Computer-Assisted Structure Identification (CASI) for high-throughput identification of small molecules using GC×GC-HRAM-TOFMS high-resolution accurate mass spectrometry

Arno Knorr, Elyette Martin, Martin Almstetter, Antonio Castellon, Pavel Pospisil, Mark Bentley

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

Overview and Objective

Compound identification is a major bottleneck for modern metabolomics approaches and high-throughput, non-targeted characterization of complex matrices.

- Computer-assisted Structure Identification (CASI)[1] accelerates and standardizes the identification of compound structures using comprehensive two-dimensional gas chromatography with time-of-flight mass spectrometry (GC×GC-TOFMS) and unit mass resolution electron ionization mass spectra.
- GC×GC coupled with accurate mass spectrometry is a promising approach for further increasing the confidence for structural proposals derived from CASI.
- The objective of the presented work is to develop and integrate such an approach into an automated workflow, key to handling the vast amount of data produced from complex samples successfully.

Methods

Smoke samples from the 3R4F reference cigarette [2] and aerosol samples from a heat-not-burn product, the Tobacco Heating System (THS) 2.2, commercialized as *IQOS*®, have been analyzed by GC×GC-TOFMS (Pegasus® IV, LECO) and GC×GC-HRAM-TOFMS (Pegasus® GC-HRT 4D, LECO)[3]. Unit mass resolved data were processed using a non-targeted workflow for in-depth chemical characterization. Structural proposals for the complete dataset were derived from CASI, considering mass spectral database matching, matching of chromatographic data to QSPR*-derived prediction models for first- and second-dimension separations and boiling point, and ranks proposals according to a scoring function[4].

A workflow utilizing accurate mass electron ionization (EI) spectra combined with the existing CASI workflow has been designed to increase the confidence for structural proposals.

- Experimentally derived accurate mass spectra are compared with the output from *in silico* fragmentation of candidate structures proposed by the CASI platform using unit resolution MS data.
- ACD/MS Fragmenter[5] software applies in-built fragmentation rules to the structural features of the proposed candidates in order to reconstitute a theoretical accurate mass spectrum.
- The reconstituted mass spectrum is then compared with the experimental accurate mass spectrum using NIST MS Search v.2.3[6].
- The resulting spectral 'FIT' is termed 'Accurate Fragmentation Score' and is linked directly to the proportion of determined fragment ions matching those predicted for the candidate structure.
- A linear combination of Accurate Fragmentation Score and $CASI_{nominal}$ Score is used to strengthen the candidate selection process and further increase the confidence for CASI-derived structural proposals.

* QSPR: Quantitative Structure-Property Relationship

Workflow

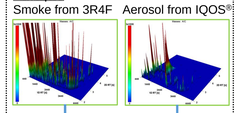
CASI (nominal mass MS)

Mass Spectra Library Search & Spectra Extraction

EI Mass Spectral Libraries comparison (NIST Search)

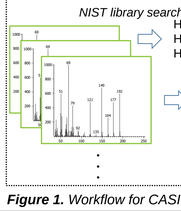
- Extract best Hits (e.g. 100)
- Extract best Hits molecular structures

Example:
Smoke from 3R4F | Aerosol from IQOS®



- Generic peak picking
- Mass spectral deconvolution

NIST library search

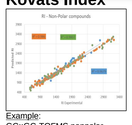


HIT 1
HIT 2
HIT 3
. . .

HIT 1
HIT 2
HIT 3
. . .

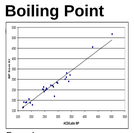
Predictive Models

Kovats Index



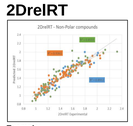
Example:
GC×GC-TOFMS nonpolar method
R^2 = 0.9807 (validation set)

Boiling Point



Example:
GC×GC-TOFMS nonpolar method
R^2 = 0.9420 (validation set)

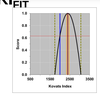
2DrelRT



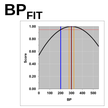
Example:
GC×GC-TOFMS nonpolar method
R^2 = 0.8555 (validation set)

Scoring Functions

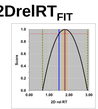
KI_{FIT}



BP_{FIT}



$2DrelRT_{FIT}$

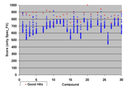


- Scoring: quadratic equations using polynomials factorization
- Optimization: all equations on-the-fly, based on result, optimization is using grid search procedure

