analytical chemistry

Article

pubs.acs.org/ac

Computer-Assisted Structure Identification (CASI)—An Automated Platform for High-Throughput Identification of Small Molecules by Two-Dimensional Gas Chromatography Coupled to Mass Spectrometry

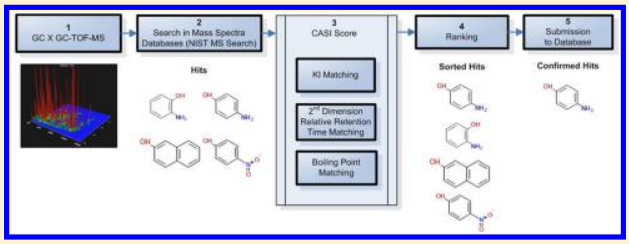
Arno Knorr,*[,†] Aurelien Monge,[†] Markus Stueber,[‡] André Stratmann,[‡] Daniel Arndt,[†] Elyette Martin,[†] and Pavel Pospisil[†]

[†]Philip Morris International R&D, Philip Morris Products S.A., 2000 Neuchâtel, Switzerland

[‡]Philip Morris International R&D, Philip Morris Research Laboratories GmbH, 51149 Köln, Germany

Supporting Information

ABSTRACT: Compound identification is widely recognized as a major bottleneck for modern metabolomic approaches and high-throughput nontargeted characterization of complex matrices. To tackle this challenge, an automated platform entitled computer-assisted structure identification (CASI) was designed and developed in order to accelerate and standardize the identification of compound structures. In the first step of the process, CASI automatically searches mass spectral libraries for matches using a NIST MS Search algorithm, which proposes structural candidates for experimental spectra from two-dimensional gas chromatography with time-of-flight mass spectrometry (GC × GC-TOF-MS) measurements, each with an associated match factor. Next, quantitative structure-property relationship (QSPR) models implemented in CASI predict three specific parameters to enhance the confidence for correct compound identification, which were Kovats Index (KI) for the first dimension (1D) separation, relative retention time for the second dimension separation (2DrelRT) and boiling point (BP). In order to reduce the impact of chromatographic variability on the second dimension retention time, a concept based upon hypothetical reference points from linear regressions of a deuterated n-alkanes reference system was introduced, providing a more stable relative retention time measurement. Predicted values for KI and 2DrelRT were calculated and matched with experimentally derived values. Boiling points derived from 1D separations were matched with predicted boiling points, calculated from the chemical structures of the candidates. As a last step, CASI combines the NIST MS Search match factors (NIST MF) with up to three predicted parameter matches from the QSPR models to generate a combined CASI Score representing the measure of confidence for the identification. Threshold values were applied to the CASI Scores assigned to proposed structures, which improved the accuracy for the classification of true/false positives and true/false negatives. Results for the identification of compounds have been validated, and it has been demonstrated that identification using CASI is more accurate than using NIST MS Search alone. CASI is an easily accessible web-interfaced software platform which represents an innovative, high-throughput system that allows fast and accurate identification of constituents in complex matrices, such as those requiring 2D separation techniques.



Recent developments in analytical techniques for comprehensive nontargeted screening of small molecules in complex matrices using chromatographic separation coupled with mass spectrometry techniques, as used in the field of metabolomics, has resulted in the generation of huge amounts of data.[1−3] Gas chromatography−mass spectrometry (GC-MS) is one of the well-established analytical techniques used to separate and detect individual compounds within complex mixtures, such as biological samples or cigarette smoke. While traditional 1-dimensional (1D) separation techniques are limited in their capacity to separate the components within highly complex matrices such as cigarette smoke, which is estimated to contain several thousands of compounds,[4] two-dimensional separation techniques (often referred to as 2D-GC-MS or GC × GC-MS)[5] were established to increase the resolution of the chromatographic separation. Structural identification of resolved components by simple mass spectral library queries provides insufficient confidence to trust in the proposed structure. In order to increase the level of confidence, manual verification, and interpretation of the mass spectral library search has to be carried out. Finally, for unequivocal compound identification, confirmation with reference standards is required.

Owing to the fact that this is very costly and time-consuming, this can, in most cases, only be performed for a limited number of compound structures.

Received: March 27, 2013
Accepted: October 25, 2013
Published: October 25, 2013

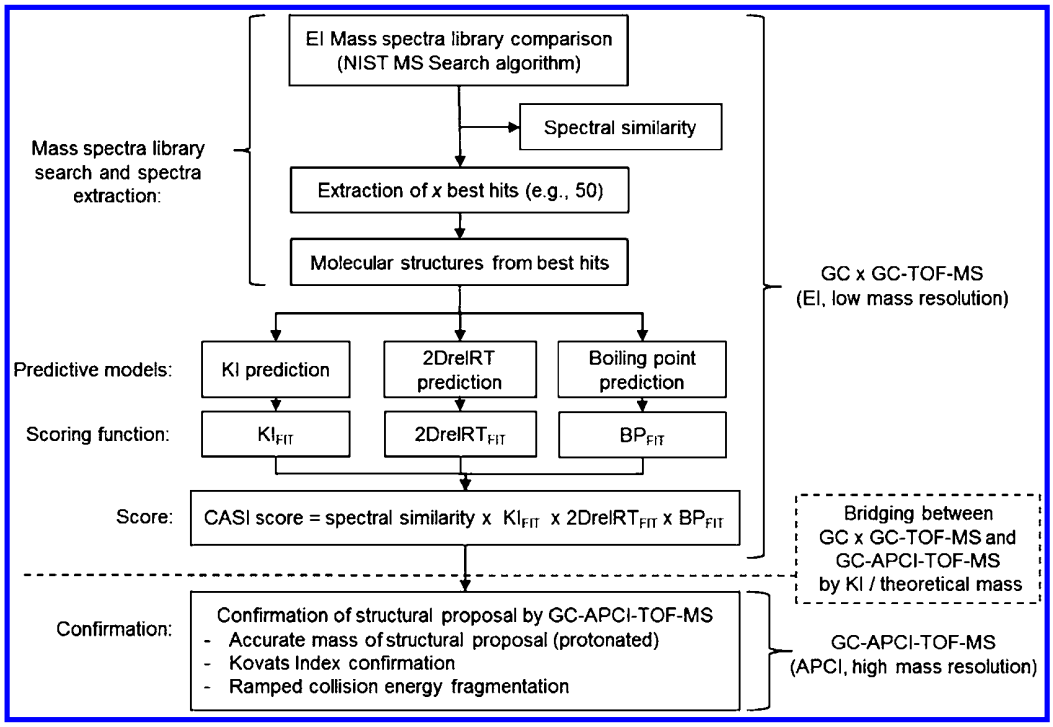


Figure 1. CASI concept. The general CASI concept is based upon the generation of proposals from mass spectra library searches followed by subsequent refinement using prediction models for separation and boiling point. Lower part represents the proof of concept for confirmation of the structural proposals using accurate mass gas chromatography atmospheric pressure chemical ionization TOF-MS (GC-APCI-TOF-MS).

