

# Retention Index Prediction Combined with *in silico* Fragmentation Spectra Comparison for Increasing Confidence in Structural Elucidation Using Non-targeted Gas Chromatography Coupled with High Resolution Mass Spectrometry

P.A. Guy, E. Dossin, E. Martin, P. Diana, P. Pospisil, M. Bentley

Complex Matrix Analysis Group, Philip Morris International R&D (part of Philip Morris International group of companies), 2000 Neuchatel, Switzerland

## Introduction and Objectives

- Unambiguous chemical characterization still remains a major hurdle for analytical chemists when performing non-targeted analyses, despite significant improvements in chromatographic separation techniques and mass spectrometric instrumentation over the last decade
- A total of 552 reference compounds, including odd n-alkanes as chemical markers (n=5 to 19), were analyzed and experimental linear retention index (LRI) values were determined (a Personal Compound Database accurate mass Library was created)
- Using these experimentally determined LRI values, RapidMiner (combined with Dragon software) and ACD/ChromGenius software were used to create two independent computational Quantitative Structure-Property Relationship (QSPR) models for the prediction of LRI values
- In parallel, targeted MS/MS spectra of a smoke sample were acquired using positive chemical ionization (GC-HR-PCI-MS/MS)
- In silico* fragmentation software MetFrag and Molecular Structure Correlator, connected to the ChemSpider database, were evaluated, and predicted LRI values for compound hit proposals generated by these *in silico* fragmentation software were calculated
- Compounds identified using NIST14 MS Search, both with and without a molecular formula constraint, were compared with proposals resulting from *in silico* fragmentation software in conjunction with (or not) predicted LRI values calculated using both RapidMiner and ACD/ChromGenius software

## Methods

**Chemicals:** All reference compounds were solubilized in appropriate solvents prior to analysis by gas chromatography with high resolution mass spectrometry (GC-HR-MS), either as mixtures or as single compounds in solution. **GC Conditions:** Separation was achieved using an Agilent 7890A instrument equipped with a J&W DB-624 ultra-inert column (30 m x 0.25 mm, 1.4 μm). The column oven was maintained at 35°C for 2 min before being ramped to 250°C at a constant rate of 10°C/min. The transfer line was set at 260°C and a constant nitrogen flow rate of 1.8 mL/min was used throughout. **MS Conditions:** Detection was carried out using a 7200A Q-TOF accurate mass spectrometer system (Agilent Technologies, Santa Clara, CA). Temperature of the ion source and the emission current were set at 230°C and 35 μA, respectively. Mass spectrometric data were acquired in full scan mode by scanning *m/z* from 22 to 500 using positive electron (+EI), and positive chemical (PCI) ionization modes.

### Retention Index Modeling:

**1) RapidMiner (RM).** All reference compound structures were drawn using Accelrys Draw 4.1. 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. 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), a population size of 100, and maximum generation of 5,000. The Pareto algorithm (NSGA-II) was used as a scoring method with adjusted R-square. Multilinear regression (MLR) with 20 major descriptors was used for the final optimized prediction model.

**2) ACD/ChromGenius (CG).** The same training and test sets were used to optimize the LRI prediction model using ACD/ChromGenius Batch software (version 2014, ACD/Labs, Toronto, CA). In this case, the prediction was based upon calculated physicochemical parameters and structural similarity with known retention index values contained within a knowledge base. The calculated physicochemical parameters used were boiling point (BP), logP, polar surface area (PSA), molecular volume (MV), molecular weight (MW), molar refractivity (MR), number of hydrogen donors (ND) and number of hydrogen acceptors (NA).

