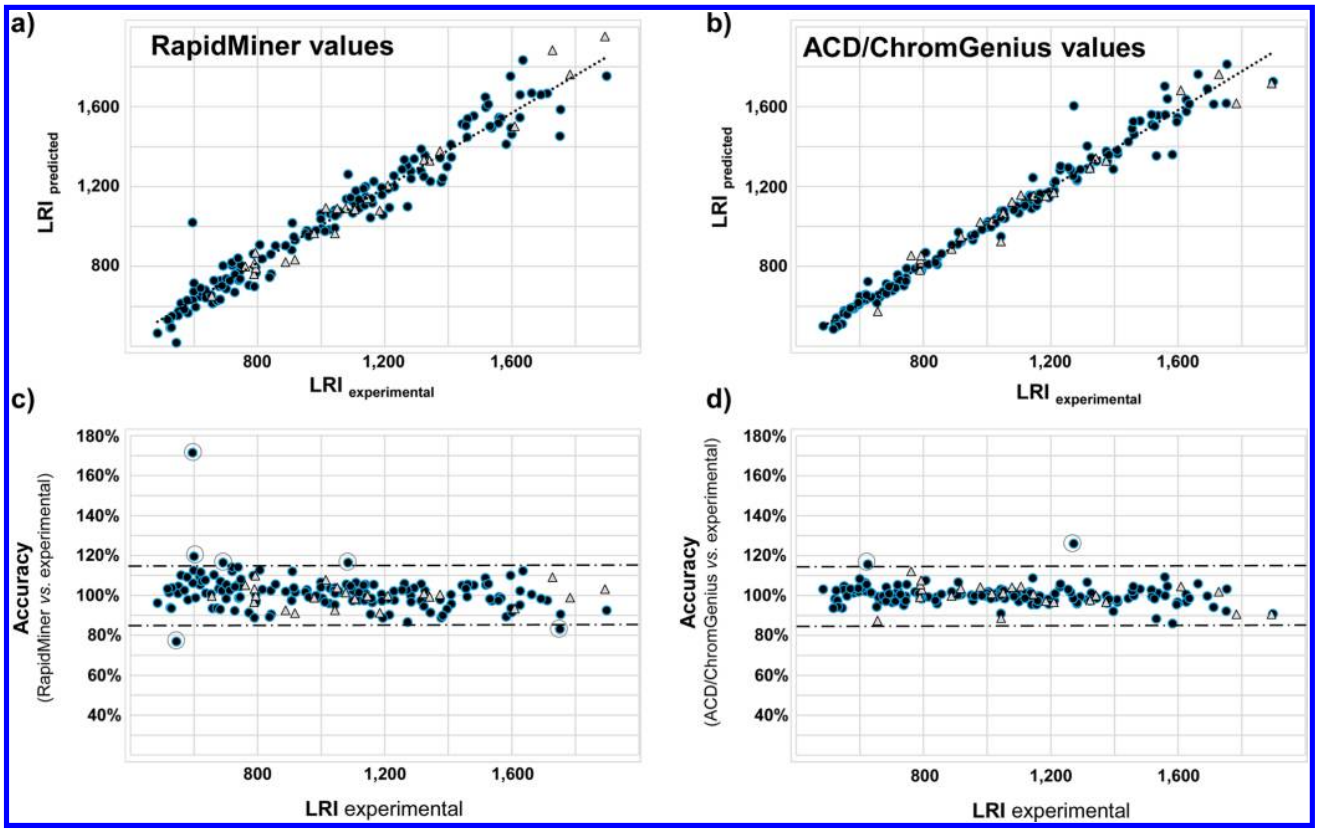


Figure 2. Correlation plots of calculated linear retention indices obtained from (a) RapidMiner and (b) ACD/ChromGenius software against experimentally derived values for the analysis of volatile and semivolatile reference standards (test set) using GC-HR-MS. Reference standards used for the test set have been assigned in a circle whereas those in a triangle refer to additionally added, confirmed compounds. Accuracy data of predicted LRI versus experimental data for (c) RapidMiner and (d) ACD/ChromGenius for the reference compounds used in the test set. A dashed line has been included to delimit compounds outside an arbitrary accuracy limit values of 85−115%.