A strategy to enhance the level of confidence in structural proposals from mass-spectral library searches is the use of additional parameters (orthogonal information) derived from the chromatographic analysis, such as the retention time or Kovats Index (KI).[6] Therefore, the inclusion of prediction model parameters such as KI can dramatically improve the certainty for compound identification. Several such models, predicting KI values based upon molecular structures and properties, and termed as quantitative structure−property relationship (QSPR) models, have been published.[7−9]

In the same way, second dimension retention times from GC × GC-TOF-MS can be used as additional information. The correlation of analytes retention times and their boiling points using nonpolar columns and linear ramped temperature gradients for separation is known for a long time and can provide additional information in confirming a structural proposal.[10−12] Consequently, a high-throughput computer-assisted system that incorporates models for KI, second dimension relative retention time (2DrelRT) and boiling point has been developed, which increases confidence in the accuracy of compounds identified by GC × GC-TOF-MS. The system is termed "computer-assisted structure identification" (CASI), with the objective to accelerate, standardize and ensure the reproducibility for the identification of compound structures.

■ CONCEPT

The concept of CASI is based upon the extraction of proposed molecular structures (hits) from mass spectral libraries, which have the highest match with the queried experimental spectra, and subsequent refinement of the proposed hits by matching experimental chromatographic parameters with predicted parameters derived from their chemical structures (Figure 1).

First, mass spectra are submitted to a search for structural candidates and their associated match factors in mass spectra databases using NIST MS Search version 2.0g[13] (see mass spectra databases used for the validation of CASI in Supporting Information).

For the next step, specific QSPR models were developed using different molecular descriptors to predict essential parameters for enhancing the confidence in compound identification: KI values for the first dimension separation and 2DrelRT values for the second dimension separation. In addition, predicted boiling points are calculated by ACD/PhysChem Batch software.[14] The predicted values for each hit are then directly compared with experimentally determined values, with the exception of boiling point, which is compared with a value derived from the first dimension retention time. Finally, CASI combines each match result from NIST MS Search with the corresponding matches for the aforementioned predicted parameters to create a new match score, referred to as the CASI score. False positive (incorrect hits) identifications are minimized by ensuring that absolute score values exceed a specific threshold.

In the second step, we were questioning if by adding accurate and high-resolution mass information would further improve confidence in the proposed structure (see Confirmation in Figure 1).

■ MATERIALS AND METHODS

Instrumentation and Analytical Methods. *Data Generation.* The experiments were performed using a LECO Pegasus IV[15] GC × GC-TOF-MS system and a Bruker micrOTOF-Q II[16] orthogonal accelerated TOF mass spectrometer coupled to GC using an APCI source. Cigarette smoke, collected on glass-fiber filter pads, was extracted with dichloromethane/acetone (80:20) and fortified with a mixture of several

dx.doi.org/10.1021/ac4011952 | Anal. Chem. 2013, 85, 11216−11224

deuterated internal standards and retention time marker compounds. The cigarette smoke extracts were analyzed (i) using dichloromethane/water partitioning and injection of the dichloromethane extract and (ii) as crude extract derivatized using N,O-bis(trimethylsilyl)trifluoroacetamide (BSTFA) + trimethylsilyl chloride (TMSCI) (99:1) reagents and subsequent injection using cool-on-column mode.

The experiments performed using GC × GC-TOF-MS and GC-APCI-TOF-MS are described in detail in Supporting Information section.

Data Processing. Data processing was performed using a nontargeted screening setup with LECO ChromaTOF software version 3.34[17] for automatic peak finding, spectral deconvolution, and peak alignment. Subsequent data evaluation, with a focus on the most relevant differences in chemical composition was performed.

Data Sets. The data sets used for the development of the CASI platform were generated using a method for nontargeted comparison of different cigarette smoke samples. In order to cover a wide range of polarities, two independent analyses per aerosol sample were performed to generate each data set, the first for nonpolar compounds and the second for polar compounds derivatized by trimethylsilylation. In total, the data sets used comprised chromatograms and spectra for 218 structures confirmed by reference compounds, plus data for a further 176 unknown compounds. The procedure that was applied to categorize a chromatographic peak as unknown (not identified) was based on a mass-spectrometry and chromatography expert knowledge based decision process (see Supporting Information).

The structural diversity of the data sets used is demonstrated by the range of chemical compound classes covered: aliphatic and alicyclic hydrocarbons, aromatic hydrocarbons, aliphatic and alicyclic alcohols, polyols, polyolesters, chlorinated polyols, terpenes, O-heterocycles, phenols, quinones, ketones and aldehydes, hydroxyketones, small organic acids and fatty acids, fatty acids alkyl esters, phthalate esters, nitrocompounds, nitrosamines, N-heterocycles, hydroxy-N-heterocycles, steroids, aromatic amines, imides, and siloxanes (detailed description in Supporting Information).

Generation of the Second Column Relative Retention Time. For the calculation of 2DrelRT a novel experimental model was developed. In the CASI approach, 2DrelRT is derived from second dimension peaks in relation to hypothetical reference points based upon linear regressions of deuterated *n*-alkanes (Figure 2). The *n*-alkanes are used to generate a hypothetical second dimension retention time reference system, compensating for systematic shifts (such as different column length or gas flow) but not for any shifts related to analyte-stationary phase interaction, as these shifts are dependent upon individual compound properties.

The second dimension relative retention time of a compound is calculated as follows:

$$2\mathrm{DrelRT} = \frac{\mathrm{abs2DRT}}{2\mathrm{DRT_{hrf}}} \tag{1}$$

where abs2DRT is the measured second dimension retention time of the compound and $2\mathrm{DRT_{hrf}}$ is the second dimension retention time of the hypothetical reference point.

