Retention Index Prediction Combined with *In Silico* Fragmentation Spectra Comparisons for Increasing Confidence in Structural Elucidation using Non-Targeted Gas Chromatography coupled with High Resolution Mass Spectrometry

“NonTarget 2016 Conference, Ascona, May, 30th 2016”

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

*Philip Morris International R&D*
Outline

- Generation of aerosol sample / chemical complexity / GC-HR-MS analysis
- Building linear retention index (LRI) prediction models
  - RapidMiner - Dragon software (RM)
  - ACD/ChromGenius software (CG)
  - LRI modeling assessment & usage to characterize aerosol constituents (library database)
- Non-targeted screening workflow for aerosol characterization
- Case studies
- Conclusion and next steps
PMI Science

• PMI is working on various Reduced Risk Products (RRP) delivering nicotine containing aerosols.

• In this context, it is important to fully characterize the chemical composition of RRP aerosols in comparison to smoke produced from cigarettes.

• For analytical method development purposes we use a reference cigarette (3R4F).
Generation of Smoke Samples from a Reference Cigarette

• Reference cigarette: 3R4F*
• Smoking regimen: Health Canada
  - 2 sticks accumulation
  - Puff volume: 55 mL
  - Puff duration: 2 sec
  - Frequency: 2 puffs / min
  - Puff count (butt length)

* University of Kentucky (Kentucky Tobacco R&D Center). http://www2.ca.uky.edu/refcig/

- Cambridge filter is combined with the impingers
- Addition of retention index chemical markers (n-alkanes) & isotopically labeled internal standards

Total Particulate Matter (TPM) → Cambridge filter is extracted → 2 cold impingers in series → Gas Vapor Phase (GVP) → Whole smoke
Unique Compounds & Spectra Database (UCSD)

11,567 molecules are registered in our in-house database:
- Over 7,000 chemicals reported as present in tobacco and tobacco smoke¹
- Over 3,000 molecules associated with flavor properties²-³

1,013 (+EI) accurate mass spectra

- Hydrocarbon (n=1’081)
- Oxygen-containing functions (n=7’427)
- Nitrogen-containing functions (n=1’788)
- Nitrogen heterocyclic functions (n=2’715)
- Sulfur-containing functions (n=897)
- Miscellaneous functions (n=1195)

³ EFSA flavoring substances database.

Goal is to screen the broadest range of smoke constituents in a “non-targeted screening” approach.
Building Linear Retention Index Models using QSPR

**Model 1**

**Structural descriptors**

- Experimental LRI data (n=552 ref. stds) DB-624 GC column
- Rapid Miner software
- Model optimization
- ACD/ChromGenius software

**Training set**
Reference chemicals (n=401, 2/3)

**Test set**
Reference chemicals (n=151, 1/3)

**Model 2**

**Compound similarities**

- TIC
- EI-MS
- Validation set
Reference chemicals (n=23)

- Model Assessment

- **Model Optimization**
  - \( y = ax + b \)
  - LRI\(_{predicted}\)
  - LRI\(_{experimental}\)
Accuracy Data for Predicted versus Experimental LRI Values

Accuracy (ACD/ChromGenius vs. experimental)

1,600
1,200
800
LRI experimental

180%
160%
140%
120%
100%
80%
60%
40%

Accuracy (RapidMiner vs. experimental)

1,600
1,200
800
LRI experimental

180%
160%
140%
120%
100%
80%
60%
40%

n=151 reference standards (Test set)
△ n=23 reference standards (Validation set)

r² = 0.949
Q² = 0.96

r² = 0.976
Residual std error = 53

Submitted in peer-reviewed Journal
LRI Prediction for the Complete UCSD Compound Library

- 6,053 molecules were predicted with LRI values between 500 - 1,900 (targeted for DB-624 GC column)
- 3,646 molecules (60%) have an EI Mass Spectra (NIST or Wiley)
- LRI values can be predicted from any compound databases
Non-targeted Screening Workflow for Aerosol Characterization

Data Acquisition

• CI full scan MS
• Targeted MS/MS

Data Acquisition (+EI)

Deconvolution

MassHunter Unknown Analysis Software

Identification

• Stop when found:
  → Fingerprinting_DB-624.xml
    (n~700 EI accurate mass & LRI_{expt.})
  → UCSD Library (n=3,646 EI nominal mass & predicted LRI)

Final Matching score

Low to MEDIUM

Final matching score

HIGH

Final Reporting

• Purchase of ref. std. if available
• Compound confirmation for new ones

Final Data review

Final report

Smoke sample

MetFrag (or MSC) in silico fragmentation software

LRI prediction of proposal hits
• Final ranking score
## Case Study 1: Compound Identification with Accurate Mass Library

Easy compound confirmation if reference standard is already present within our Personal Compound Database accurate mass Library (PCDL, n~700)
Case Study 2: Problematic Hit Proposals