## Results & Discussion

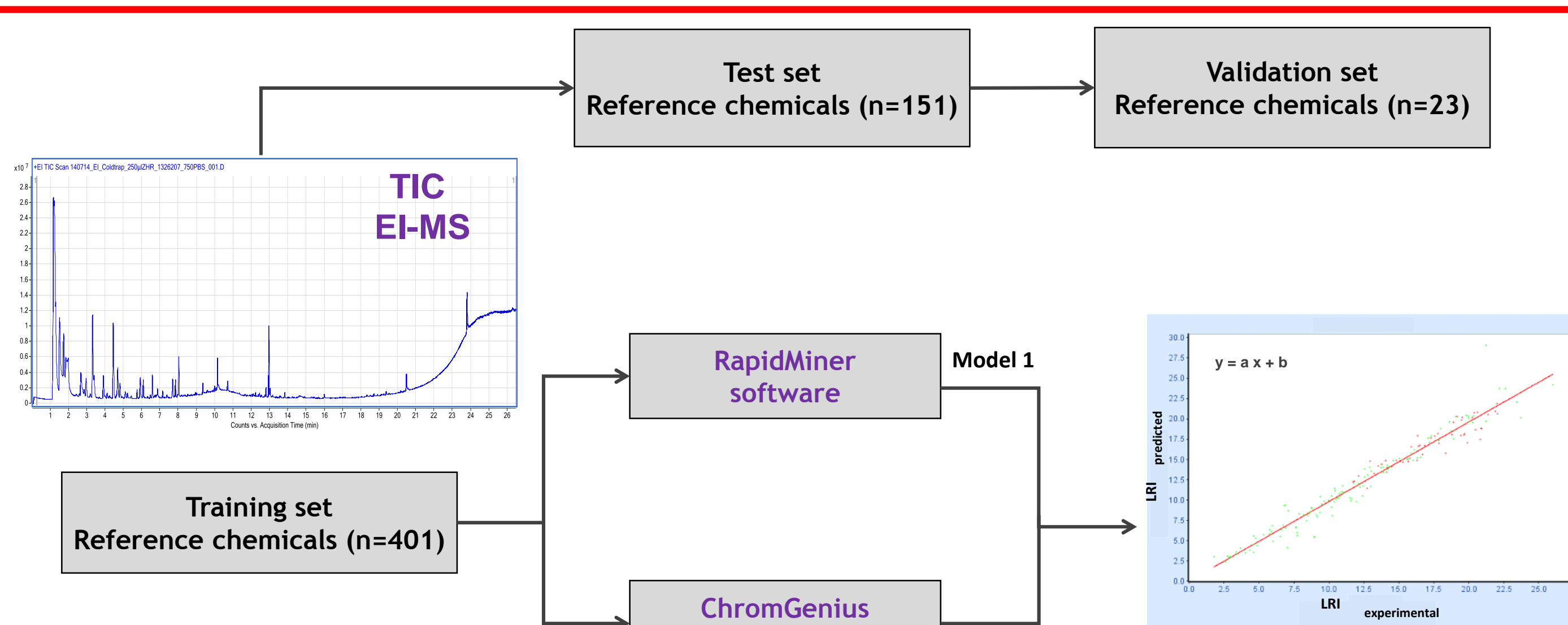


Figure 1. Workflow used to build and validate the retention index prediction models. Correlation coefficients for experimental versus predicted LRI values calculated for test set compounds were 0.949 and 0.976 using RM and CG software, respectively

A cross-validation correlation for RM was calculated with a square correlation Q2 of 0.96 and the residual standard error value obtained from CG was 53.6 (Figure 2).

