Deriving confidence metrics for automatic peak assignment through n-D Kendrick defect inference networks

Or “Getting software to do the hard work—so you don’t have to!”

David Kilgour[1]; C. Logan Mackay[2]; Pat Langridge-Smith[2] & Peter B. O'Connor[1]

1) University of Warwick, Coventry, UK; 2) SIRCAMS, University of Edinburgh, Edinburgh, UK

Overview

The data volume produced by Fourier Transform Ion Cyclotron Resonance Mass Spectrometers, analysing complex, small molecule samples (e.g. crude oil, natural organic matter, food products) requires fast and accurate automatic peak assignment algorithms in order to allow efficient interpretation. Kendrick mass defect and peak mass difference formulaic inference techniques have been developed to allow the elemental formulae of unassigned peaks to be inferred from those of assigned peaks – but, how confident can one be in the predictions of these methods? We have developed a novel, n-dimensional Kendrick defect inference network algorithm which uses artificial intelligence methods to produce improved confidence metrics of the resulting peak assignments.

2D KMD Space

Plotting complex mass spectra in multiple Kendrick mass defect (KMD) dimensions (Eqns [1] and [2][1]), using different Kendrick mass bases (e.g. CH_2, C or O) can make it easier to manually interpret the data. Commonly, 2D KMD plots have proven the most useful, but the approach can be easily extended to higher dimensions if required.

$$KM_{base} = M_{peak}.\frac{[M_{base}]}{M_{base}} \quad [1] \qquad KMD_{base} = [KM_{base}] - KM_{base} \quad [2]$$

Pairs of peaks which can be connected by the same transformation vectors, in this Kendrick space, will have the same relative formulaic change between them as shown in Figure 1. Correlating these transformation vectors to formulaic differences allows an analyst to infer the elemental composition of previously unknown peaks from known peaks.

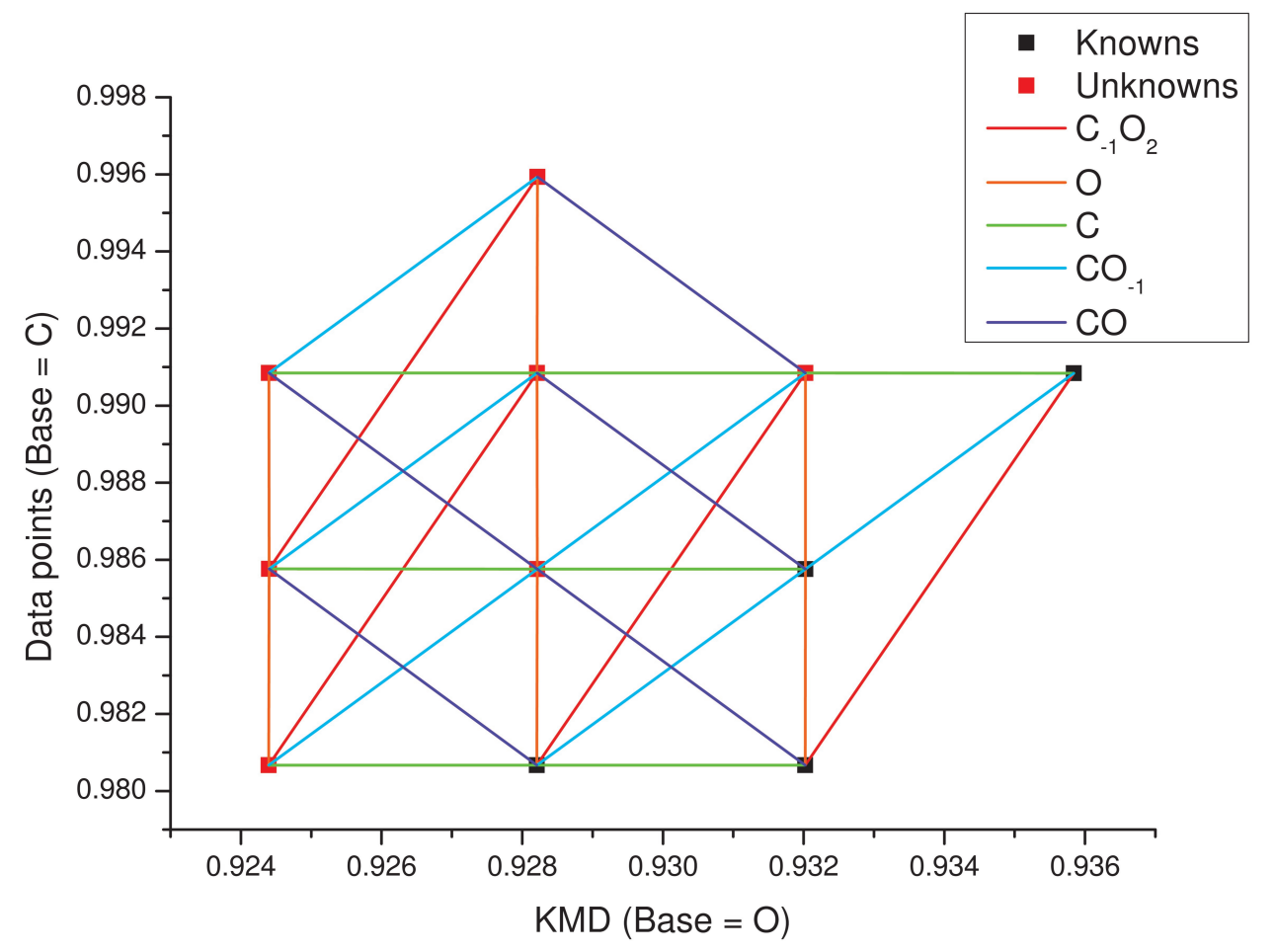
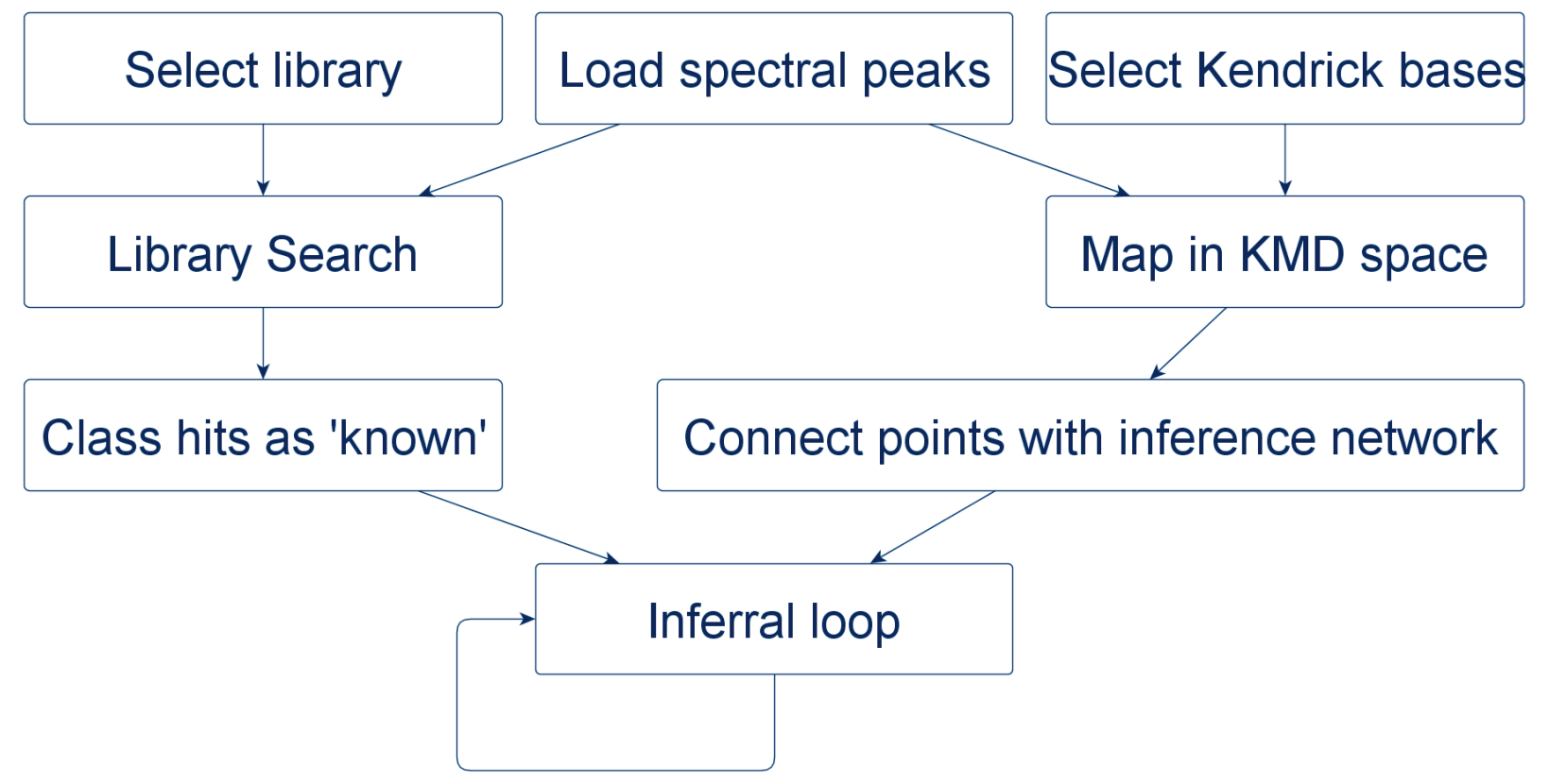