assessed in order to optimize the modeling. The correlation coefficients of both training and test sets, as well as their corresponding residual standard error values were calculated from different number and/or percentage of compounds similarity (see optimization results in Supporting Information). The model with the best correlation and lower residual standard error value for the test set was selected for use, since ACD/ChromGenius software was not able to perform a cross-validation. The best results were obtained using a structural similarity search option using the Dice coefficient, with automatic software selection of the 15 most similar compounds.[42] These option settings provided a coefficient of correlation for the training and test sets of 0.9767 and 0.9762, respectively, with a residual standard error value of 53.635 (Table 1).

To our knowledge, according to the current literature, these are the most accurately predictive LRI models built so far, for the largest set of diverse molecules. The test set of compounds (n = 151 and n = 149 used for RapidMiner and ACD/ChromGenius software, respectively) was used to confirm the performance of the prediction models (Figure 2a,b). In order to further investigate the performance of the two models, predicted LRI values were compared with experimental values and expressed in terms of accuracy (percentage). Figure 2c,d show the accuracy data plotted against the experimental LRI values. This representation rapidly highlights outlying compounds and pinpoints weakly predicted classes of compound for each software. For the reference standards used in the test set, prediction accuracy values ranged between 76.8 and 171.4% and between 86.0 and 126.0%, for RapidMiner and ACD/ChromGenius, respectively. All compounds were classified according to their chemical properties into one or more of the 47 classes as defined by Perfetti and Rodgman,[36,39,40] according to the nature of the compounds (see Supporting Information). From the 151 compounds used in the test set, 6 generated accuracy values outside the 85−115% limit using RapidMiner prediction (values arbitrarily defined), namely, acetonitrile (76.8%), triethylcitrate (83.0%), 2-methyl-*p*-benzoquinone (116.3%), 2-methyltetrahydrofuran (116.3%), diisopropylether (119.5%), and 2-bromo-2-chloro-1,1,1-trifluoroethane (171.4%). These results, combined with those obtained for the training set (n = 401), revealed 9 additional compounds falling outside this expected range, namely, divinylacetylene (69.4%), 1H-imidazole-1,2-dimethyl (81.1%), 3-methylpyridazine (83.0%), furanone (83.3%), *N*,*N*-dimethylformamide (84.8%), 2-methylfuran (117.8%), isopropyl formate (119.5%), salicylaldehyde (120.4%), and diacetyl (123.4%). Some of these outliers can be explained as either having insufficient or no representative compounds present in the training set for the related compound class (n = 0 for quinone or mixed halogens; n = 1 for imidazole, pyrazidine, amide, or aldehyde-phenol as combined functions; n = 2 for alkene-alkyne or alcohol-ester as combined functions, respectively).

Using ACD/ChromGenius prediction, only 2 out of the 149 compounds used in the test set had accuracy values outside 85−115%, namely, 1-hexyne (115.7%) and 3-pyridinol