For a given compound that elutes between deuterated n-alkane standard compound 1 and compound 2, the $2\mathrm{DRT_{hrf}}$ is calculated using the linear equation $y = ax + b$

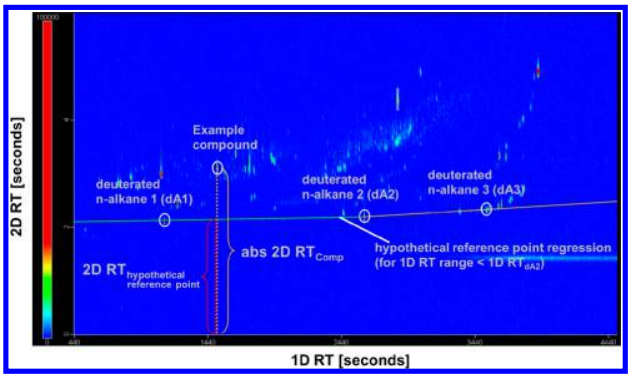


Figure 2. Principals of the second dimension relative retention time generation. Hypothetical reference points (full line) are derived from the linear regression of the experimentally measured retention times of deuterated n-alkanes in the two-dimensional separation space.

$$2\mathrm{DRT_{hrf}} = \frac{2\mathrm{DRT_{dA2}} - 2\mathrm{DRT_{dA1}}}{1\mathrm{DRT_{dA2}} - 1\mathrm{DRT_{dA1}}} \times 1\mathrm{DRT} + \left(2\mathrm{DRT_{dA1}} - \frac{2\mathrm{DRT_{dA2}} - 2\mathrm{DRT_{dA1}}}{1\mathrm{DRT_{dA2}} - 1\mathrm{DRT_{dA1}}} \times 1\mathrm{DRT_{dA1}}\right) \tag{2}$$

where $a = (2\mathrm{DRT_{dA2}} - 2\mathrm{DRT_{dA1}})/(1\mathrm{DRT_{dA2}} - 1\mathrm{DRT_{dA1}})$ is describing the slope and b is calculated by substituting x using the known values for dA1.

The equation resolved using the known values for dA1 for $b = y_{\mathrm{dA1}} - ax$ leads to

$$b = 2\mathrm{DRT_{dA1}} - \frac{2\mathrm{DRT_{dA2}} - 2\mathrm{DRT_{dA1}}}{1\mathrm{DRT_{dA2}} - 1\mathrm{DRT_{dA1}}} \times 1\mathrm{DRT_{dA1}}$$

where dA1 and dA2 are deuterated *n*-alkane 1 and 2, respectively, and 1DRT and 2DRT are the first and second dimension retention time of the respective molecules.

QSPR Modeling Methods for Predicting KI and 2DrelRT. The retention times for 218 unique commercially sourced compounds analyzed using GC × GC-TOF-MS were used to build the KI and 2DrelRT models. The compounds were split randomly into a training set (118 compounds), a test set (40), and a validation set (60). The training set was used to select the descriptors for QSPR modeling and to build the models. Performances of the predictions on training set and test set were used to identify the best models.

The chemical structures of the 218 compounds were standardized using Pipeline Pilot,[18] the process for which being described later in this publication. From each of these standardized structures, 2489 two-dimensional descriptors were computed using Dragon.[19] Noninformative descriptors (constant and near-constant variables and pair correlation with a threshold of 0.98) were subsequently excluded, after which 370 descriptors remained. For each model a smaller number of descriptors were then selected using genetic algorithms. Models were built using three learning algorithms, *k*-nearest neighbors (*k*-NN), multilinear regression (MLR) and support vector regression (SVR) (see Supporting Information section).

The *k*-NN, MLR, and SVR learning algorithms were used within the RapidAnalytics 5[20] software environment. Genetic algorithms were developed in Java to select the descriptors to be used in each model. Scoring was executed using a RapidAnalytics protocol with a cross validation squared correlation (Q^2) function used for *k*-NN and MLR and root mean squared

error (RMSE) function for SVR (detailed description in Supporting Information).

Derivation of Boiling Point from Kovats Index. Boiling points for structural candidates are calculated from their proposed structures using ACD/Labs PhysChem Batch software. These calculated boiling points for structural candidates (hits) are then matched with boiling points derived from experimentally measured KI values.[21] A correlation between measured KI and calculated boiling point was developed using the training set of compounds.

Algorithm for Scoring Structure Candidates. Scores are calculated from the NIST MS Search match factor, predicted KI, predicted 2DrelRT and the calculated boiling point, using hyperbolic equations. The general principle is to factor scores for similarity of experimental mass spectra to library mass spectra, with scores derived from each analytical property (KI, 2DrelRT, ...). The analytical property scores (KI_{FIT}, $2DrelRT_{FIT}$, ...) are normalized from 0 (no similarity) to 1 (perfect match) and are based on quadratic equations using polynomials factorization (Figure 3). A detailed explanation can be found in the Supporting Information.

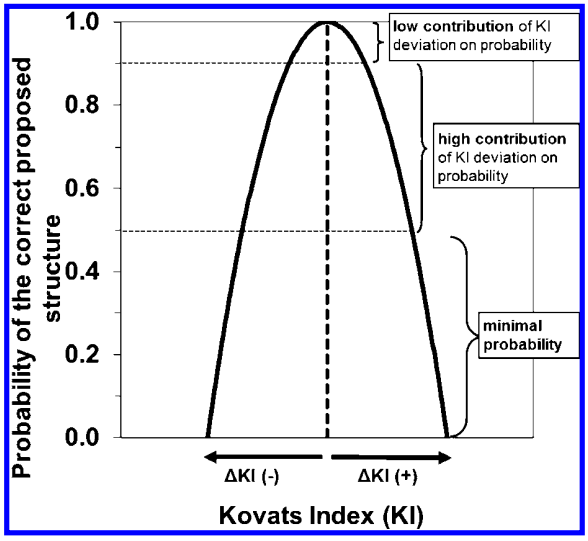


Figure 3. Contribution equation of a scoring module: example for KI_{FIT}. The curve shows the dependency of the probability for the correct proposal (*y* axis) based upon the deviation of predicted KI (*x* axis). The greater the deviation between experimental and predicted KI, the lower the probability for the proposal, depending upon the curve fitting function used. The greater the steepness of the curve, the greater the effect of any deviation between experimental and predicted values on the probability for the proposal, and the higher the impact on the overall CASI score.