Figure 1: 2D KMD plot showing a subset of the ‘known’ and ‘unknown’ peaks from a negative mode electrospray mass spectrum of diluted malt whisky. The Kendrick bases selected are O and C. The connecting lines correspond to mapping vectors related to known formulaic differences.

Basic Algorithm

The basic algorithm is shown in Figure 2. The first step in the algorithm is to attempt to identify peaks by means of a library search. We use various libraries including an accurate mass version of the NIST GC/MS library and in-house developed CHO, CHONS and other task specific libraries. Peaks which can be matched to a library entry are classed as ’knowns’. The unidentified peaks are classed as ‘unknowns’.

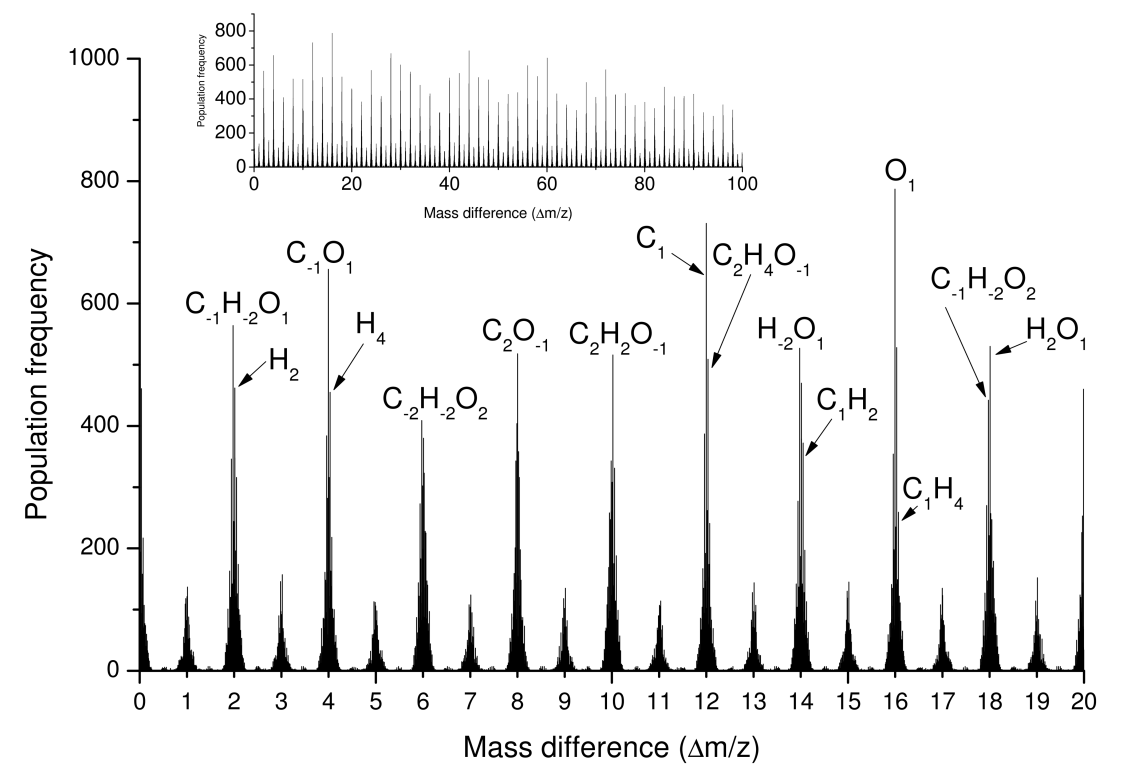
Figure 2: The basic automatic peak assignment algorithm.



The Inference Network

The next step is to generate the inference network. The mass data is re-plotted in 2D KMD space. Points in the 2D KMD space are connected if the mapping vector between them can be assigned to a known, or expected, formulaic difference—as shown in Figure 1. With a new sample type, where the homologous series present are not known, the mass spectrum can be statistically analysed to identify common mass differences within the dataset[2] as shown in Figure 3. These are matched against a molecular fragment library (commonly C, H, O, N & S). These molecular fragments are then used to generate of a list of mapping vectors in the 2D KMD space which can be saved and used to investigate other, similar samples.

Figure 3: Common mass differences in a mass spectrum can be identified and then matched against a molecular fragment library. Some of the statistically common mass differences from a whisky mass spectrum are labelled here.



The system now has a set of ‘knowns’ and a set of ‘unknowns’ connected by the inference network, an example of which is shown in Figure 1.

The algorithm can infer the formula of ‘unknowns’ directly connected to ‘knowns’; then the next layer of ‘unknowns’ can be inferred, and so on (the Inferral Loop) until all peaks which are linked into the inference network have formulae assigned to them.

This basic process can be used to automatically assign formulae in mass spectra, but the confidence one would have in the results would be highly uncertain. Therefore we apply a range of confidence metrics to improve the performance of the algorithm.

Basic Mass Spectrometric Metrics

Basic mass spectrometric confidence metrics are used to provide increased confidence in the assigned formulae; all peaks identified as library matches must be within a certain mass error of the library entry, All detected isotopologue peaks must exhibit a relative ion abundance within defined bounds, and all inferred formulae must meet stoichiometric requirements defined by a simple rings plus double bonds check.

These metrics are common to many formula assignment algorithms. However, we also apply some more novel approaches.

Artificial Immune System Derived Metrics

Artificial immune systems are a type of adaptable artificial intelligence which are useful for classification, anomaly detection and other recognition tasks. Artificial neural networks mimic the processing abilities of the brain while artificial immune systems mimic the control mechanisms thought to be important in the distributed intelligence of the mammalian immune system.[3-6]

A key part of the control mechanism in the immune system is provided by the B cells and the degree to which they are stimulated[6] which is moderated by the closeness of the match between the B cell antigens and the pathogens which it encounters. The stimulation level triggers a clonal selection process; insufficiently stimulated cells are culled from the immune system (by apoptosis) whereas B cells which are stimulated above a threshold are set to multiply and mutate, to better detect the pathogens.