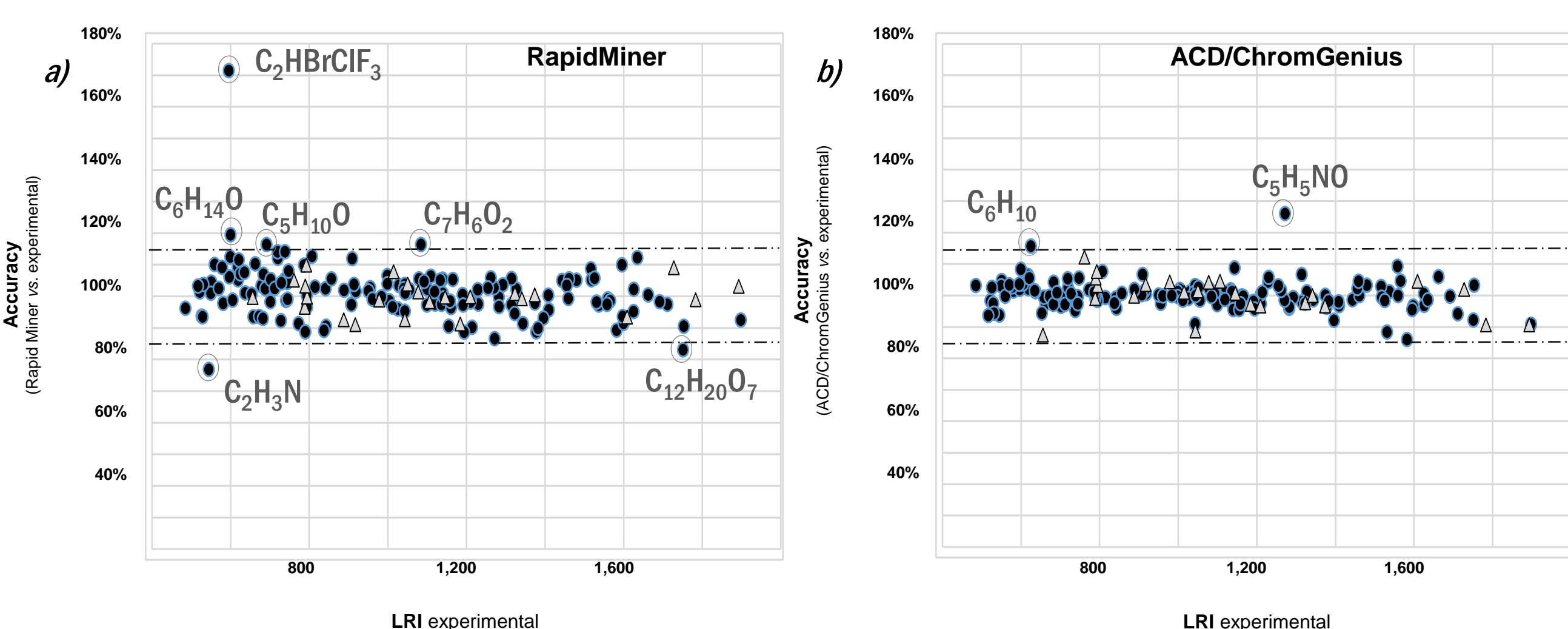


Figure 2. Accuracy data for predicted LRI versus experimental data for a) RM and b) CG for the reference compounds used in the test set (n=151) and validation set (n=23, triangles). Elemental compositions for the outlier reference compounds have been presented (arbitrarily defined as outside an accuracy value of 85-115%)

Figures 3a-b represent deconvoluted chromatograms for a complex and solvent blank samples.

As illustrated in Figure 3a, the component highlighted in blue was not registered in our accurate mass library (currently contains 669 reference standards).

MassHunter Unknown Analysis software highlighted several deconvoluted ions originating from the same component (Figure 3c) and a NIST14 MS library search gave several compound hit proposals (Figures 3d-e).

GC-HR-MS analysis performed in positive chemical ionization acquisition mode determined an empirical formula  $C_{11}H_{14}N_2O$  for this unknown constituent, which is in agreement with the first compound hit proposed from NIST14 (Figure 3e and Figure 4).

Targeted MS/MS spectra were acquired for *m/z* 191.1184 (Figure 5). Fragment ions with intensities above 10% of the most intense peak (i.e. n=6) were used to evaluate two *in silico* fragmentation software, both of which were connected to the ChemSpider database (MetFrag and Molecular Structure Correlator).