Optimization of the CASI Score. To optimize the contribution of each module to the final score (CASI Score), a weighting scheme was developed. The value at which the hyperbolic curve crossed the X axis for each module, as defined by the steepness of the hyperbolic function, was used to weight each module's contribution to the final CASI Score. A grid search procedure was established in order to define optimal values for n_{KI}, $n_{2DrelRT}$, and n_{BP} and all possible solutions were generated. The solution score was the number of correct hits achieved and the solution with the highest number of correct hits was selected for use. The algorithm is described in Supporting Information.

Software Development and Architecture. To automate the entire process, the CASI software platform was developed. The software is accessible via a web interface to enter mass spectra as a multi JDX file, KI values, 2DrelRT values, and any additional information required to describe the experiment. Each submitted mass spectrum is queried versus commercially available mass spectra databases using NIST MS Search (see Supporting Information). Chemical names for the hits are then converted into chemical structures using an Accelrys Pipeline Pilot workflow (accelrys.com), which combines searches in the PMI corporate chemical registry database (UCSD),[22] Pubchem[23] and ChemSpider[24] with the functionality of ACD/Laboratories Name-to-Structure Batch software version 12.[25] Models are applied to predict the Kovats index, 2DrelRT, and boiling point for each hit, which are then compared with experimentally measured values to create match scores. These calculated match scores are combined with the match factor from NIST MS Search, as described previously, to give a CASI Score (Figure 1). Finally the hits for each query are listed in order of decreasing CASI Score, which the user can then view via a dedicated web interface (see Supporting Information).

The CASI platform was developed using several external software components, the architecture for which is presented in Supporting Information.

RESULTS AND DISCUSSION

Second Column Relative Retention Time Performance. The conventional procedure to utilize chromatographic information from the second dimension of GC × GC-TOF-MS analysis is based upon absolute retention times. The advantage of having a relative model for the second dimension retention time (currently no known method available) is the ability to compensate for any chromatographic fluctuations.

The performance of this model for 2DrelRT was tested by evaluating the reproducibility of absolute retention times versus relative retention times for merged data sets from three independent nontargeted screening studies comparing different smoke extract samples. The focus of this evaluation was made using data from a reference cigarette, which was used as a quality standard and was analyzed in triplicate within each study (detailed description in Supporting Information Section). The evaluation of the nine independent chromatograms was performed in a nontargeted way, with peaks having a signal-to-noise ratio exceeding 250 being selected. The total number of evaluated compounds, regardless if putatively identified or not and without outlier correction, was 1219. The results show that the relative standard deviation for the 90th percentile of all evaluated compounds for the entire data set was reduced from 4.3% for the second dimension absolute retention time (2 DabsRT) data down to 2.5% when using the 2DrelRT system (Figure 4).

Results for the Prediction of KI and 2DrelRT using QSPR models. The predictive models for Kovats Indices were generated using genetic algorithms in combination with either MLR or *k*-NN algorithms. The best model for MLR and *k*-NN algorithms were selected for comparison and the best of them was selected to be included for the CASI Score function calculation. The best results were obtained using MLR with seven descriptors, where the squared correlation (r^2) was 0.981 and the relative error was 5.18% for the validation set (see results and the correlation for KI prediction in Supporting Information Section). The main contributing descriptor was the number of C−C bonds (F01[C−C]). Other descriptors related to size and

dx.doi.org/10.1021/ac4011952 | Anal. Chem. 2013, 85, 11216−11224

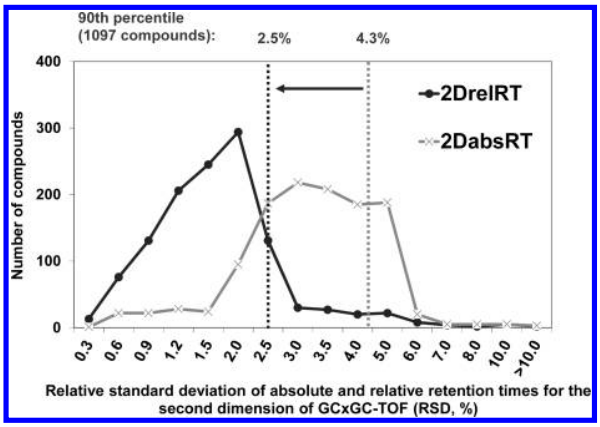


Figure 4. Number of identified compounds per relative standard deviation. Result on reproducibility ($N = 9$) using the relative retention time model compared to the absolute retention time for the second dimension of GC × GC-TOF-MS.

lipophilicity (H-047 and nCt) were part of the model and polarity descriptors were also involved (nN, EEig04d, TPSA-(NO) and B01[C−O]).

The best results for the 2DrelRT model were obtained using support vector machine with 12 descriptors, where the squared correlation (r^2) was 0.855 and the relative error was 6.76%. The nature of the descriptors used for the 2DrelRT model was more complex than for KI prediction (see list of descriptors in Supporting Information). A descriptor related to the number of benzene rings (nBnz) was present which was consistent with the presence of 50% of phenyl-groups in the used DB-17 column. Polar descriptors were also included (B03[C−O], F02[C−O], and Me). Although the predictive power of this model is high, it is not as accurate as the KI model due to the fact that the second dimension separation is affected by chromatographic variances in both dimensions as retention shifts in the first dimension are also causing subsequent retention shifts in the second dimension separation. For the list of descriptors and their definitions, results and correlation for 2DrelRT prediction see Supporting Information.

Prediction of Boiling Points from Kovats Indices. The linear equation resulting from the correlation of calculated boiling points with experimental Kovats Indices for the training set compounds was:

$$BP = 0.1549 \times KI + 31.725 \quad (3)$$

The r_2 value for this correlation was 0.953. For the test set compounds, the squared correlation between the boiling points obtained with this equation and the computed boiling points is 0.867 (0.867 at zero intercept). For the compounds of the validation set the squared correlation is 0.942 (0.940 at zero intercept).