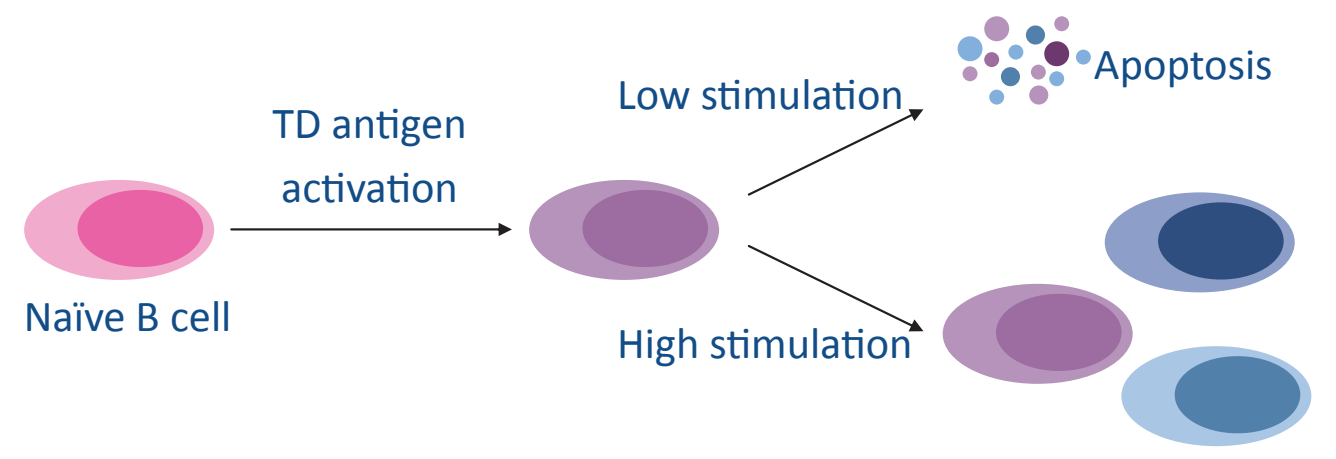
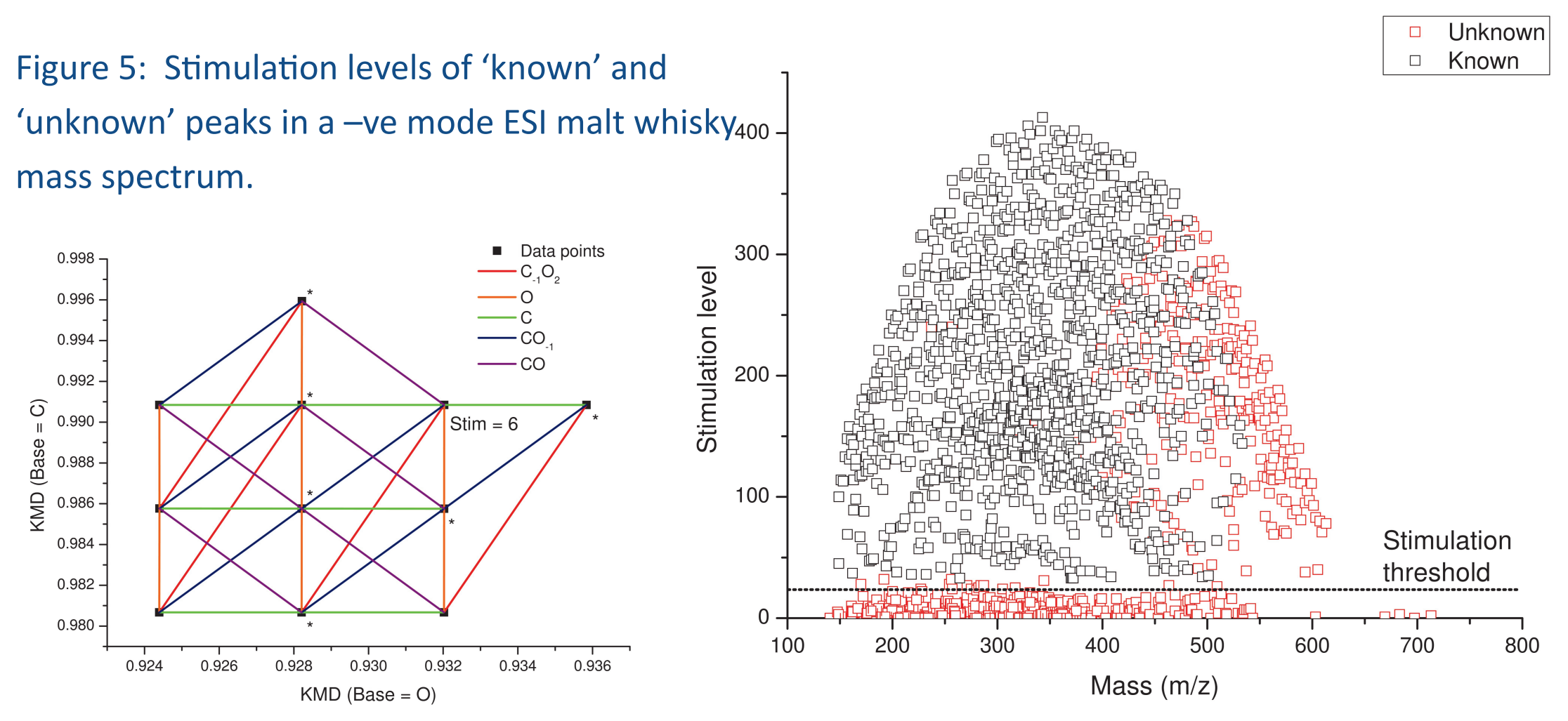


Figure 4: Part of the control system in the mammalian immune system.

Taking each peak in turn, we treat it as an artificial B cell and all other peaks in the spectrum as potential pathogens. The B cell can detect a pathogen if it is directly connected through the inference network. The total stimulation of a B cell is the sum of all the connections that cell has to potential pathogens (Figure 5, lower left). In this way, we can record the stimulation level of all peaks in the spectrum as shown in Figure 5, right.

Figure 5: Stimulation levels of ‘known’ and ‘unknown’ peaks in a –ve mode ESI malt whisky mass spectrum.



Assigned peaks, below the user defined stimulation threshold, cannot act as inference sources—this prevents the errors associated with potentially mis-assigned or artefact peaks being propagated through large parts of the inference network.

Additionally, ‘inferred’ peaks cannot be assigned until their stimulation level crosses a separate, user defined threshold. This can be thought of as analogous to the inverse of the culling process in the biological immune system.

Uniqueness Metric

The uniqueness metric is intended to reduce the possibility of mis-assigned library matches. Not only must a library match be within a certain accuracy of the peak mass, but that there must be only that single library hit within an even larger mass range—the Uniqueness Range (see Figure 6). Therefore it is very unlikely that that peak has been mis-assigned.