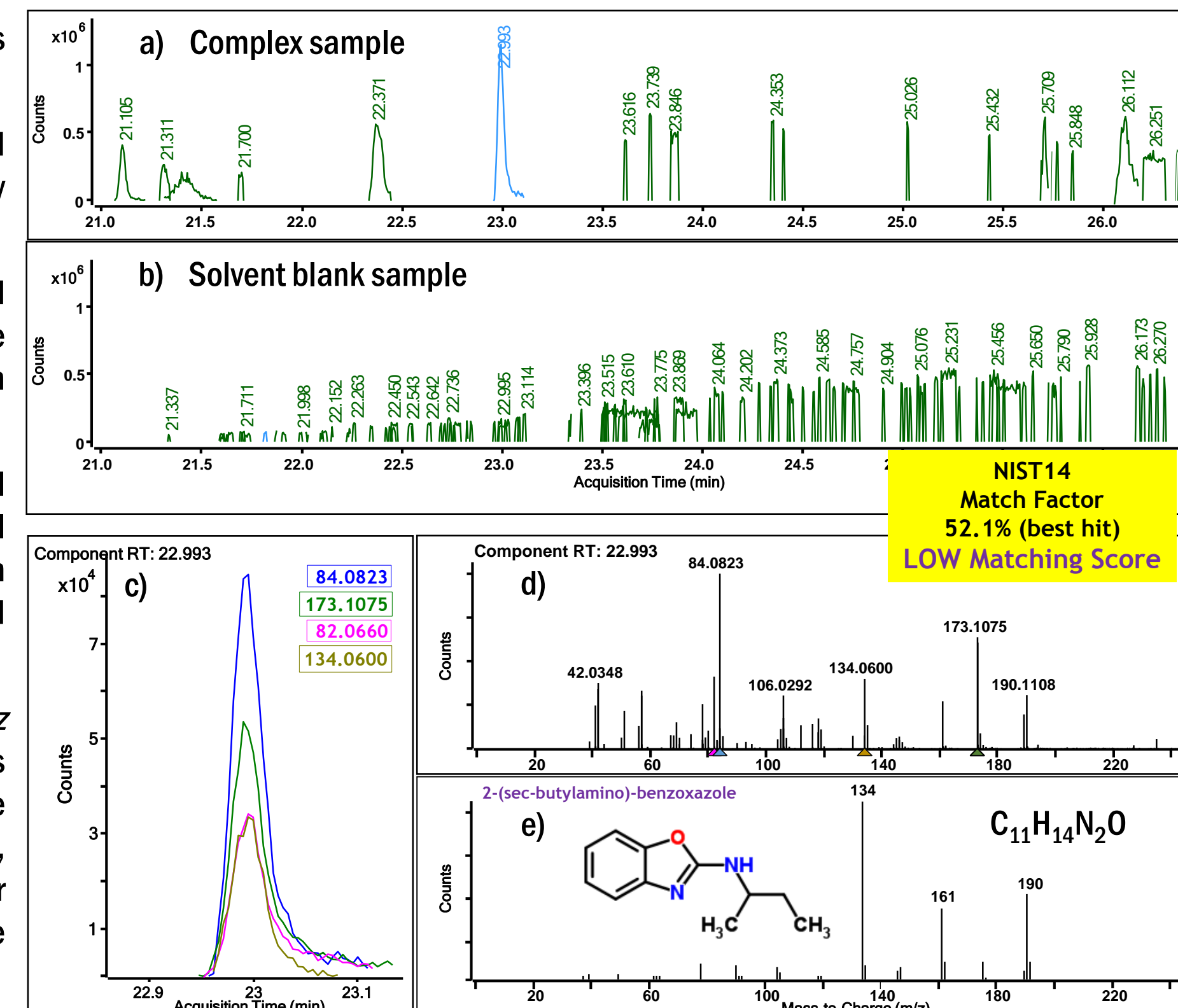


Figure 3 Deconvoluted chromatograms of a trapped smoke (a) and blank (b) samples. Overlaid EICs of a component (c) with corresponding EI accurate mass spectrum (d), and associated NIST14 library search first hit proposal (e).