DOI: 10.1021/acs.analchem.6b00868
Anal. Chem. 2016, 88, 7539−7547

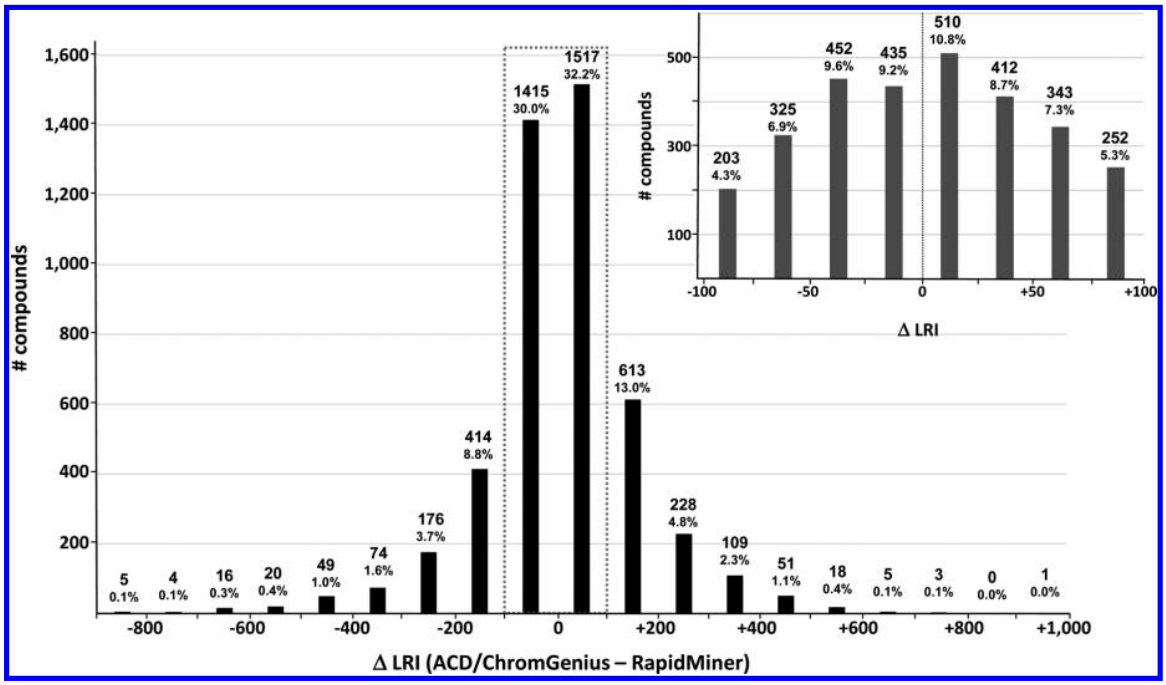


Figure 3. Compounds clustered according to the difference in LRIs predicted by ACD/ChromGenius and RapidMiner software. LRI values for all chemicals related to tobacco, tobacco aerosol, and flavors were put in bins of ΔLRI values by size of ±100 units. Inset in the right corner of the figure depicts the binning of compounds from ΔLRI values within the ±100 units (main graph) in bins by 25 units.

(126.0%), due to the absence of any compound representative of the alcohol/pyridine class. An additional 23 reference compounds were analyzed as a validation set. All accuracy values (both models) fitted well within expectations (depicted as triangles in Figure 2 and Supporting Information). In conclusion, 96.55% of all compounds tested using RapidMiner software (169 out of 174 reference standards) fitted well within a prediction accuracy of 85−115%. For ACD/ChromGenius, this percentage was slightly higher (98.84%) with 170 out of 172 reference standards falling within a prediction accuracy of 85−115% (Table 1).

Assessment of LRI Data Collected by External Laboratories. The main advantage of using LRI values instead of absolute retention times is the possibility to compare results across different instruments and laboratories. Indeed, these values are unaffected by different analytical conditions such as one continuous temperature gradient and column length; however, comparability is more reliable when similar column stationary phase materials are used (e.g., methyl phenyl cyanopropyl polysiloxane, in our case). To corroborate this hypothesis, experimentally determined LRI values were benchmarked against published data for the analysis of volatile organic compounds using a DB-624 GC column.[43−45] These data revealed a strong correlation (r^2 above 0.9958) between LRI values across laboratories (see the Supporting Information). From this data comparison, the correlation could be further improved by including even *n*-alkanes (only odd *n*-alkanes were used in our work). From these results, it appears extremely beneficial to use the retention indexing feature within NIST 14 mass library search option (that contains currently 385 872 RI values for 82 868 compounds) in order to reduce false positive compound identification.[46] Obviously, attention must be paid to published LRI values with respect to the similarity of GC column stationary phase material used.