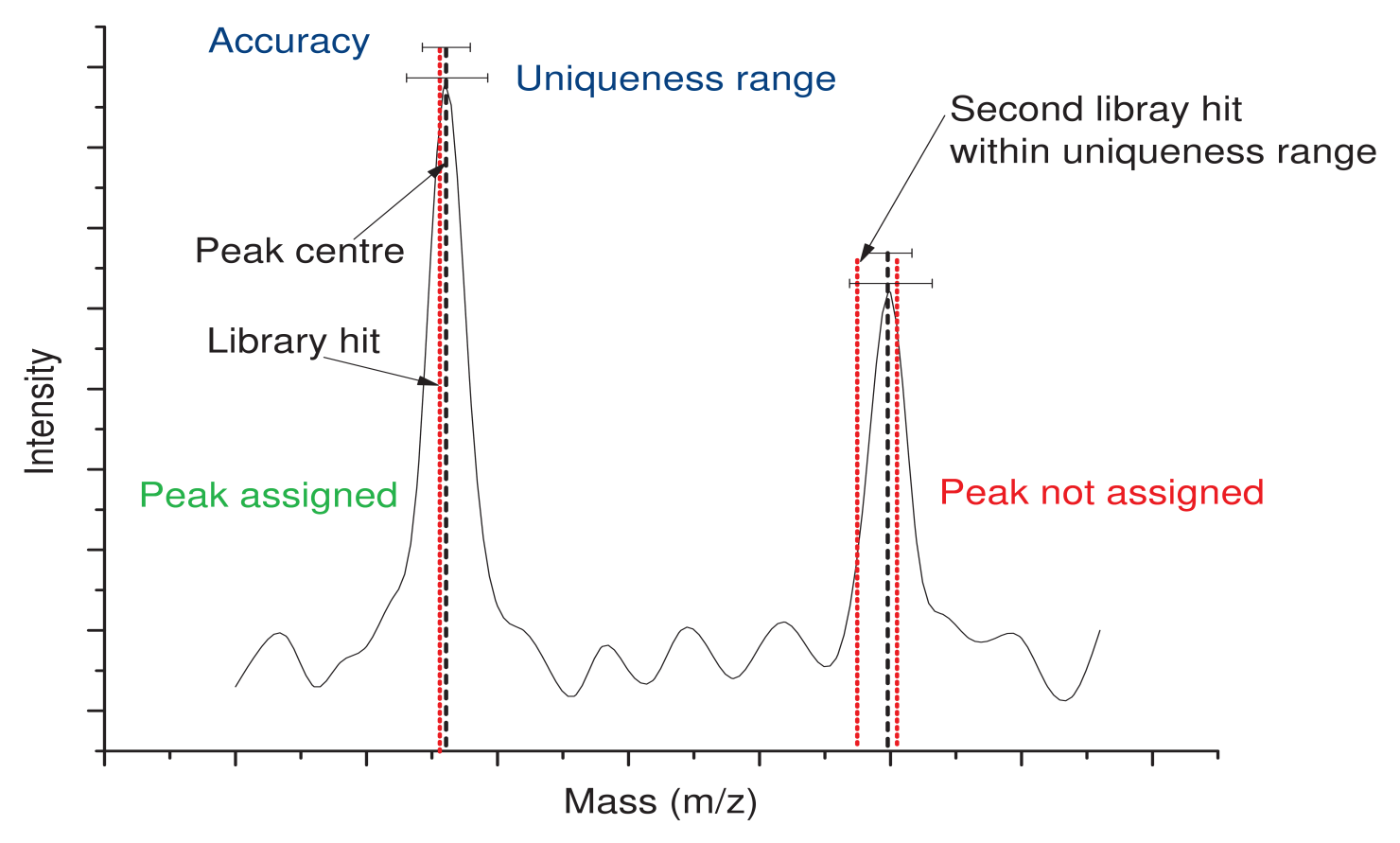


Figure 6: Showing the concept behind the Uniqueness Range.

For peaks where there is more than one library hit within the Uniqueness Range, no formula assignment will be returned for that peak from the library search—the peak, will, most likely be assigned through the Inferral Loop though (see Figure 7).