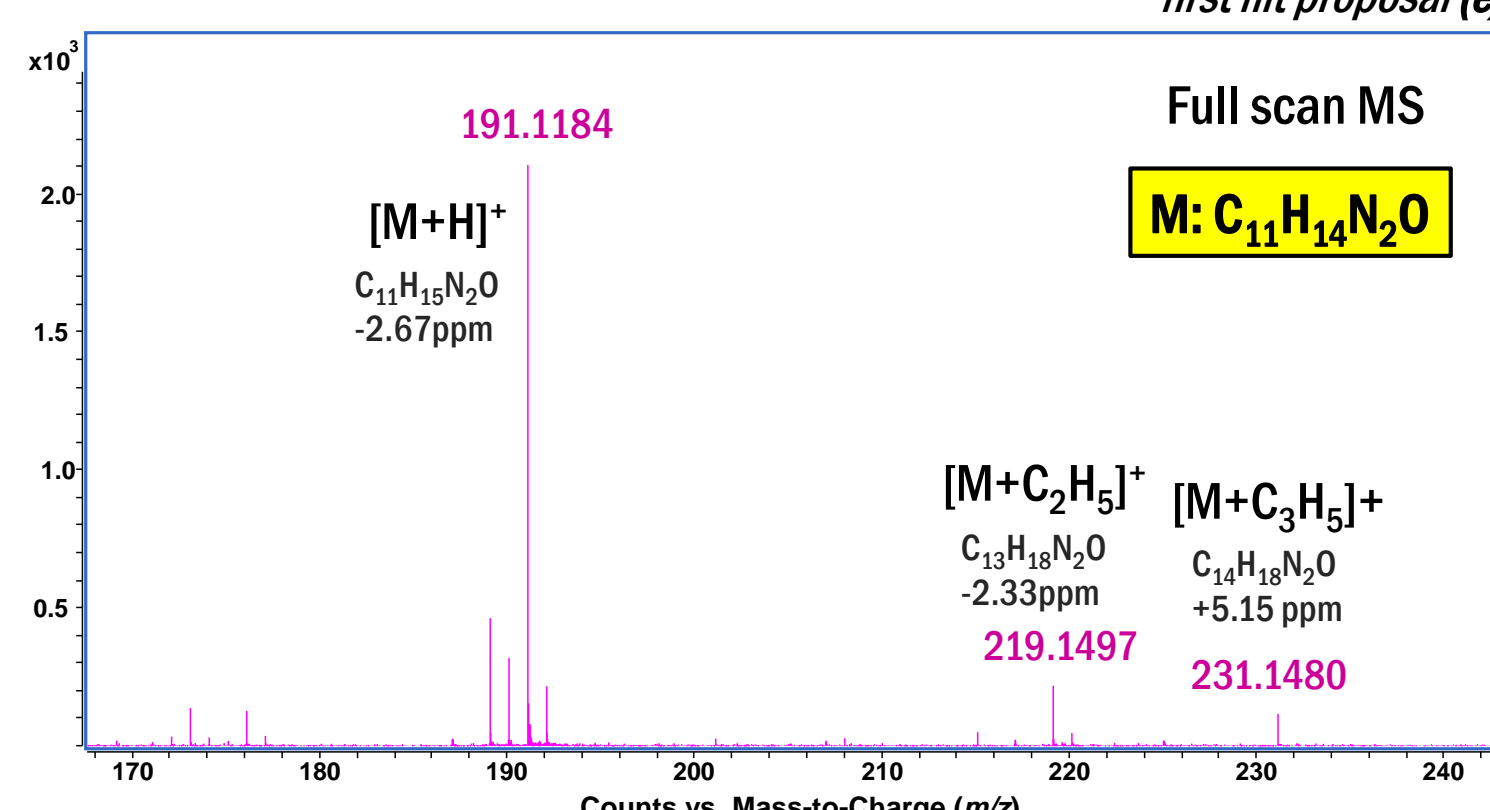


Figure 4: Subtracted PCI accurate mass spectrum obtained at RT 22.993 min. Protonated and specific adduct ion species are highlighted.