LRI Prediction Models for Unknown Analysis in Accurate Mass Chemical Ionization and/or EI Acquisition Modes. All subtracted EI mass spectra generated from the analysis of reference standards were uploaded into an in-house PCDL accurate mass database library (n = 607 compounds). The identification of compounds from the analysis of complex matrix samples is realized using Unknown MassHunter software. The deconvoluted peaks are first compared to our accurate mass PCDL library for compound identification with the experimental LRI tolerance value set within 10 units. For the remaining unknowns, the EI mass spectra were compared against commercialized libraries. Although the ultimate compound identification step comprises matching the results of putative identification with reference compounds analyzed under identical analytical conditions, this may not be practically feasible due to either cost or commercial availability of reference compounds due to troublesome chemical synthesis or compound stability issues. Moreover, these practicalities are even more pronounced for screening analysis in complex matrixes, due to the numbers of compounds involved. Therefore, there is an overall need to strengthen compound identification using alternative solutions, despite the availability of accurate mass instrumentation on the market, which are able to greatly reduce the number of possible hits. As a complementary tool, LRI prediction values from other databases (e.g., flavors and tobacco-related compounds, in our research context) should be available to increase the confidence level for compound identity. Ideally, LRIs could be predicted for millions of compounds present in public databases such as PubChem[2,3] and ChemSpider,[4] especially considering the low quantity of relevant data available in Wiley and NIST 2014 mass spectral databases (82 337 compounds).[1] As a proof of concept, all compounds registered in an in-house tobacco, tobacco-related compounds, as well as characteristic flavor compounds (n = 11 000), known as Unique Compound

DOI: 10.1021/acs.analchem.6b00868
Anal. Chem. 2016, 88, 7539−7547

Compound name	CAS #	NIST14 Agilent Score	LRI_expt	LRI_RM (RM - expt)	LRI_CG (CG - expt)	Final Score NIST14, $LRI_{(RM)}$, $LRI_{(CG)}$
1,1a,6,6a-tetrahydrocycloprop[a]indene	15677-15-3	92.1	1204.02	1136.62 (-67.4)	1209.15 (+5.1)	86.84
1,2-dihydronaphthalene	447-53-0	91.4		1262.91 (+58.9)	1211.44 (+7.4)	86.59
2-methyl-1H-indene	2177-47-1	91.0		1218.53 (+14.5)	1210.10 (+6.1)	89.50
3-methyl-1H-indene	767-60-2	90.7		1213.31 (+9.3)	1210.10 (+6.1)	**89.58**
1-methyl-4-prop-1-ynylbenzene	2749-93-1	90.2		1207.60 (+3.6)	1065.30 (-138.7)	80.64
1-cycloprop-2-en-1-yl-2-methylbenzene	65051-83-4	89.3		1158.66 (-45.4)	1071.24 (-132.8)	77.54
1,4-dihydronaphtalene	612-17-9	89.3		1255.34 (+51.3)	1238.46 (+34.4)	83.24