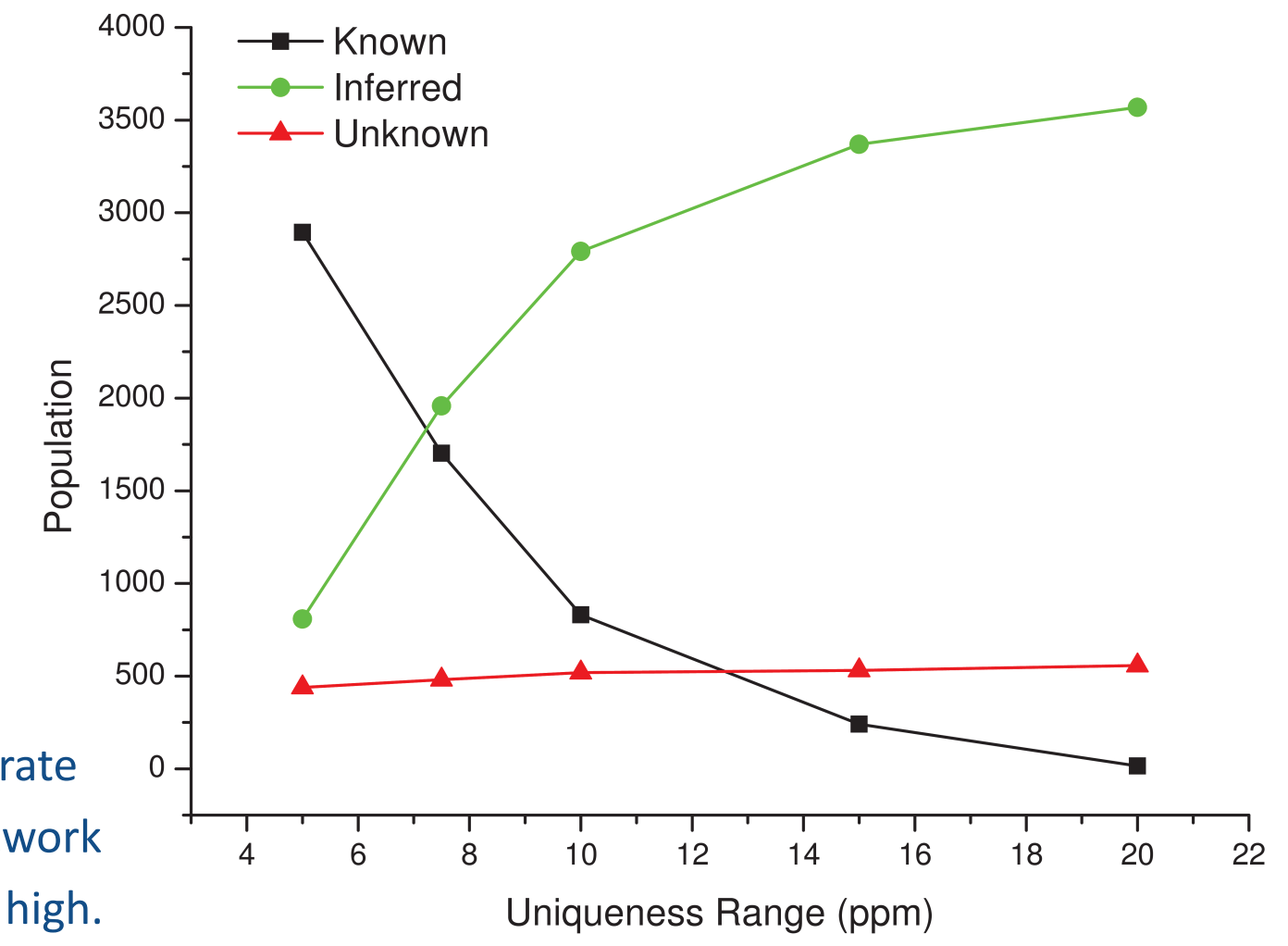


Figure 7: Showing that the total assignment rate can be maintained through the inference network even when the Uniqueness Range is set very high. (-ve mode ESI FT-ICR MS of Fulvic acids)

Consistency Metric

There will be many routes through the inference network between any two points. The Consistency Metric allows the user to require that all routes between two points must result in the same formulaic difference—i.e. the inference network must be 100% internally consistent, as shown in Figure 8.

This metric can be used to probe the validity of the confidence settings used to process a given spectrum. For example, consider the user manipulation of the required accuracy of linkages in the inference network — has this been set too loosely? In the example shown in Figure 9, the accuracy requirement for valid linkages is systematically degraded. As the accuracy requirement is degraded, the number of connections in the inference network increases. However, after a threshold, the network ceases to be 100% internally consistent and the peak assignment rate begins to drop off. Therefore the spectrum accuracy requirement setting must be maintained higher than this threshold.

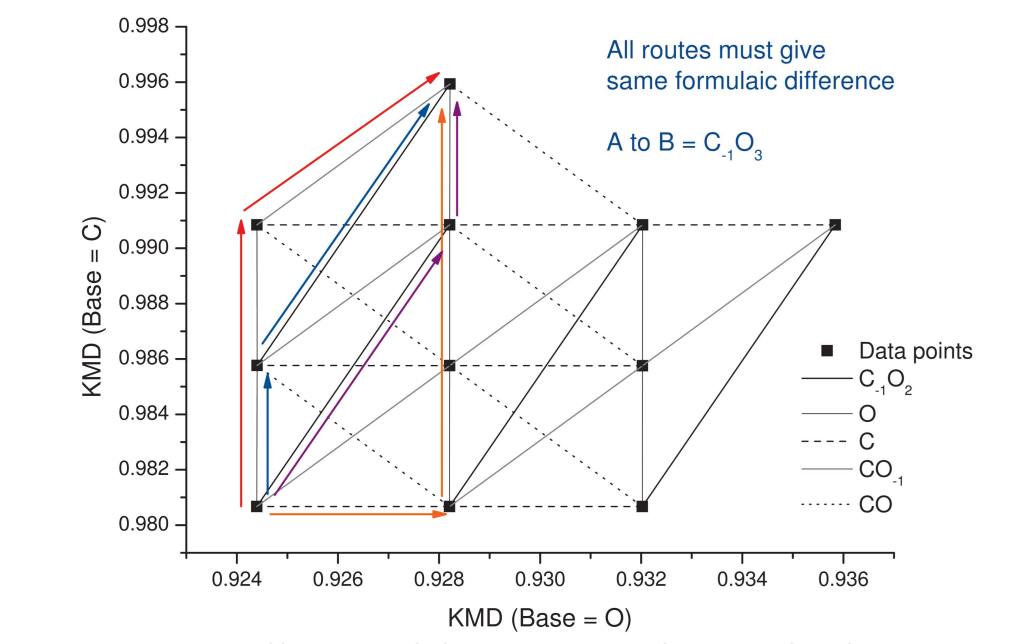


Figure 8: All possible routes through the inference network must result in the same formulaic difference.