Global Validation: CASI versus NIST. *Ability to Correctly Rank True Hits.* Optimization of the CASI score function was performed by varying the weighting for each of the prediction models (KI, 2DrelRT, and boiling point), and subsequently assessing the number of correct hits achieved for compounds contained within the training and test set. The standard error of prediction values for KI, 2DrelRT, and boiling point were also integral to the calculation of the CASI score, and were calculated using data from the same sets of compounds (SEP_{KI} = 82.57, $SEP_{2DrelRT}$ = 0.0771, and SEP_{BP} = 23.05). Each weighting combination applied to the CASI score equation resulted in a solution, and with weighting values represented by integers (n) between 1 and 50 for each prediction model, this represented a total of 125 000 solutions applied to each spectrum proposed by NIST MS Search. The test set compounds were used to perform a first pass evaluation of the entire range of solutions, and 93 of the 125 000 solutions enabled a maximum of 35 hits to be sorted correctly for a total of 40 queries (88%). These 93 solutions were then further filtered by applying them to the training set compounds, where a maximum of 94 compounds out of 118 (80%) were correctly identified using 11 of these solutions. The weighting values for KI and 2DrelRT were constant for these 11 solutions (n_{KI} = 11; $n_{2DrelRT}$ = 10) and the weighting values for boiling point (n_{BP}) were greater than or equal to 36. The solution with the lowest value for n_{BP} (= 36) was chosen to maintain the highest selectivity for this parameter. To ensure that this selection had no influence on the number of correctly identified results, we calculated the results for each solution for the validation set. Each of the solutions gave the same results, with 52 compounds being correctly identified from the 60 present within the validation set (87%), which indicates the lower impact of the BP module compared to KI and 2DrelRT modules.

Noteworthy, 2 hits were removed from the evaluation of ranking of correct hits for the validation set (complete set of hits can be seen in Supporting Information section). The first was a duplicate entry for stigmasterol using a different name. The second was a stereoisomer of campesterol which had the same score as the correct isomer. These examples illustrate the challenges and limitations of the selected techniques used with generic chromatographic systems.

Using the NIST MS Search match factor (NIST MF) alone, the number of correct hits for compounds within the validation set was 45 (75%), whereas the number achieved using the CASI Score was 52 (87%) (Table 1 and Figure 18 in Supporting Information).

The diversity of compounds contained within the training set and the validation set was evaluated by comparing calculated ECFP6 fingerprints (extended connectivity fingerprints with a diameter of 6)[26] using Tanimoto indices (for visualization see Supporting Information). One compound from the validation set, namely isobornyl acetate, was ranked in 27th position by CASI Score compared to first place ranking by NIST MF, clearly indicating an outlier compound for the CASI ranking procedure (Table 1). The predicted retention times and boiling points are the reason for the poor ranking (the variances were 19.3% for KI, 24.3% for 2DrelRT, and 8% for BP). An analysis of similarity between each compound of the validation set with each compound of the training set clearly showed that isobornyl acetate had the lowest similarity to any compound in the training set, most probably being outside the applicability domain of the models.

In addition, the individual impact of each of the modules KI, 2DrelRT, and boiling point upon the outcome of the CASI Score was evaluated, the parameters for each module having been optimized with the NIST MF in isolation. The results are presented in Table 2.

All three components KI, 2DrelRT, and BP improve the identification of correct hits in comparison to NIST MF alone. Best results were obtained using all three modules KI, 2DrelRT, and BP on all three data sets (training, test, and validation sets). The contribution of modules in decreasing order are KI, 2DrelRT and BP. The BP was kept as a component of the score in addition of KI because of the applicability domain of the BP model. BP is calculated with ACD/Labs PhysChem software

dx.doi.org/10.1021/ac4011952 | Anal. Chem. 2013, 85, 11216−11224

Table 1. List of Correct Structures Retrieved for Mass Spectra of the Validation Set[a]