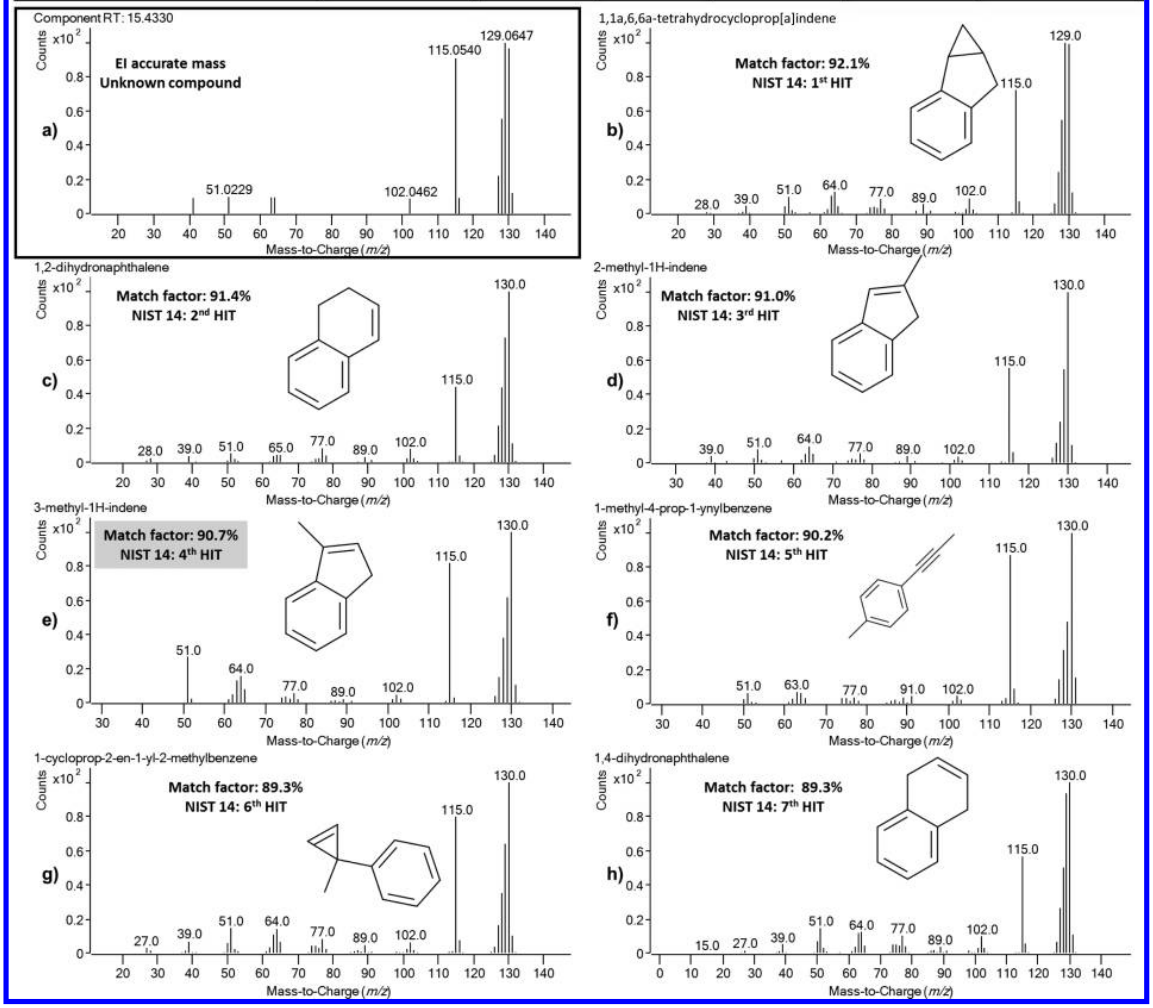


Figure 4. Case study for the assessment of unknown compound ($C_{10}H_{10}$) observed during the analysis of a 3R4F aerosol reference cigarette analyzed by GC-HR-MS. The calculation of final matching score takes into account both NIST 14 Agilent score and precision of LRI experimental values against predicted ones calculated from RapidMiner and ACD/ChromGenius models (see calculation formula in the Supporting Information figures). (a) EI accurate mass spectrum of the unknown compound eluted at a retention time of 15.433 min (LRI = 1204.016) and (b–h) EI nominal mass spectra of NIST 14 candidates.

Spectra Database (UCSD), were used.[38] LRI values were predicted for each compound, using both ACD/ChromGenius and RapidMiner approaches. Only chemicals with predicted LRI values between 500 and 1 900 units (i.e., matching values

