USING COLLISION CROSS SECTION PREDICTIONS TO SUPPORT CHARACTERISATION OF UNKNOWNS

Russell Mortishire-Smith[1], Sarah Dowd[2], Mike McCullagh[1], Jeff Goshawk[1] and Johannes P.C. Vissers[1]

[1]Waters Corporation, Wilmslow SK9 4AX, UK [2]Waters Corporation, Milford, MA 01757. USA

INTRODUCTION

The identification of unknowns using LC-MS has applications across a broad range of fields from natural product characterisation, drug metabolism, extractable and leachable identification and impurity profiling. A common approach to identification is the use of high-resolution mass spectrometry to generate plausible elemental compositions, followed by library matching to identify possible matches, and subsequent product ion matching to determine which of the matches is the most likely explanation for the experimental data. We describe here the use of collision cross section (CCS) prediction as an additional characterisation endpoint, and its application to reduce false positives in the matching set.

METHODS

Experimental LC-IM-MS data was acquired on Q-IMS-oaToF and IMS-Q-oaTof geometries. Following data processing, selected peaks were subject to characterisation using UNIFI or Progenesis QI Discovery tools. Briefly, this approach involves determination of the most likely elemental composition (including isotope pattern matching) and is then used to search chemical knowledge databases for matching compounds. For each match, the software calculates a 'spectral matching %', which is the fraction of the intensity of the observed product ion spectrum which can be plausibly derived from each match. Here, we additionally predict the CCS value of each match using a machine learning model called CCSondemand and derive the % difference between the predicted and observed CCS values.

RESULTS

Here, we evaluate the ability of predicted CCS values to reduce the number of plausible matches for putative unknowns, as an adjunct to the more routinely used inputs of m/z, isotope pattern, product ion spectrum and citation counts.

Consider the case of 2-hydroxy-4-octyloxy benzophenone (octabenzone), a commonly present additive in plastics, as a hypothetical unknown (Figure 1). That is to say, experimental data for octabenzone was used as the input to the Discovery workflow, and an evaluation was made of the ability of the workflow to return octabenzone as a high scoring match. In the first step in the workflow we determine $C_{21}H_{26}O_3$ as the most plausible elemental composition for this m/z value, based on the precursor ion. This elemental composition is passed to ChemSpider which returns more than 700 putative matches which are ranked on the basis of number of citations and product ion matching (Figure 2).

Only five of these are plausible fits to the observed product ion spectrum. Two of these five are the same structure but from different entries, leaving four matches (Figure 3).

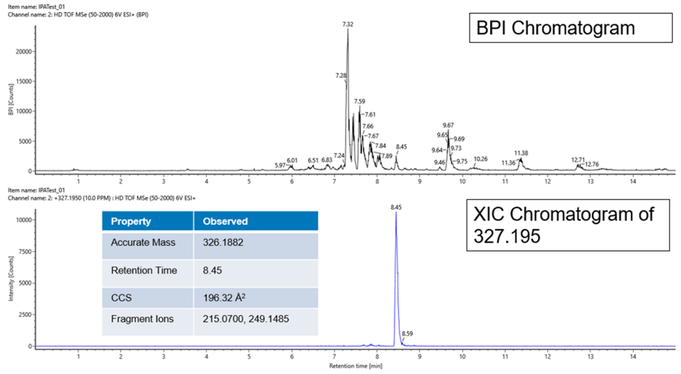

Figure 1. *Extracted ion chromatogram for a putative unknown (here, octabenzone) which was spiked in as the 'unknown') in an*

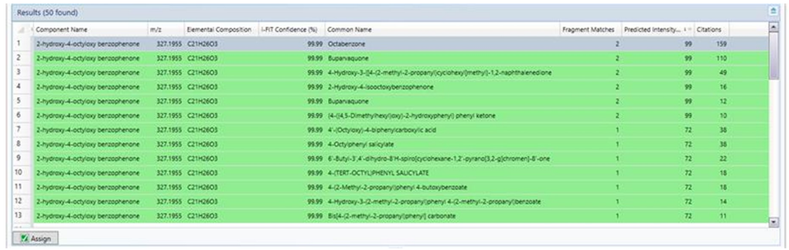

Figure 2. *UNIFI Discovery Tool outcomes.*

Common Name	Fragment Matches	Predicted Intensity (%)	Citations
Octabenzone	2	99	159
Buparvaquone	2	99	110
4-Hydroxy-3-{[4-(2-methyl-2-propanyl)cyclohexyl]methyl}-1,2-naphthalenedione	2	99	49
2-Hydroxy-4-isooctoxybenzophenone	2	99	16
(4-((4,5-Dimethylhexyl)oxy)-2-hydroxyphenyl) phenyl ketone	2	99	10

Figure 3. *Highest probability matches returned by the UNIFI.Discovery tool to search ChemSpider for a search of this m/z value. The predicted intensity denotes the % of the observed product ion spectrum which can reasonably be expected from the proposed structure.*