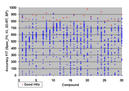
$CASI_{nominal}$ Score

$CASI_{nominal}$ = **Spectral Similarity × KI_{FIT} × BP_{FIT} × $2DrelRT_{FIT}$**

NIST Match Factor (MF)
(set of 30 example compounds)



$CASI_{nominal}$ SCORE



- Characterized in depth (JACS 2013, Knorr et al.)[1]
- Predictive precision rate: $CASI_{nominal}$: 81%
  NIST MF: 58%
- Confidence categories:
  - HIGH: ≥795
  - MEDIUM: 700 – 795
  - LOW: <700

CASI (accurate mass MS)

Extract Structures with Highest $CASI_{nominal}$ Score

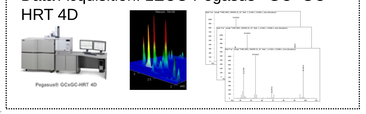
Example:
3-Acetylpyrrole

HIT No.	COMPOUND NAME	Structure	NIST MF	$CASI_{nominal}$ Score	MW	FORMULA
1	3-Acetylpyrrole (confirmed by reference standard)		941	938	109.0526	C6H7NO
2	5-Bromo-3-Methylidene-1-methoxycyclohexane		933	933	204.0148	C8H13BrO
3	Ethanone, 1-(1H-pyrrol-2-yl)-		897	879	109.0526	C6H7NO
4	3-Ethyl-4-methylpyrrole		932	855	109.0891	C7H11N
5	1-Cyano-2-methylbuten-3-one		900	838	109.0526	C6H7NO
6	Pyridine, 3-methoxy-		725	721	109.0526	C6H7NO

+

Experimental Accurate Mass Spectrum

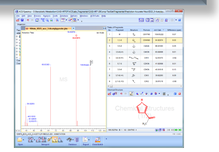
Data Acquisition: LECO Pegasus® GC×GC-HRT 4D



ACD Fragmenter Processing

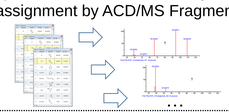
Assignment of structural features to experimental accurate mass spectrum

- Define fragmentation parameters
- Assign fragments to structural features based on ACD/Labs *in silico* fragmentation rules
- Process all candidate structures

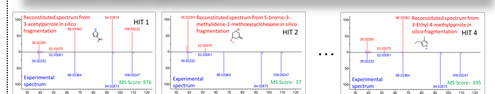


Reconstitute Mass Spectrum

Creation of reconstituted mass spectra based on *in silico* fragments assignment by ACD/MS Fragmenter



Matching Reconstituted Mass Spectra with Experimental Accurate Mass Spectrum



Final Score

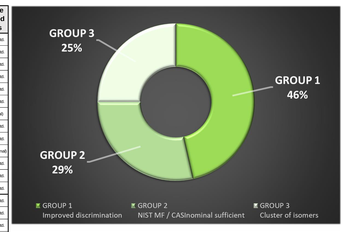
Final Score = $CASI_{nominal}$ Score × *In silico* Accurate Fragmentation Score

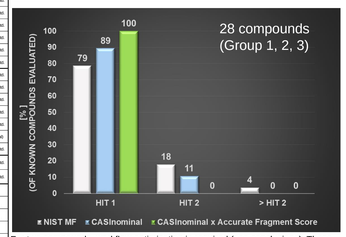
Figure 1. *Workflow for CASI using nominal mass resolution spectra.*

Figure 2. *Workflow for CASI using accurate mass resolution spectra and in silico assignment of fragments to structural features.*