DOI: 10.1021/acs.analchem.6b00868
Anal. Chem. 2016, 88, 7539–7547

within C5 and C19 odd *n*-alkanes) were used as relevant plausible compounds, observable under our analytical conditions (i.e., DB-624 column and temperature gradient conditions). Beside the compounds used within the training and test sets (n = 552), an additional 4 718 chemicals were suitable candidates for the chromatographic conditions. Figure 3 represents the number of compounds clustered according to LRI difference (ACD/ChromGenius values minus RapidMiner values) using bin sizes of 100 units. Overall, 62.2% of the compounds (n = 2 931) had LRI values predicted by both software approaches within 100 absolute units of each other. A script in Pipeline Pilot was developed to back calculate absolute retention time values from the mean of the LRI values predicted by both software approaches and odd *n*-alkanes reference standard retention times. A comma separated value (.csv) file was then created and used in combination with processed CI and/or EI accurate mass raw data obtained from GC-HR-MS analysis (Figure 1b) to increase the confidence level for compound identification and, most importantly, to lower the rate of false positive results. An additional set of 23 reference standards analyzed recently provided a good match between experimental and predicted values, within accuracy between 87 and 112% (Figure 2c,d). Figure 4 highlights a particular example where 3-methyl-1H-indene was recently confirmed with the analysis of reference standard. The EI accurate mass spectrum (Figure 4a) of the compound eluted at a retention time of 15.433 min (experimental LRI of 1 204) was search against the NIST 14 database library through MassHunter Unknown Analysis software. 1,1a,6,6a-Tetrahydrocycloprop[a]-indene was retrieved as the most probable compound, and we have listed six alternative proposal hits matching the unknown MS spectrum with good scoring from 89.3 up to 92.1% (Agilent scoring). Nominal EI mass spectra of all these compounds, having similar formulas of $C_{10}H_{10}$, revealed a close fragmentation pattern (Figure 4b–h). LRI values of all these candidates were predicted with both RapidMiner and ACD/ChromGenius models, and a final candidates scoring was calculated by combining NIST 14 score and differences between experimental LRI value to the predicted ones. Under these considerations, 3-methyl-1H-indene was ranked as the first candidate with an improved discriminatory power (final score of the seven proposals ranging from 77.54 up to 89.58%). Moreover, this example illustrates the impact of keeping the two prediction models for unknown analysis. Indeed, 1,1a,6,6a-tetrahydrocycloprop[a]-indene or 1-methyl-4-prop-1-ynylbenzene would have been ranked as first proposal, respectively, using ACD/ChromGenius or RapidMiner models only (Figure 4). In cases where no positive hit could be confirmed with a reference standard, other databases (e.g., specific flavor databases, ChemSpider or PubChem) could be used to identify compounds using elemental formulas determined from CI data as a query constraint. The application of LRI prediction models to large sets of molecules, together with a comparison of experimental with *in silico* fragmentation spectra will provide a highly relevant approach to strengthen the confidence level for compound identification before final confirmation.

CONCLUSION

A GC-HR-MS method using a DB-624UI column with headspace and liquid injection modes was developed to monitor volatile and semivolatile compounds present in tobacco aerosol and aerosol fraction samples. A set of reference standards (n = 552), covering a broad range of chemical diversity, was analyzed and experimental data were used to build a personal compound database accurate mass library. Linear retention index values were calculated by bracketing with odd *n*-alkanes, from pentane up to nonadecane. Two software approaches, RapidMiner (coupled to Dragon) and ACD/ChromGenius, were used to build independent LRI prediction models using a training set of compounds (n = 401). The accuracy of the models was assessed using a test set of compounds (n = 151) plus an additional set of 23 compounds analyzed recently. Both prediction models performed very well with correlation coefficient for the training and test set of 0.966 and 0.949 for RapidMiner and 0.977 and 0.976 for ACD/ChromGenius, respectively. Although these models could be improved by increasing the number of representative compound classes, all chemicals used in the test set fitted within an accuracy limit of 85−115% (except with one mixed halogen compound: 2-bromo-2-chloro-1,1,1-trifluoroethane) when using the mean predicted value of both software approaches. Indeed, mean LRI predicted values for putative compounds, where the smallest differences between the two predicted values occurred, was found to be a reliable tool for complementing accurate mass measurements and strengthening the confidence level for compound identification during the analysis of complex matrixes. These models will be dynamically improved when additional experimental data become available. In addition, it has been demonstrated that sharing retention index data across laboratories is feasible and will greatly assist in the difficult task to unambiguously identify compounds present in complex matrixes. Overall, the use of predicted linear retention index enables one to shorten the list of putative chemicals to order or synthesize for full confirmation.

ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.analchem.6b00868.

Plot of chemical space of compounds reported in tobacco and tobacco smoke *A* log *P* against molecular weight and boiling point against *A* log *P*; data comparison of LRI values across different laboratories plotted against our experimental results; and final compound scoring (PDF)

Descriptions of the reference standards used for the training set, test set, and identified compounds and results; calculation of LRI values; and description of the reference standards used as benchmarking our experimental LIR against other laboratories (XLSX)

List of Dragon descriptors used to build RapidMiner model and results for the model optimization; results for the model optimization for ChromGenius software; and classification of the reference standards used for the test and validation sets (XLSX)

RapidMiner protocol (PDF)

AUTHOR INFORMATION

Corresponding Author

*E-mail: philippealexandre.guy@pmi.com. Phone: 00-41-58-242-2622.