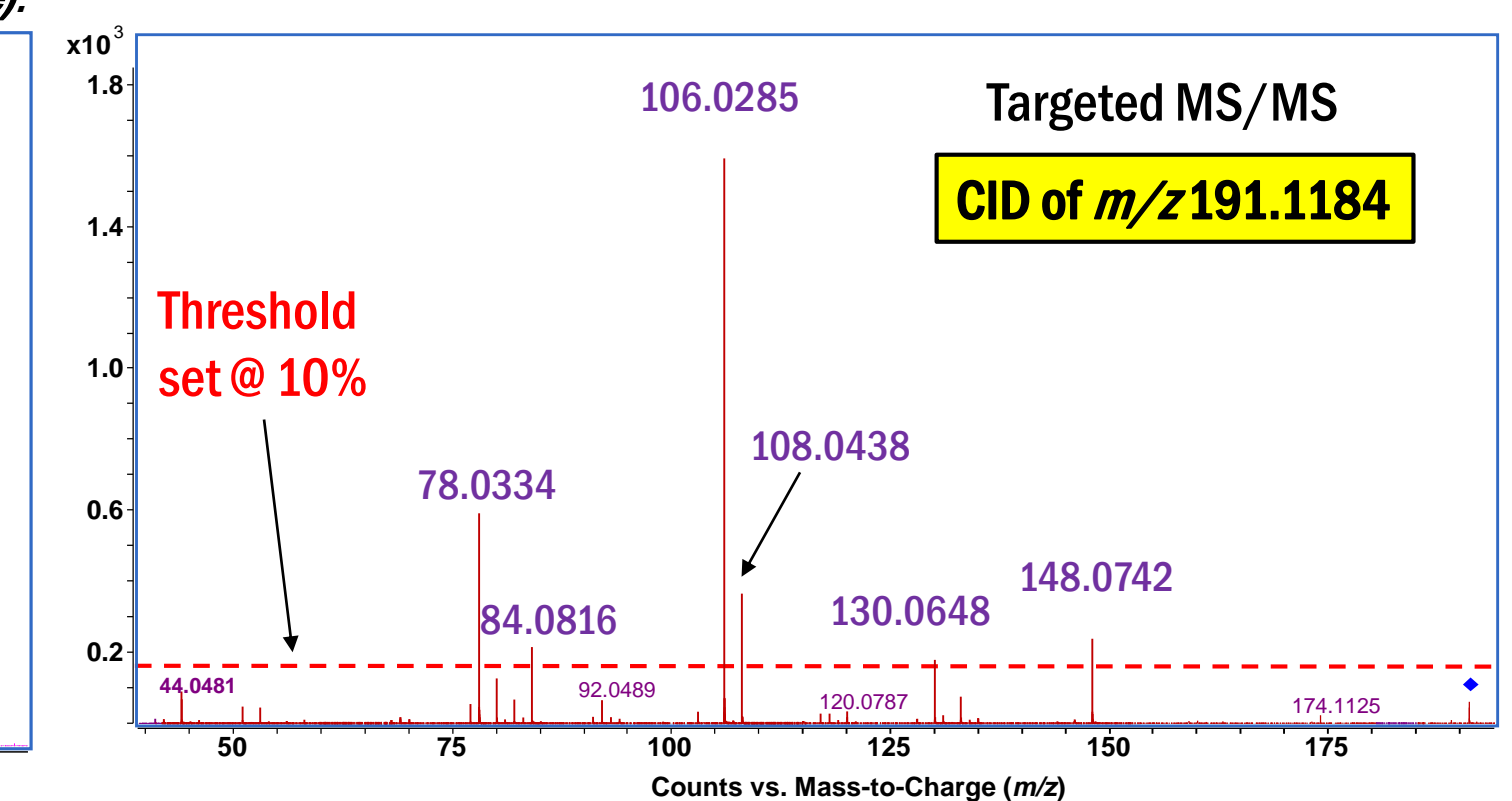


Figure 5: PCI-MS/MS spectrum of 191.1184 (RT 22.993 min) with a collision energy set at 25 eV. Fragment ions selected for *in silico* fragmentation software assessment are highlighted in bold (i.e. above 10% of major ion).

Amongst all the compounds registered in ChemSpider database, 3'651 matched the elemental formula of  $C_{11}H_{14}N_2O$ . (R,S)-1-Methyl-3-nicotinylpyrrolidine was confirmed (matching LRI and EI mass spectrum). This compound was ranked by MetFrag and Molecular Structure Correlator software as the 5<sup>th</sup> and 43<sup>rd</sup> hit, respectively. LRI values were subsequently predicted for all compound proposals made by MetFrag using both RM and CG models.