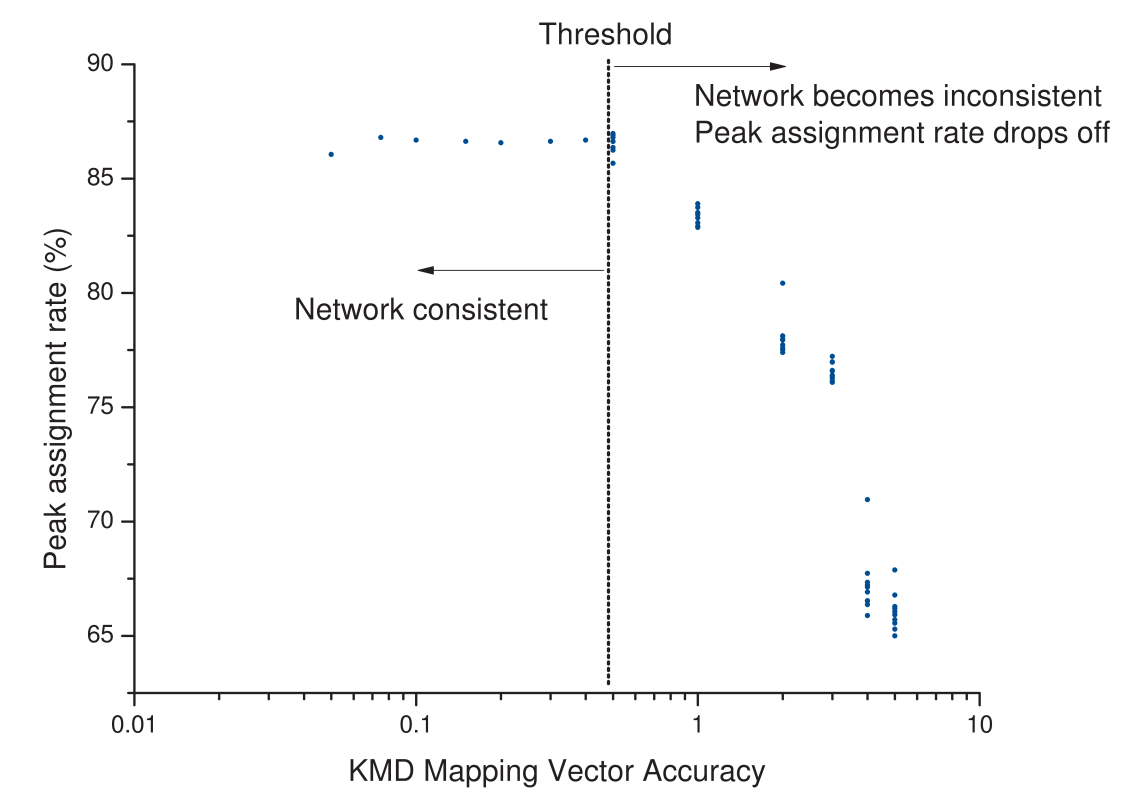


Figure 9: Showing the growing inconsistency of the inference network above a certain accuracy threshold resulting in reduced peak assignment rates for malt whisky mass spectra (10 replicates).

Charge State Deconvolution—With No Isotope Peaks

Knowing the charge state of an ion is crucial to being able to assign a formula to it. However, for many peaks in complex mass spectra, the intensity of the isotopic peaks may be too low to allow the charge state to be calculated from isotopic spacings.

We have found that it is possible to use the Inference Network to estimate the charge state of peaks in complex mass spectra which allows the peak assignment rate to be greatly improved.

Assumptions: This process is based on the assumption that the majority of the peaks in the spectrum are singly charged small molecule ions and that members the same homologous series will be present in all charge states.

Method: Build inference networks for all charge states of interest; use the same homologous series that were discovered to build the inference network, as before, but reduce the connection vector lengths proportional to the charge state of interest—i.e. vector length for +2 will be half that for +1. Record the stimulation level of each ion within the different inference networks. The ion stimulation level should be highest for one charge state over the others. That is when the ion fits into the most homologous series and is likely to reflect the actual charge state of the ion.

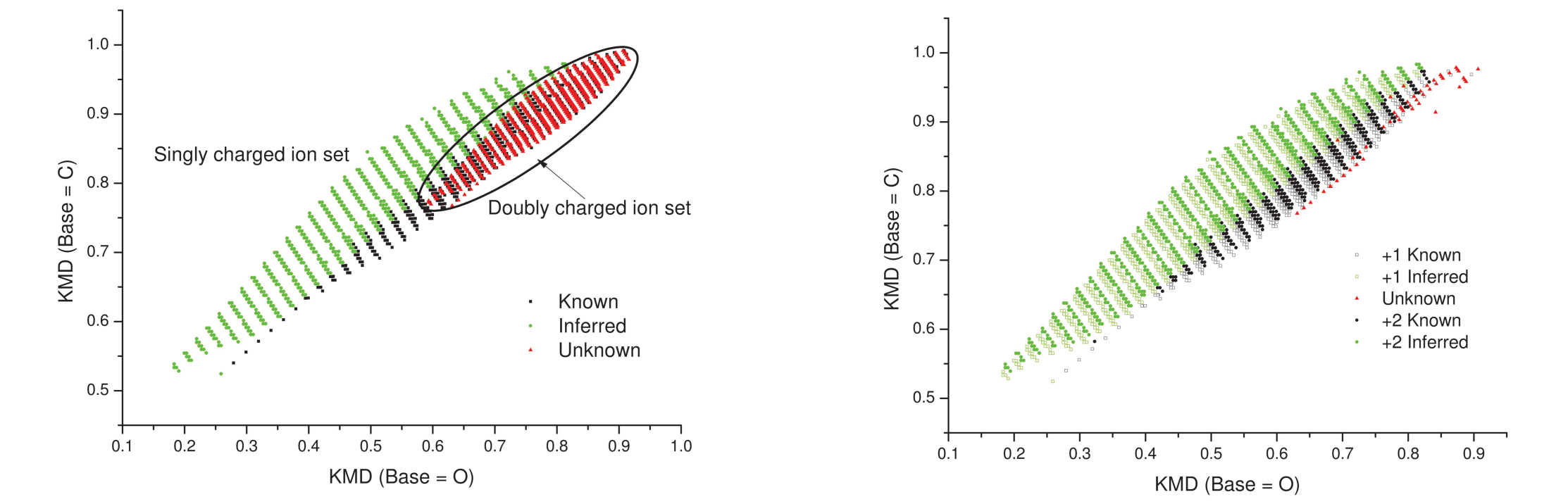


Figure 10: Left: Prior to deconvolution, doubly charged ions are not identified. Right: After deconvolution, most ions are identified.

Deconvolution Example—Synthetic Data

A synthetic data set was created using the first 65 series of fulvic acids presented by Stenson *et al.*[7] This dataset contains both singly (monosodiated) and doubly (disodiated) ions for all members. This dataset was analysed both before and after deconvolution and the results (Figure 10) show a marked improvement in the peak assignment rate after deconvolution.