Notes

The authors declare no competing financial interest.

DOI: 10.1021/acs.analchem.6b00868
Anal. Chem. 2016, 88, 7539−7547

■ ACKNOWLEDGMENTS

The authors would like to acknowledge Arno Knorr for critical reviewing of the manuscript.

■ REFERENCES

(1) *Wiley Registry: Mass Spectral Library*, 10th ed./*NIST 2014*, 2014; http://eu.wiley.com/WileyCDA/WileyTitle/productCd-1119004144.html.
(2) Substance record page release, *PubChem blog*, 2015; http://pubchemblog.ncbi.nlm.nih.gov/2015/04/09/substance-record-page-released/.
(3) Kim, S.; Thiessen, P. A.; Bolton, E. E.; Chen, J.; Fu, G.; Gindulyte, A.; Han, L.; He, J.; He, S.; Shoemaker, B. A.; Wang, J.; Yu, B.; Zhang, J.; Bryant, S. H. *Nucleic Acids Res.* **2016**, *44*, D1202−D1213.
(4) Royal Society of Chemistry, 2015; http://www.chemspider.com.
(5) *Mass Frontier software*, version 7.0, 2014; http://highchem.com/index.php/support/.
(6) Zhou, J.; Weber, R. J. M.; Allwood, J. W.; Mistrik, R.; Zhu, Z.; Ji, Z.; Chen, S.; Dunn, W. B.; He, S.; Viant, M. R. *Bioinformatics* **2014**, *30* (4), 581−583.
(7) NIST Mass Spectrometry Data Center. *Mass Spectrum Interpreter*, ver. 2; 2011; http://www.chemdata.nist.gov/mass-spc/interpreter/.
(8) *ACD/MS Fragmenter*, version 12; 2012; http://www.acdlabs.com/products/adh/ms/ms_frag/.
(9) Heinonen, M.; Rantanen, A.; Mielikainen, T.; Kokkonen, J.; Kiuru, J.; Ketola, R. A.; Rousu, J. *Rapid Commun. Mass Spectrom.* **2008**, *22*, 3043−3052.
(10) *Fragment Identificator Algorithm*, version 1.3; 2008; http://www.cs.helsinki.fi/group/sysfys/software/fragid/.
(11) Wegner, A.; Weindl, D.; Jager, C.; Sapcariu, S. C.; Dong, X.; Stephanopoulos, G.; Hiller, K. *Anal. Chem.* **2014**, *86* (4), 2221−2228.
(12) Fragment Formula Calculator (FFC) Algorithm; 2013; http://www.ffc.lu/.
(13) Hill, A. W.; Mortishire-Smith, R. J. *Rapid Commun. Mass Spectrom.* **2005**, *19*, 3111−3118.
(14) Wolf, S.; Schmidt, S.; Muller-Hannemann, M.; Neumann, S. *BMC Bioinf.* **2010**, *11*, 148.
(15) In silico fragmentation for computer assisted identification of metabolite mass spectra, *MetFrag Algorithm*; 2011; http://msbi.ipb-halle.de/MetFrag/.
(16) Schymanski, E. L.; Gallampois, C. M.; Krauss, M.; Meringer, M.; Neumann, S.; Schulze, T.; Wolf, S.; Brack, W. *Anal. Chem.* **2012**, *84* (7), 3287−3295.
(17) Ruttkies, C.; Schymanski, E. L.; Wolf, S.; Hollender, J.; Neumann, S. *J. Cheminf.* **2016**, *8*, 3.
(18) Schymanski, E. L.; Meringer, M.; Brack, W. *Anal. Chem.* **2009**, *81* (9), 3608−3617.
(19) McLafferty, F. W.; Turecek, F. *Interpretation of Mass Spectra*, 4th ed.; University Science Books, Sausalito, CA, 1994.
(20) Kim, S.; Zhang, X. *J. Chemom.* **2015**, *29* (2), 80−86.
(21) Schymanski, E. L.; Meringer, M.; Brack, W. *Anal. Chem.* **2011**, *83* (3), 903−912.