MetFrag Only					MetFrag + LRI prediction						
PNG_Image	Comment	ChemSpider ID	Mass	MetFrag Score	Rank	PNG_Image	LRI_pred CG	LRI_pred RM	LRI_exp	MetFrag & LRI pred. SCORE	Rank
	unspecified stereochem.	1221410 (2Z form) 1221411 (2E form) 4603758 (2E form)	190.1106	1.0000	1 <sup>st</sup>		1701.97 $\Delta$ LRI=-82.3	1763.39 $\Delta$ LRI=-20.9	1'784	0.930	1 <sup>st</sup>
	unspecified stereochem.	2045246	190.1106	1.0000	2 <sup>nd</sup>		1793.5298 $\Delta$ LRI=+9.3	1898.80 $\Delta$ LRI=+114.55	1'784	0.920	2 <sup>nd</sup>
		1259330	190.1106	0.9860	3 <sup>rd</sup>		1811.87 $\Delta$ LRI=+27.6	1891.95 $\Delta$ LRI=+107.7	1'784	0.916	3 <sup>rd</sup>
		1256481	190.1106	0.9860	4 <sup>th</sup>		1820.33 $\Delta$ LRI=+36.1	1893.98 $\Delta$ LRI=+109.7	1'784	0.910	4 <sup>th</sup>
		3716473	190.1106	0.9840	5 <sup>th</sup>		1637.96 $\Delta$ LRI=-146.3	1702.82 $\Delta$ LRI=-81.4	1'784	0.884	5 <sup>th</sup>
		963178	190.1106	0.9840	6 <sup>th</sup>		1634.80 $\Delta$ LRI=-149.5	1699.52 $\Delta$ LRI=-84.7	1'784	0.881	6 <sup>th</sup>

The combination of both MetFrag data and predicted LRI values (RM & CG) enabled the correct compound to be ranked as first hit compared to MetFrag alone (5<sup>th</sup>). The following table reports the ranking scores obtained from additional compounds investigated (NIST with and without formulae constraint, MetFrag or Molecular Structure Correlator in combination (or not) with predicted LRI values.

TRUE COMPOUND	(R,S)-1-methyl-3-nicotinylpyrrolidine	2,3-pentanedione	2-pentanone	3-penten-2-one
Formula	$C_{11}H_{14}N_2O$	$C_5H_8O_2$	$C_5H_{10}O$	$C_5H_8O$
RANKING NIST14 nominal classical search	not registered	Not present in hit list	1 <sup>st</sup>	Not present in hit list
RANKING NIST14 with formula constraint	-	2 <sup>nd</sup>	1 <sup>st</sup>	Not present in hit list
# Cpds NIST14	38	50	55	34
# Cpds ChemSpider	3,651	243	125	120
# of Fragment ions (above 10%)	6	3	4	7
RANKING MetFrag	5 <sup>th</sup> ranking	15 <sup>th</sup> ranking	17 <sup>th</sup> ranking	12 <sup>th</sup> ranking
RANKING MSC	43 <sup>th</sup> ranking	34 <sup>th</sup> ranking	6 <sup>th</sup> ranking	15 <sup>th</sup> ranking
LRI expt	1'783	738	730	792
LRI (RM)	1763 ( $\Delta$ LRI=-20)	842 ( $\Delta$ LRI=+104)	714 ( $\Delta$ LRI=-16)	746 ( $\Delta$ LRI=-46)
LRI (CG)	1702 ( $\Delta$ LRI=-81)	771 ( $\Delta$ LRI=+33)	732 ( $\Delta$ LRI=+2)	770 ( $\Delta$ LRI=-22)
RANKING MetFrag & LRI pred.	1 <sup>st</sup>	7 <sup>th</sup>	3 <sup>rd</sup>	4 <sup>th</sup>

## Conclusions

- RapidMiner combined with Dragon software and ACD/ChromGenius software both demonstrated an excellent ability to predict LRI values, with correlation coefficients of 0.949 and 0.976 for predicted vs. experimental values calculated for the test set (n=151).
- Predicted LRI values can be calculated for any compound and close agreement between values calculated by the two models will enhance confidence in compound proposal.
- NIST14 and/or other existing MS libraries are not exhaustive and additional strategies have to be developed to reduce false positive chemical identification.
- CID MS/MS experiments in combination with *in silico* fragmentation prediction software is an elegant approach to fill this gap, but chromatographic considerations should be integrated to reduce the list of putative compound hits.
- We have demonstrated that prediction of LRI values from putative compound hits generated by MetFrag software can help in strengthening the confidence level where most of true compounds were ranked below the top 7<sup>th</sup>.
- The presented approach here should reduce the number of putative compounds requiring confirmation using reference standards, thereby leading to a reduction in total cost for ordering chemicals, reducing the time for compound identification and minimizing the rate of false positive compound identification.