There is a need to develop alternative approaches when compounds are not registered in existing MS libraries.
Case Study 2:
GC-HR-MS in Chemical Ionization Mode & MS/MS

GC-HR-MS (Full Scan MS)
Positive Chemical Ionization (PCI)

**Determination of elemental formula**
(adduct ion species)

\[ M: C_{11}H_{14}N_{2}O \]

GC-HR-MS (Full Scan MS/MS)
PCI data acquisition CID of 191.1184

**MS/MS data processed using a larger chemical database**
with *in silico* predicted fragmentation software

Ion threshold above 10% (n=6 ions)
**In Silico Theoretical Fragmentation Software Evaluation: MetFrag**

1) LRI values were predicted for all 100 proposals
2) Final ranking SCORE was calculated using:
   - MetFrag Score
   - LRI \text{expt.} Against LRI \_RM
   - LRI \text{expt.} Against LRI \_CG ...

3,932 hits! Search performed May 5\textsuperscript{th} 2016

4,048 hits! Search performed May 19\textsuperscript{th} 2016
In Silico Theoretical Fragmentation Software Evaluation: Molecular Structure Correlator (MSC)

1) MS/MS accurate mass spectra exported as .cef files
2) Open in MSC software
3) Several databases are available

- Elucidation of Product Ion Connectivity (EPIC) based-approach
- Systematic bond cleavages with a score penalty function

Calculated Elemental formula

Fragment ions interpretation

List of putative compounds

Different chemical databases available for search

$m/z$ fragment ions & intensities imported values

True compound was ranked in 43rd position

**Assessment for MetFrag *In Silico* Fragmentation**

<table>
<thead>
<tr>
<th>PNG Image</th>
<th>Comment</th>
<th>ChemSpider ID</th>
<th>Mass</th>
<th>MetFrag Score</th>
<th>Rank</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image1.png" alt="" /></td>
<td>unspecified stereochem.</td>
<td>1221410 (2z &amp; 2E) 1221411 (2Z form) 4603758 (2E form)</td>
<td>190.1106</td>
<td>1.0000</td>
<td>1st</td>
</tr>
<tr>
<td><img src="image2.png" alt="" /></td>
<td>unspecified stereochem.</td>
<td>2045246</td>
<td>190.1106</td>
<td>1.0000</td>
<td>2nd</td>
</tr>
<tr>
<td><img src="image3.png" alt="" /></td>
<td></td>
<td>1259330</td>
<td>190.1106</td>
<td>0.9860</td>
<td>3rd</td>
</tr>
<tr>
<td><img src="image4.png" alt="" /></td>
<td></td>
<td>1256481</td>
<td>190.1106</td>
<td>0.9860</td>
<td>4th</td>
</tr>
<tr>
<td><img src="image5.png" alt="" /></td>
<td></td>
<td>3716473</td>
<td>190.1106</td>
<td>0.9840</td>
<td>5th</td>
</tr>
<tr>
<td><img src="image6.png" alt="" /></td>
<td></td>
<td>963178</td>
<td>190.1106</td>
<td>0.9840</td>
<td>6th</td>
</tr>
</tbody>
</table>

**Usefulness to combine LRI prediction with MetFrag score**

<table>
<thead>
<tr>
<th>PNG Image</th>
<th>LRI_pred CG</th>
<th>LRI_pred RM</th>
<th>LRI_exp</th>
<th>MetFrag &amp; LRI_pred. SCORE</th>
<th>Rank</th>
</tr>
</thead>
<tbody>
<tr>
<td>![image7.png]</td>
<td>1701.97</td>
<td>1763.39</td>
<td>1'784</td>
<td>0.930</td>
<td>1st</td>
</tr>
<tr>
<td>![image8.png]</td>
<td>1793.5298</td>
<td>1898.80</td>
<td>1'784</td>
<td>0.920</td>
<td>2nd</td>
</tr>
<tr>
<td>![image9.png]</td>
<td>1811.87</td>
<td>1891.95</td>
<td>1'784</td>
<td>0.916</td>
<td>3rd</td>
</tr>
<tr>
<td>![image10.png]</td>
<td>1820.33</td>
<td>1893.98</td>
<td>1'784</td>
<td>0.910</td>
<td>4th</td>
</tr>
<tr>
<td>![image11.png]</td>
<td>1637.96</td>
<td>1702.82</td>
<td>1'784</td>
<td>0.884</td>
<td>5th</td>
</tr>
<tr>
<td>![image12.png]</td>
<td>1634.80</td>
<td>1699.52</td>
<td>1'784</td>
<td>0.881</td>
<td>6th</td>
</tr>
</tbody>
</table>

5th proposal confirmed (ref. standard)

1st proposal confirmed (ref. standard)