(22) Boswell, P. G.; Abate-Pella, D.; Hewitt, J. T. *J. Chromatogr. A* **2015**, *1412*, 52−58.
(23) Hall, L. M.; Hill, W.; Menikarachchi, L. C.; Chen, M.-H.; Hall, L. H.; Grant, D. F. *Bioanalysis* **2015**, *7* (8), 939−955.
(24) Quilliam, M. A.*Retention Index Standards for Liquid Chromatography*, PTC International Application WO 2013134862, March 15, 2013.
(25) Menikarachchi, L. C.; Cawley, S.; Hill, D. W.; Hall, L. M.; Hall, L.; Lai, S.; Wilder, J.; Grant, D. F. *Anal. Chem.* **2012**, *84* (21), 9388−9394.
(26) Menikarachchi, L. C.; Hamdalla, M. A.; Hill, D. W.; Grant, D. F. *Comput. Struct. Biotechnol. J.* **2013**, *5* (6), 1−7.
(27) Kumari, S.; Stevens, D.; Kind, T.; Denkert, C.; Fiehn, O. *Anal. Chem.* **2011**, *83* (15), 5895−5902.
(28) Kupska, M.; Chmiel, T.; Jedrkiewicz, R.; Wardencki, W.; Namiesnik, J. *Food Chem.* **2014**, *152*, 88−93.
(29) Knorr, A.; Monge, A.; Stueber, M.; Stratmann, A.; Arndt, D.; Martin, E.; Pospisil, P. *Anal. Chem.* **2013**, *85*, 11216−11224.
(30) Garkani-Nejad, Z.; Karlovits, M.; Demuth, W.; Stimpfl, T.; Vycudilik, W.; Jalali-Heravi, M.; Varmuza, K. *J. Chrom. A* **2004**, *1028*, 287−295.
(31) Stein, S. E.; Babushok, V. I.; Brown, R. L.; Linstrom, P. J. *J. Chem. Inf. Model.* **2007**, *47*, 975−980.
(32) ACD/Labs. *ChromGenius software*; 2015; http://www.acdlabs.com/products/com_iden/meth_dev/chromgen/index.php.
(33) Accelrys Inc. *Pipeline Pilot*; http://accelrys.com.
(34) Talete SRL. *Dragon*; 2013; http://www.talete.mi.it/.
(35) RapidMiner Ltd.; http://rapidminer.com.
(36) Rodgman, A.; Perfetti, T. A. *The Chemical Components of Tobacco and Tobacco Smoke*, 2nd ed.; CRC Press: Boca Raton, FL, 2013.
(37) Leffingwell, J. C.; Young, H. J.; Bernasek, E. *Tobacco Flavoring for Smoking Products*; R. J. Reynolds Tobacco Company: Winston-Salem, NC, 1972.
(38) Martin, E.; Monge, A.; Duret, J. A.; Gualandi, F.; Peitsch, M. C.; Pospisil, P. *J. Cheminf.* **2012**, *4* (1), 1−11.
(39) Perfetti, T. A.; Rodgman, A. *Beitr. Tabakforsch. Int.* **2011**, *24* (5), 215−232.
(40) Rodgman, A.; Perfetti, T. A. *Beitr. Tabakforsch. Int.* **2009**, *23* (5), 277−333.
(41) Elkan, C. *Evaluating Classifiers*, 2012; http://cseweb.ucsd.edu/~elkan/250Bwinter2012/classifiereval.pdf.
(42) Dice, L. R. *Ecology* **1945**, *26* (3), 297−302.
(43) Sharp, M.-E. *J. Anal. Toxicol.* **2001**, *25* (7), 631−636.
(44) Seeley, J. V.; Seeley, S. K. *J. Chrom. A* **2007**, *1172*, 72−83.
(45) Rood, D. *Solvent retention data DB-624, DB-1, DB-WAX columns specified in USP Method 467*, Agilent Technologies Application Note, 2002; http://www.chem.agilent.com/cag/cabu/pdf/b-0292.pdf.
(46) Wei, X.; Koo, I.; Kim, S.; Zhang, X. *Analyst* **2014**, *139* (10), 2507−2514.

DOI: 10.1021/acs.analchem.6b00868
Anal. Chem. 2016, 88, 7539−7547