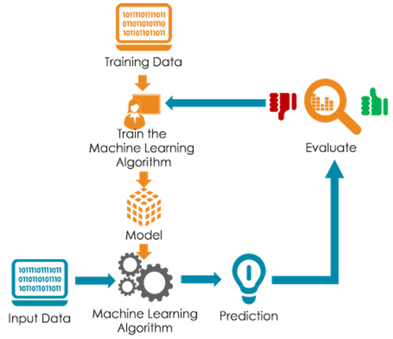

Figure 4. *Schematic of CCSondemand.*

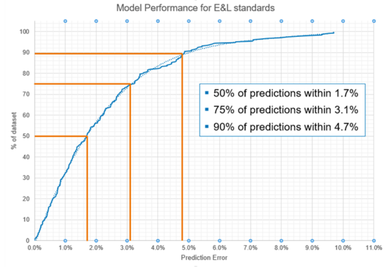

Figure 5: *Performance of CCSondemand for a set of 500 extractables and leachables which were not in the machine learning training set.*

The structures of all four were used as input to in silico CCS prediction using a platform we call CCSondemand[1,2,3] (Figure 4), which generates a machine learning model from a set of more than 5000 input datapoints and 200 two dimensional molecular descriptors. The performance of this model was evaluated by predicting CCS values for 500 extractable and leachables, none of which were used in the training model (Figure 5). Currently around 75% of predictions are within 3% of experimental values, and 90% are within 5% of experimental values

The predictions for each of the putative matches are shown in Figure 6. Of the matches, one (buparvaquone) has a predicted CCS value which is 6.7% different (Δ(CCS)) from the observed CCS value (, suggesting that this is a false positive match. Two of the four matches have intermediate

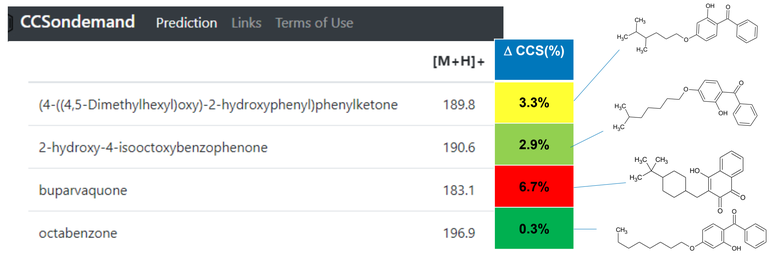

Figure 6. *CCS predictions from CCSondemand.*

Compound ID	Description	Predicted CCS	ΔCCS(A2)	Fragmentation Score
HMDB80060722	2-pyridyl acetic acid	121.90	0.47	89.9
HMDB80000875	trigonelline	123.20	-0.83	83.8
HMDB80001891	m-aminobenzoic acid	127.89	-5.52	83.9
HMDB80001392	p-aminobenzoic acid	130.37	-8.00	83.9
HMDB80001123	o-aminobenzoic acid	131.22	-8.85	83.9
HMDB80015687	salicylamide	122.94	-0.57	0

Figure 7. *Use of CCS prediction to support discrimination between putative matches for the elemental composition $C_7H_7NO_2$, after a search of HMDB...*

Δ(CCS) values of 2.9 and 3.3%, meaning that they are less plausible explanations, but which should not be ruled out. The combination of citation scoring, fragment ion matching, and D (CCS) then leads to octabenzone as being the most likely explanation for these data.

The application of this approach to a metabolomics example is summarised in Figure 7, in which a representative biomarker was characterised by LC/IMS/MS on the VION platform and the data analysed in Progenesis QI, which is able to search HMDB for elemental composition matches. This returns 6 isomeric structures are returned. One of these, salicylamide can be rejected on the basis of a very poor match between the structure and the observed product ion spectrum. Calculation of DCCS using CCSondemand allows three of the matches to be downgraded in likelihood because the observed CCS value is very different from the predicted value. Of the two remaining matches, trigonelline is more likely, since it represents a primary metabolite, and this is the correct answer.

CONCLUSION

In our investigations of the utility of CCS prediction to support the characterization of unknowns, not every case was as clear cut as the above examples. In some cases, CCS predictions did not discriminate between sets of matches. CCSondemand performs best with proton adducts, and less well with sodium and potassium adducts. Where the chemistries involved are not well represented in the model basis set, predictions are also likely to be poorer. Nonetheless, predictions of collision cross section seem to afford a valuable addition to the unknown characterisation toolkit, as long as circumspect use is made of the predictions alongside hard experimental data points.

References

1. Broeckling, et al. Application of Predicted Collisional Cross Section to Metabolome Databases to Probabilistically Describe the Current and Future Ion Mobility Mass Spectrometry. https://pubs.acs.org/doi/10.1021/jasms.0c00375

2. Yanling et al/ Investigation and Performance Evaluation of a Research Prototype Tool for CCS Prediction ((2020).Waters Application Note.

3. Connolly et al,. Investigation into Small Molecule Isomeric Glucuronide Metabolite Differentiation Using In Silico and Experimental Collision Cross-Section Values, J Am Soc Mass Spectrom. 2021, 32(8):1976-1986

TO DOWNLOAD A COPY OF THIS POSTER, VISIT WWW.WATERS.COM/POSTERS