Better discriminatory power
Interpretation of (1-Methyl-3-pyrrolidinyl)(3-pyridinyl)methanone MS/MS Spectrum Using MetFrag Software

5 out of 6 fragment ions were assigned by MetFrag software
## MetFrag vs. Molecular Structure Correlator Software

<table>
<thead>
<tr>
<th>TRUE COMPOUND</th>
<th>(R,S)-1-methyl-3-nicotinoylpyrrolidine</th>
<th>2,3-pentanedione</th>
<th>2-pentanone</th>
<th>3-penten-2-one</th>
</tr>
</thead>
<tbody>
<tr>
<td>Formula</td>
<td>C_{11}H_{14}N_{2}O</td>
<td>C_{5}H_{8}O_{2}</td>
<td>C_{5}H_{10}O</td>
<td>C_{5}H_{8}O</td>
</tr>
<tr>
<td>RANKING NIST14 nominal classical search</td>
<td>not registered</td>
<td>Not present in hit list</td>
<td>1\textsuperscript{st}</td>
<td>Not present in hit list</td>
</tr>
<tr>
<td>RANKING NIST14 with formula constraint</td>
<td>-</td>
<td>2\textsuperscript{nd}</td>
<td>1\textsuperscript{st}</td>
<td>Not present in hit list</td>
</tr>
<tr>
<td># Cpds NIST14</td>
<td>38</td>
<td>50</td>
<td>55</td>
<td>34</td>
</tr>
<tr>
<td># Cpds ChemSpider</td>
<td>3,651</td>
<td>243</td>
<td>125</td>
<td>120</td>
</tr>
<tr>
<td># of Fragment ions (above 10%)</td>
<td>6</td>
<td>3</td>
<td>4</td>
<td>7</td>
</tr>
<tr>
<td>RANKING MetFrag</td>
<td>5\textsuperscript{th} ranking</td>
<td>15\textsuperscript{th} ranking</td>
<td>17\textsuperscript{th} ranking</td>
<td>12\textsuperscript{th} ranking</td>
</tr>
<tr>
<td>RANKING MSC</td>
<td>43\textsuperscript{th} ranking</td>
<td>34\textsuperscript{th} ranking</td>
<td>6\textsuperscript{th} ranking</td>
<td>15\textsuperscript{th} ranking</td>
</tr>
<tr>
<td>LRI expt</td>
<td>1'783</td>
<td>738</td>
<td>730</td>
<td>792</td>
</tr>
<tr>
<td>LRI (RM)</td>
<td>1763 (\Delta LRI=-20)</td>
<td>842 (\Delta LRI=+104)</td>
<td>714 (\Delta LRI=-16)</td>
<td>746 (\Delta LRI=-46)</td>
</tr>
<tr>
<td>LRI (CG)</td>
<td>1702 (\Delta LRI=-81)</td>
<td>771 (\Delta LRI=+33)</td>
<td>732 (\Delta LRI=+2)</td>
<td>770 (\Delta LRI=-22)</td>
</tr>
<tr>
<td>RANKING MetFrag &amp; LRI pred.</td>
<td>1\textsuperscript{st}</td>
<td>7\textsuperscript{th}</td>
<td>3\textsuperscript{rd}</td>
<td>4\textsuperscript{th}</td>
</tr>
</tbody>
</table>
Conclusions & Next Steps

➢ Advantageous to combine state-of-the-art instrumentation with advanced chemoinformatic tools
  - LRI prediction models using both RM & CG software (algorithms) showed great results
  - Low differences between the two LRI models enhanced the confidence level for compound identification

➢ Existing MS libraries are not exhaustive and additional strategies need to be developed

➢ Targeted MS/MS combined with software to predict in silico fragmentation is mature
  - MetFrag software seems to be more reliable than Molecular Structure Correlator
  - Addition of LRI prediction values demonstrated a greater potential to correctly rank putative hits than in silico fragmentation alone
Conclusions & Next Steps (continued)

- This combined approach significantly reduces the amount of compounds purchased for absolute confirmation
  - Reducing the overall time for compound identification
  - Reducing the cost for purchasing chemicals
  - Minimizing the rate of false positive compound identification

- Complete automated data-processing has to be developed and validated in order to reduce the workload for Non-Targeted Screening applications
  - Final Ranking SCORE to be calculated on the fly (accurate mass results - LRI predictions)
  - Data fusion across volatile - semi-volatile & polar - apolar methods
Acknowledgments

Agilent Technologies
• Joerg Riener
• Tomi Hamalainen

Philip Morris International R&D*
• Complex Matrix Analysis (M. Bentley)
• Fingerprinting & Special Analysis Team
  • E. Dossin
  • P. Diana
• Computational Chemistry Team (P. Pospisil)
  • E. Martin
  • A. Castellon
• Aerosol generation staff (R. Reis Pires)

* Philip Morris Products S.A. (part of Philip Morris International group of companies)