no.	correct hit name	NIST MF	NIST rank	CASI score	CASI rank	KI exp	KI pred	2DrelRT exp	2DrelRT pred	BP from KI	BP pred
1	hexadecanoic acid, methyl ester	896	1	881	1	1936	1959	1.13	1.23	332	332
2	1-pentene, 2,3-dimethyl-	915	1	906	1	579	560	0.94	0.87	121	85
3	fluoranthene	871	2	869	1	2129	2136	1.84	1.87	362	375
4	2*H*-1-benzopyran-2-one, 7-hydroxy-6-methoxy-	923	1	858	1	1969	1759	2.13	2.05	337	413
5	acetic acid, phenyl ester	923	2	877	2	1089	1199	1.52	1.67	200	195
6	benzofuran, 2,3-dihydro-	868	1	853	2	1125	1193	1.58	1.66	206	188
7	*p*-benzoquinone, 2-methyl-	929	1	875	1	1038	1239	1.62	1.69	193	187
8	benzaldehyde, 4-hydroxy-3-methoxy-	632	3	620	1	1480	1446	1.81	1.91	261	283
9	d12-benz(a)pyren	542	7	431	1	2859	2660	2.32	2.01	475	495
10	*N*-nitrosonornicotine	821	1	797	1	1817	1864	2.1	1.98	313	369
11	1*H*-indole, 3-methyl-	939	1	922	1	1474	1365	1.78	1.73	260	265
12	stigmasta-5,22-dien-3-ol, 3α,22E)-	809	1	804	1	3159	3140	1.8	1.74	521	501
13	9,12-octadecadienoic acid, methyl ester, (E,E)-	883	1	856	1	2110	2216	1.2	1.3	359	373
14	d7-isoquino line	879	1	871	1	1341	1362	1.7	1.77	239	243
15	1-tetradecanol	878	12	873	1	1746	1696	1.13	1.13	302	263
16	2-dodecanone	924	1	917	1	1483	1448	1.16	1.1	261	248
17	anthracene	913	1	909	1	1867	1837	1.74	1.78	321	337
18	2-hexene, 2-methyl-	906	1	901	1	622	638	0.98	0.94	128	96
19	2-propanone, 1-chloro-	798	1	724	1	604	490	1.33	1.55	125	120
20	oleic acid	875	1	866	1	2158	2226	1.2	1.26	366	360
21	pentanoic acid, 4-oxo-, trimethylsilyl ester	613	8	582	7	1177	1139	1.34	1.17	214	202
22	2-cyclopenten-1-one, 3-methyl-	916	1	868	1	964	928	1.69	1.52	181	158
23	ergost-5-en-3α-ol	829	1	810	1	3135	3013	1.75	1.7	517	489
24	benzene, nitro-	909	1	591	6	1150	1321	1.59	2.03	210	211
25	3,6-dioxa-2,7-disilaoctane, 2,2,4,7,7-pentamethyl-	929	1	903	1	1006	926	1	0.89	188	167
26	phenol, 2,6-dimethoxy-	939	1	915	1	1427	1297	1.71	1.77	253	264
27	sorbic acid, trimethylsilyl	893	1	880	1	1235	1137	1.19	1.15	223	193
28	pyridine, 3-(3,4-dihydro-2*H*-pyrrol-5-yl)-	905	1	861	1	1526	1490	1.69	1.86	268	245
29	silane, (2-methoxyphenoxy)trimethyl-	919	1	917	1	1286	1246	1.31	1.32	231	215
30	3,6,9,12-tetraoxa-2,13-disilatetradecane, 2,2,13,13-tetramethyl-	895	1	852	1	1572	1374	1.12	1.13	275	271
31	naphthalene, 1-ethyl-	932	1	930	1	1489	1502	1.55	1.58	262	259
32	furan-2-carboxylic acid, 3-methyl-, trimethylsilyl ester	878	1	872	1	1274	1217	1.35	1.38	229	197
33	nonacosane	881	8	879	2	2850	2822	1.01	1.01	473	441
34	hydroquinone (2TMS)	901	2	885	1	1471	1397	1.12	1.04	260	246
35	triacontane	902	2	900	1	2929	2915	1.02	1.01	485	450
36	silane, (2-furanylmethoxy)trimethyl-	914	1	897	1	1003	995	1.21	1.11	187	172
37	2-propanone, 1-hydroxy-	879	1	836	1	580	637	1.3	1.46	122	145
38	hexadecane, 2,6,10,14-tetramethyl-	900	2	893	1	1845	1779	0.99	1.02	318	322
39	furfural	945	1	871	1	790	879	1.6	1.8	154	162
40	1-dodecanol	893	3	888	1	1560	1498	1.14	1.12	273	258
41	octanoic acid, ethyl ester	921	1	908	1	1253	1281	1.16	1.25	226	208
42	2-cyclopenten-1-one	942	1	936	1	787	850	1.65	1.68	154	136
43	2-hydroxydecanoic acid (2TMS)	916	1	910	1	1692	1755	1.03	1.02	294	321
44	D-ribopyranose, 1,2,3,4-tetrakis-*O*-trimethylsilyl-	700	1	608	1	1684	1770	1.24	0.98	293	368
45	1-butanamine, *N*-butyl-*N*-nitroso-	852	1	829	1	1347	1383	1.32	1.45	240	251
46	9,12-octadecadienoic acid (*Z*,*Z*)-, trimethylsilyl ester	999	1	988	1	2236	2319	1.09	1.12	378	402
47	pentanal	935	1	915	1	618	652	1.16	1.27	127	104
48	oleic acid, trimethylsilyl ester	834	1	832	1	2236	2278	1.09	1.09	378	401
49	hexadecanoic acid, ethyl ester	909	1	902	1	2001	2051	1.11	1.16	342	342
50	1,2-benzenedicarboxylic acid, bis(2-ethylhexyl) ester	880	6	746	3	2550	2826	1.37	1.56	427	385
51	dodecanoic acid, ethyl ester	906	1	876	1	1661	1675	1.13	1.26	289	269
52	cholesta-3,5-diene	616	10	560	1	2875	2716	1.38	1.57	477	458
53	docosane	933	1	930	1	2224	2170	1	1.01	376	368
54	dotriacontane	889	5	878	1	3085	3100	1.08	1	510	467
55	d19-decanoic acid	908	1	903	1	1418	1411	1.22	1.17	251	270
56	benzeneacetaldehyde	937	1	913	1	1078	1219	1.59	1.56	199	198
57	isobornyl acetate	854	1	653	27	1363	1627	1.29	1.6	243	223

dx.doi.org/10.1021/ac4011952 | *Anal. Chem.* 2013, 85, 11216−11224

Table 1. continued

no.	correct hit name	NIST MF	NIST rank	CASI score	CASI rank	KI exp	KI pred	2DrelRT exp	2DrelRT pred	BP from KI	BP pred
58	1,2,3-propanetriol, monoacetate	772	1	766	1	1125	1147	1.64	1.59	206	253
59	naphthalene, 1-methyl-	926	2	925	2	1414	1409	1.54	1.51	251	243
60	methyl triacontanoate	862	1	802	1	3206	3263	1.32	1.12	528	476

[a]no.: Position of the compound in the validation set. Correct Hit Name: Correct compound proposed by NIST MS Search. NIST MF: NIST match factor. NIST Rank: Rank of the correct structure according to NIST Match Factor. CASI Score: Score calculated by CASI. CASI Rank: Rank of the correct structure according to CASI score. KI exp.: Experimental Kovats Indices. KI pred.: Kovats Indices predicted from the structure of the correct hit. 2DrelRT exp.: Experimental 2D relative retention time. 2DrelRT pred: 2D relative retention time predicted from the structure of the correct hit. BP from KI: Boiling point calculated with a linear equation using experimental Kovats Index value. BP pred: Boiling point predicted from the structure of the correct hit using ACD/Laboratories Physchem Batch.

Table 2. Number of Correct Hits Using Different Combinations of CASI Score Modules[a]

	number of correct hits and % of the total (*n*)						
	training set (*n* = 118)		test set (*n* = 40)		validation set (*n* = 60)		overall (*n* = 218)
combination of modules	hits	%	hits	%	hits	%	%
NIST MF, KI, 2DrelRT, BP	94	79.7	35	87.5	52	86.7	83.0
NIST MF, KI, 2DrelRT	93	78.8	35	87.5	52	86.7	82.6
NIST MF, KI, BP	93	78.8	33	82.5	48	80.0	79.8
NIST MF, KI	92	78.0	33	82.5	48	80.0	79.4
NIST MF, 2DrelRT, BP	92	78.0	33	82.5	47	78.3	78.9
NIST MF, BP	92	78.0	32	80.0	47	78.3	78.4
NIST MF, 2DrelRT	90	76.3	31	77.5	45	75.0	76.1
NIST MF	87	73.7	30	75.0	45	75.0	74.3

[a]Results for the CASI score based on NIST MF and all three components are presented in the first line.

which uses a model based on a training set of 8500 molecules, thus it represents more robust model applicable for future analysis of different samples for example.