Results

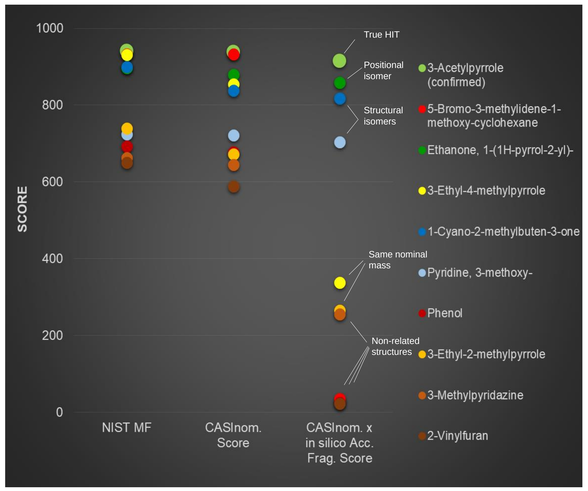
Category	Compound Name	Structure	NIST MF	$CASI_{nominal}$ Score	$CASI_{nominal}$ x in silico Frag. Score	Ranked HIT	Confidence in Proposed Structures
GROUP 1 Improved discrimination of wrong Hits	2-Furanmethanol		935	899	809	HIT 1	Confirmed by ref. std.
	5-Hydroxymethylfurfural		881	858	829	HIT 1	Confirmed by ref. std.
	Cotinine		935	864	833	HIT 1	Confirmed by ref. std.
	2-Methoxy-4-vinylphenol		911	902	863	HIT 1	Confirmed by ref. std.
	(E)-Solanone		906	894	869	HIT 1	Confirmed by ref. std.
	Pyranone		904	800	799	HIT 1	Confirmed by ref. std.
	2,3-Pentanedione		930	835	807	HIT 1	HIGH (CASInominal)
	Methyl 2-furoate		923	843	538	HIT 1	Confirmed by ref. std.
	3-Acetylpyrrole		941	938	915	HIT 1	Confirmed by ref. std.
	3-Hydroxy-β-damascone		796	752	704	HIT 1	MEDIUM (CASInominal)
	Phenylacetaldehyde		906	882	851	HIT 1	Confirmed by ref. std.
	Methyl pyruvate		952	778	735	HIT 1	Confirmed by ref. std.
	2-Ethylpyrazine		832	798	754	HIT 1	Confirmed by ref. std.
GROUP 2 No additional benefit compared to NIST MF / CASI(nominal)	Vitamin E		875	874	777	HIT 1	Confirmed by ref. std.
	3-Phytol		915	906	903	HIT 1	Confirmed by ref. std.
	Butanoic acid, 2-methyl-		873	848	846	HIT 1	Confirmed by ref. std.
	2-Methylpyrazine		962	927	891	HIT 1	Confirmed by ref. std.
	Farnesyl acetone		868	796	781	HIT 1	Confirmed by ref. std.
	Neophytadiene		912	909	897	HIT 1	Confirmed by ref. std.
	Palmitic acid		892	889	876	HIT 1	Confirmed by ref. std.
	Eicosane		890	889	871	HIT 1	Confirmed by ref. std.
GROUP 3 Cluster of structural closely related compounds (isomers)	Pyridine, 3-methyl-		950	937	903	HIT 1	Confirmed by ref. std.
	p-Cresol		933	932	695	HIT 1	Confirmed by ref. std.
	3-Pyridinol		832	788	743	HIT 1	Confirmed by ref. std.
	Scopoletin		900	842	758	HIT 1	Confirmed by ref. std.
	Megastigmatrienone		799	798	775	HIT 1	Confirmed by ref. std.
	3-Vinylpyridine		925	905	856	HIT 1	HIGH (CASInominal)
	Cyclotene		948	935	906	HIT 1	Confirmed by ref. std.
Problem Cases (=>workflow optimization)	Stigmasterol		897	896	433	HIT 2	Confirmed by ref. std.
	trans-Caryophyllene		High number of isomers, correct structure not ranked into TOP 10 by CASI (nominal) => not submitted for further processing.				Confirmed by ref. std.
"NOT IDENTIFIED" Re-evaluation Not Identified	ID-UNK-01_033_Unknown-004		689	644	625		Not Identified (CASInominal)
	ID-UNK-04_191_Unknown-044		735	724	586		Not Identified (CASInominal)
	ID-UNK-05_040_Unknown-006		747	713	612		Not Identified (Scientist review)
	ID-UNK-07_117_Unknown-023		714	695	579		Not Identified (CASInominal)
	ID-UNK-02_118_Unknown-022		558	511	76		Not Identified (CASInominal)
	ID-UNK-06_063_Unknown-009		674	673	462		Not Identified (CASInominal)





For two compounds, workflow optimization is required (see conclusions). These compounds were therefore excluded from further evaluation.