Deconvolution Example—Whisky

The same method was used to deconvolve a +ve mode ESI spectrum of malt whisky. Prior to deconvolution, the algorithm returns an assignment rate of 74% (2164 peaks out of 2516 peaks being assigned unique formulae) with a mass accuracy requirement of 200ppb, a uniqueness threshold of 400ppb and the requirement that the inference network be 100% internally consistent.

As a result of deconvolution, 2203 peaks are most stimulated in (and hence assigned to) charge state +1 and 313 to charge state +2. Using the same formula assignment set-up as for the un-deconvolved spectrum, the assignment rate for the +1 peaks rises to 84%, but the +2 peaks achieve poor assignment rate. Further investigation, undertaken by adjusting the mass accuracy and uniqueness range of the assignment algorithm, reveals that the poor assignment rate of the deconvolved +2 peaks is a result of the fact that they suffer an apparent 2nd order systematic mass calibration error as shown in Figure 11.

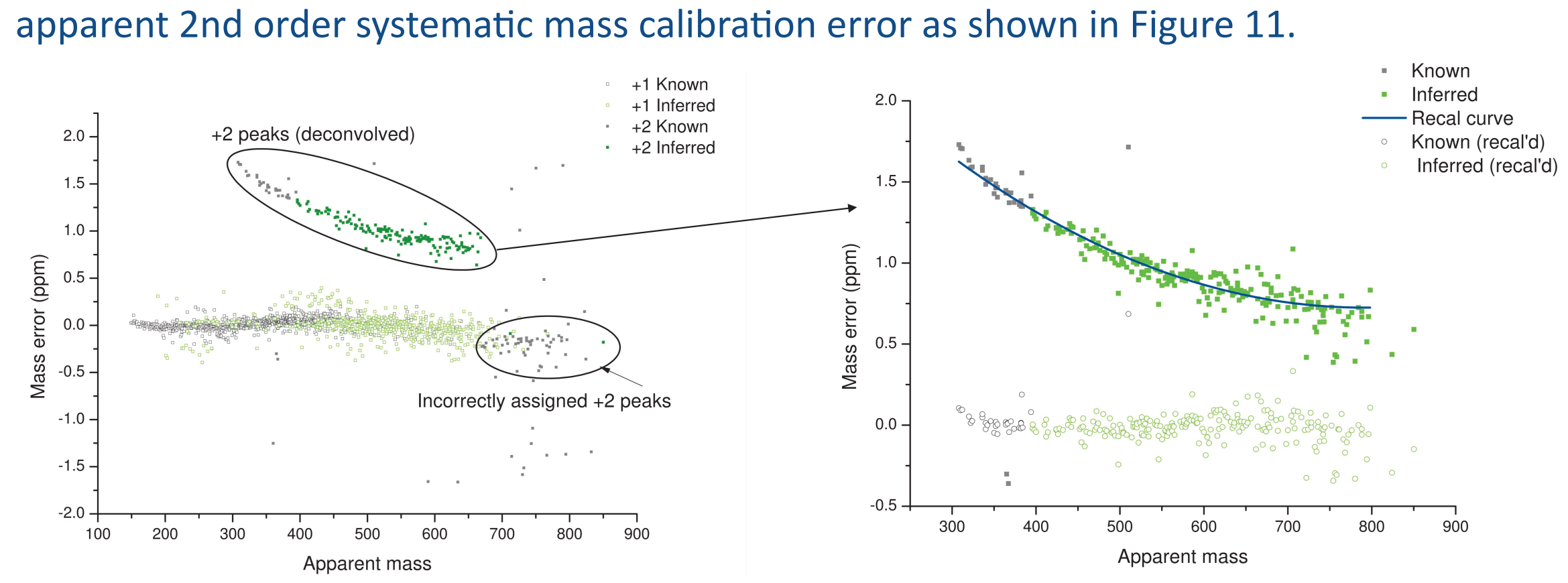


Figure 11: Results of deconvolution of +ve mode ESI mass spectrum of malt whisky. +2 charge assigned peaks suffer an apparent systematic mass error believed to be a result of the differential effect of electric field imperfections on ions of identical mass but different charge.

This systematic mass error is thought to be due to the greater relative effect that electric field imperfections (space charge, image charge or static trapping field imperfections) will have on doubly charged ions as opposed to singly charged ions of the same mass. In an FT-ICR, this effect would be predicted to have a linear effect on the frequency of the ions which would convert to a quadratic effect on the mass to charge[8]. As can be seen in Figure 11, the systematic mass error does indeed follow a second order polynomial relationship as a function of mass and this can be used to recalibrate the doubly charged ions to correct this effect. After deconvolution and recalibration, 87% of the doubly charge ions can be confidently assigned giving a total assignment rate of 84%.

Conclusions

New metrics of confidence and a novel artificial intelligence method of assigning charge states to ions can improve your ability to automatically assign formulae to peaks in complex mass spectra—this could greatly improve the rate at which such spectra can be processed.

Acknowledgments

We would like to acknowledge Matthias Witt for providing the malt whisky spectra used to test the algorithm and Mark Neal (University of Aberystwyth) for useful discussions on artificial immune systems.

References

1. E. Kendrick, Anal. Chem. 1963, 35. 2146-2154.
2. E. V. Kunenkov *et al.*, Anal. Chem. 2009, 81. 10106-10115.
3. L. N. de Castro, J. Timmis, Artificial Immune Systems. Springer-Verlag: Berlin, Germany, 2002.
4. J. D. Farmer, N. H. Packard, A. S. Perelson, Phys. D 1986, 22. 187-204.
5. E. Hart, J. Timmis, Applied Soft Computing 2008, 8. 191-201.
6. J. Timmis, M. Neal, Knowledge-Based Systems 2001, 14. 121-130.
7. A. C. Stenson *et al.,* Anal. Chem. 2003, 75. 1275-1284.
8. P. K. Taylor, I. J. Amster, Int. J. Mass Spectrom. 2003, 222. 351-361.