Ability to Discriminate True Hits from Unknowns. Ranking by itself is not sufficient to correctly identify structures and in cases where the correct structure is not present in the reference spectra database, any structure proposed by a mass spectral libraries-based search (CASI included) will be incorrect. A score threshold can assist the user in the decision making process and the CASI process was developed to combine a score threshold alongside the ranking procedure. In order to show the discriminative power of this threshold, correct hits from the validation set (60) and compounds with the highest score from a set of unknown compounds (176) using both NIST MF and CASI Score were compared (Figure 5). A clear separation between correct hits and unknowns for both scores was observed.

The comparative performance of CASI versus NIST MS Search was evaluated using the hits ranked in the first position with their respective score values above a certain threshold. The thresholds were selected from the crossover points of the curves representing correct and incorrect proposals (825 for NIST MF and 795 for CASI Score), see Figure 5. Using these thresholds, the ability of CASI and NIST MS Search to discriminate between true and false hits was compared (Table 3). True/false positives/negatives were defined as follows: True positives were correct hits ranked first, having a score above or equal to a predefined threshold, false positives were hits from the unknown set having a

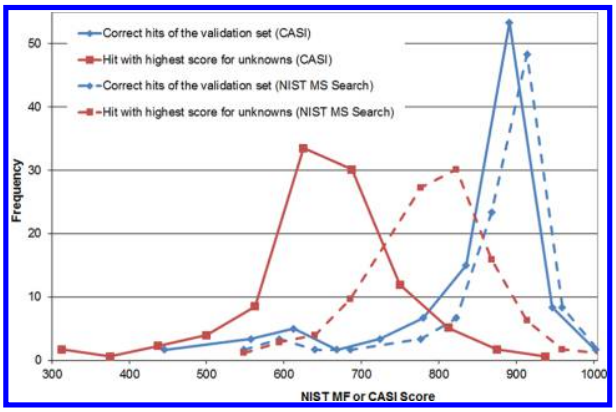


Figure 5. Comparison of hit frequency by NIST MF and CASI Score. Distribution of the NIST MF (dashed lines) and CASI Score (plain lines) for correct hits from the validation set and for hits selected by default (highest score) from a set of 176 unknown compounds.

Table 3. Comparative Power of CASI and NIST MF Scores to Discriminate between Correct and Incorrect Structural Identification

	CASI score		NIST MF	
number of hits	true	false	true	false
positive	46	11	40	29
negative	165	14	147	20
total	89%	11%	79%	21%

score above the threshold. True negatives corresponded to hits from the unknown set having a score below the threshold, and false negatives were correct hits from the validation set with a score below the threshold and hits from the validation set which did not correspond to the correct structure.

Using the CASI score resulted in 46 (77%) correct hits (true positives) being identified compared with NIST MF which delivered 40 correct hits (67%). However, when the number of false positives identified was also taken into consideration, the superior predictive power of CASI over NIST MF became apparent.

$$\text{predictive precision rate} = \frac{\text{number of true positives}}{\text{number of true positives} + \text{number of false positives}} \quad (4)$$

Using this calculation for predictive precision rate, the CASI Score had a greatly improved predictive precision rate (46/(46 + 11) = 81%) than NIST MF (40/(40 + 29) = 58%).

dx.doi.org/10.1021/ac4011952 | Anal. Chem. 2013, 85, 11216−11224

Confirmation of Structural Proposals using Accurate Mass Measurements. The proposed structural candidates were confirmed using accurate mass measurements from GC-APCI-TOF-MS, and the measured masses were compared with the theoretical monoisotopic masses of the compounds. As this confirmation required the use of two different chromatographic systems, it was important to be able to match their respective retention indices. To test the feasibility of this approach, two scenarios were evaluated: First, a test with a sample of simple chemical composition, comprising internal standards and retention index marker compounds, and second, a test using a smoke sample from a reference cigarette representing a complex matrix, see Figure 6.

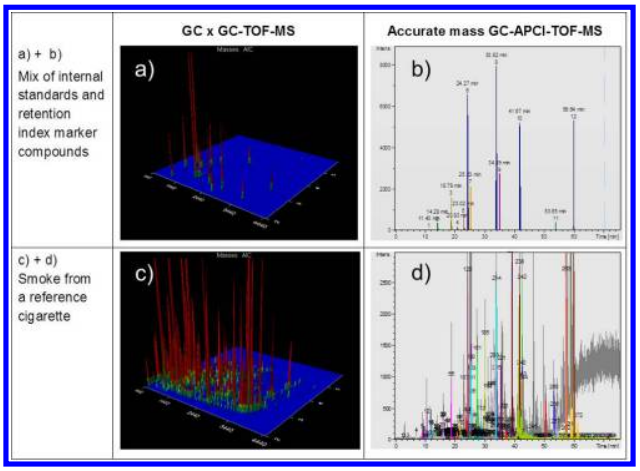


Figure 6. Total ion chromatograms of two types of sample from two systems as a confirmation of CASI performance. The confirmation strategy entailed bridging the data obtained from GC × GC-TOF-MS (Leco Pegasus IV, left part) to accurate mass GC-APCI-TOF-MS (Bruker micrOTOF-Q II, right part) using a mix of internal standards and retention index marker compounds (a, b) and a complex matrix represented by a smoke extract generated from a reference cigarette (c, d).

In the first scenario (Figure 6a, b), the deviation of the retention indices between both systems was found to be ≤1% (range 0.2%−1.0%) for all 7 internal standard compounds. The deviation between the theoretical and measured masses did not exceed 1mDa (range 0.1−1.0 mDa) for the 7 internal standard and 4 retention marker compounds using GC-APCI-TOF-MS.

In the second scenario (Figure 6c, d), a smoke extract sample generated from a reference cigarette was analyzed by accurate mass GC-APCI-TOF-MS with the focus upon 97 compounds that were confirmed by reference compounds. Out of these 97 compounds, 80 were proposed to be potentially ionizable by the APCI ionization technique as they contain heteroatoms. Seventy-three (91%) of these 80 proposed ionizable compounds were confirmed using GC-APCI-TOF-MS by means of accurate mass and retention index. The remaining 7 compounds could not be confirmed since they were not detected, most likely due to insufficient ionization.