Figure 3. *Test Set Results – ability for proposing the correct structure using either NIST match factor, $CASI_{nominal}$ Score, or linear combination of $CASI_{nominal}$ Score together with in silico Accurate Fragmentation Score.*



CASI HIT	Compound Name	Structure	NIST MF	$CASI_{nom.}$ Score	MW	Formula	in silico Acc. Frag. Score	$CASI_{nom.}$ x in silico Acc. Frag. Score
1	3-Acetylpyrrole (confirmed)		941	938	109.05276	C6H7NO	976	915
2	5-Bromo-3-methylidene-1-methoxy-cyclohexane		933	933	204.01488	C8H13BrO	37	35
3	Ethanone, 1-(1H-pyrrol-2-yl)-		897	879	109.05276	C6H7NO	976	858
4	3-Ethyl-4-methylpyrrole		932	855	109.08915	C7H11N	395	338
5	1-Cyano-2-methylbuten-3-one		900	838	109.05276	C6H7NO	976	818
6	Pyridine, 3-methoxy-		725	721	109.05276	C6H7NO	976	704
7	Phenol		693	677	94.04186	C6H6O	37	25
8	3-Ethyl-2-methylpyrrole		739	672	109.08915	C7H11N	395	266
9	3-Methylpyridazine		663	645	94.05310	C5H6N2	395	255
10	2-Vinylfuran		651	589	94.04186	C6H6O	37	22

Figure 4. *Discrimination of the correct structure from other proposals using either NIST Match Factor, $CASI_{nominal}$ Score, or linear combination of $CASI_{nominal}$ Score together with in silico Accurate Fragmentation Score, example for 3-acetylpyrrole.*

Conclusions and Discussion

- Out of approximately 400 compounds found in the aerosol of THS above 100 ng/item using GC×GC-TOFMS (more than 1,800 compounds found in 3R4F, ≥ 100 ng/item), a subset of 30 compounds was selected to test the performance of the enhanced structural identification workflow as a proof-of-concept stage.
- Diversity and complexity of structures and specificity of mass spectra were considered as inclusion criteria.
- For 28 compounds, the workflow demonstrated excellent results by ranking the correct structure as HIT 1 for all compounds in the test set, which further improves the $CASI_{nominal}$ approach:
  - **Accurate Fragmentation Score (100%), NIST MF (79%), $CASI_{nominal}$ Score (89%)**
  - For 46% of evaluated compounds, discriminatory power increased using Accurate Fragmentation Score.
  - For 25% of evaluated compounds, clusters of isomers scored closely together (considered for reporting).
- For two compounds, the workflow requires further optimization:
  - Stigmasterol: NIST MS Search generated an inconsistently low spectral FIT (ranked as HIT 2).
  - Trans-caryophyllene: correct structure not ranked within the top 10 candidates by $CASI_{nominal}$ Score and therefore not submitted for further processing (constraint for proof-of-concept stage).
- A group of six compounds classified as 'not identified' were included for evaluation of false discovery rates.
  - Four compounds were considered for further evaluation due to acceptable Accurate Fragmentation Scores.
  - Two compounds had low Scores, further confirming their absence from commercial MS libraries.
- For the next stage, full automation using an integrated workflow is planned, as established for $CASI_{nominal}$.
- Evaluation of a much larger dataset, scaled for meaningful statistical evaluation of workflow performance and qualification of our enhanced structure identification platform, will be performed.

References

[1] Knorr, A., et al., Analytical chemistry 2013, 85, 11216 https://www.ncbi.nlm.nih.gov/pubmed/24160557
[2] University of Kentucky [Internet]. Specifications of 3R4F, https://refcig.uky.edu/resources/pdf/webdocs/3R4F%20Preliminary%20Analysis.pdf
[3] LECO Corporation, https://www.leco.com/index.php/products/separation-science
[4] Knorr, A.,et al.,Intern. Patent WO 2012146787 A1, PCT/EP2012/057942, 2012
[5] ACD/MS Fragmenter, version 14.01, Advanced Chemistry Development, Inc., Toronto, On, Canada, www.acdlabs.com, 201X.
[6] NIST MS Search v.2.3, https://chemdata.nist.gov/

PMI SCIENCE
PHILIP MORRIS INTERNATIONAL

U.S. EPA: 2018 Non-Targeted Analysis Collaborative Research Trial (ENTACT) Workshop

August 13-15, 2018

Competing Financial Interest

Philip Morris International is the sole source of funding and sponsor of this project.