Overall it is feasible to bridge data between GC-APCI-TOF-MS and GC × GC-EI-TOF-MS using retention indices. As the majority of tested compounds could be confirmed (by retention index and accurate mass matching) we expect to enhance the confidence level given by CASI approach by removing false positive proposals. However, a recognized limitation of this approach is that not all compounds observed under electron impact (EI) ionization mode will be visible using atmospheric pressure chemical ionization (APCI). Alternatively, chemical ionization mode could be also interesting to overcome this issue.

The impact of using accurate mass techniques upon the CASI platform prediction performance is currently being assessed in detail.

CONCLUSION

A computer-assisted structure identification platform (CASI) was designed and developed. The CASI platform accelerates and standardizes the identification of compound structures, assures reproducibility and enables scientists to have higher confidence in the correct assignment of mass spectra to the right compounds. The CASI platform automatically identifies, on-the-fly and with highest confidence, relevant structures from mass spectra associated with chromatographic data using a novel 2-dimensional relative retention time model. This makes CASI a unique automated platform, designed for high-throughput identification of compounds from complex matrices using GC-MS data.

CASI was tested on a set of 60 compounds and demonstrated superior identification performance (87%) compared to the industry standard approach using NIST MS Search (75%). Moreover, the CASI Score was shown to have a better predictive precision rate (81%) than the NIST MF (58%). In consequence, the confidence in a proposed structure by the CASI platform is enhanced compared to NIST MS Search, essentially because of the reduced number of false positive results. On the basis of the results presented, for 10 proposed structures 4 proposals would be false when using NIST MS Search, while the CASI platform could reduce this number to 2 false proposals out of 10.

Finally, a proof of concept to further increase the confidence in the compound identification by using accurate mass measurements has been performed. The next steps are to optimize the retention index system and automatic peak confirmation techniques. Furthermore the relation of data set size to informational quality is planned for evaluation. The development of a proton affinity model is also planned in order to focus any confirmation by GC-APCI-TOF-MS on potentially ionizable molecules only.

ASSOCIATED CONTENT

Supporting Information

Additional material as described in the text. This material is available free of charge via the Internet at http://pubs.acs.org.

AUTHOR INFORMATION

Corresponding Author

*Phone: +41582422527. E-mail: arno.knorr@pmi.com.

Present Addresses

Markus Stueber: Biotest AG, Dreieich, Germany
André Stratmann: A&M StabTest GmbH, Bergheim, Germany

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

The authors thank their colleagues Philippe Guy and Mark Bentley for review and valuable comments. This work is covered by an international patent, Application No. EP12717751, Publication No. WO2012146787.

REFERENCES

(1) Tikunov, Y.; Lommen, A.; de Vos, R.; Verhoeven, H. A.; Bino, R. J.; Hall, R. D.; Bovy, A. G. *Plant Physiol.* **2005**, *139*, 1125−1137, www.plantphysiol.org.

(2) Welthagen, R.; Shellie, A.; Spranger, J.; Ristow, M.; Zimmermann, R.; Fiehn, O. *Metabolomics* **2005**, *1*, 1.

(3) Kell, D. B. *Curr.Opin. Microbiol.* **2004**, *7*, 296−307.

(4) Perfetti, T. A.; Rodgman, A. *The Chemical Components of Tobacco and Tobacco Smoke*; CRC Press: Boca Raton, FL, 2008.

(5) Venkatramani, C. J.; Phillips, J. B. *J. Microcolumn* **1993**, *5*, 511−516.

(6) Kovats, E. *Helv. Chim. Acta* **1958**, *41* (7), 1915−1932.

(7) Mihaleva, V. V.; Verhoeven, H. A.; de Vos, R. C.; Hall, R. D.; van Ham, R. C. *Bioinformatics* **2009**, *25* (6), 787−94.

(8) Garkani-Nejad, Z.; Karlovits, M.; Demuth, W.; Stimpfl, T.; Vycudilik, W.; Jalali-Heravi, M.; Varmuza, K. *J. Chromatogr. A* **2004**, *1028* (2), 287−95.

(9) Seeley, J. V.; Seeley, S. K. *J. Chromatogr. A* **2007**, *1172* (1), 72−83.

(10) Baumann, F.; Straus, A. E.; et al. *J. Chromatogr. A* **1965**, *20* (0), 1−8.

(11) Soják, L.; Kruťcík, J.; et al. *J. Chromatogr. A* **1972**, *65* (1), 93−102.

(12) Donnelly, J. R.; Abdel-Hamid, M. S.; et al. *J. Chromatogr. A* **1993**, *642* (1−2), 409−415.

(13) Stein, S. E.; Scott, D.R. J. *Am. Soc. Mass Spectrom.* **1994**, *5*, 859−866.

(14) ACD/PhysChem Batch v. 12, ACD/Labs, http://www.acdlabs.com.

(15) LECO, http://www.leco.com/products/sep_sci/pegasus_4d/Pegasus4D.html.

(16) BRUKER, http://www.bruker.com/de/products/mass-spectrometry-and-separations/lc-ms/o-tof/microtof-q/overview.html.

(17) LECO ChromaTOF vers. 3.34, http://www.leco.com/products/sep_sci/chromaTOF/chromaTOF.html.

(18) Pipeline Pilot 8.0.1, Accelrys, Inc., http://accelrys.com.

(19) Dragon, Talete SRL. http://www.talete.mi.it/products/dragon_description.htm.

(20) RapidMiner and RapidAnalytics, Rapid-I GmbH, http://rapid-i.com.

(21) Eckel, W. P.; Kind, T. *Anal. Chim. Acta* **2003**, *494*, 235−243.

(22) Martin, E.; Monge, A.; Duret, J.; Pospisil, P. *J. Cheminf.* **2012**, *4*, 11.

(23) Pubchem, http://pubchem.ncbi.nlm.nih.gov/.

(24) ChemSpider, Royal Society of Chemistry, http://www.chemspider.com.

(25) ACD/Labs Name-to-Structure Batch, ACD/Labs, http://www.acdlabs.com.

(26) Rogers, D.; Hahn, M. *J. Chem. Inf. Model.* **2010**, *50*, 742−54.