

Helen Fuell<sup>1</sup>, Manfred Beckmann<sup>2</sup>, John Draper<sup>2</sup>, Oliver Fiehn<sup>3</sup>, Nigel Hardy<sup>1</sup> and Birgit Linkohr<sup>3</sup>

3. Department of Computer Science, University of Wales, Penglais, Aberystwyth, SY23 3DB

4. Institute of Biological Sciences, University of Wales, Penglais, Aberystwyth, SY23 3DB

5. Max Planck Institute of Molecular Plant Physiology, 14424 Potsdam, Germany

## 1. Introduction

ArMet (Architecture for Metabolomics) aims to provide a standard representation of the data associated with metabolomics research. It is designed to encompass the entire timeline of metabolomics experiments from descriptions of the biological source material and the experiments themselves through, sample growth, collection and preparation for analysis by technologies such as GC-MS, NMR, FTIR, to the results of those analyses. It does this by way of nine packages that define core data applicable to a wide range of experiments. Extensive requirements analysis has resulted in sub-packages for ArMet which describe sample growth, collection and preparation for GC-MS analyses for plant metabolomics experiments. The aim of these sub-packages is to provide sufficient data on each experiment to enable meaningful comparison of samples analysed in different laboratories on different GC-MS platforms using statistical and data mining techniques. They, therefore, maintain data on as many of the sources of variability evident in the experimental process as possible.

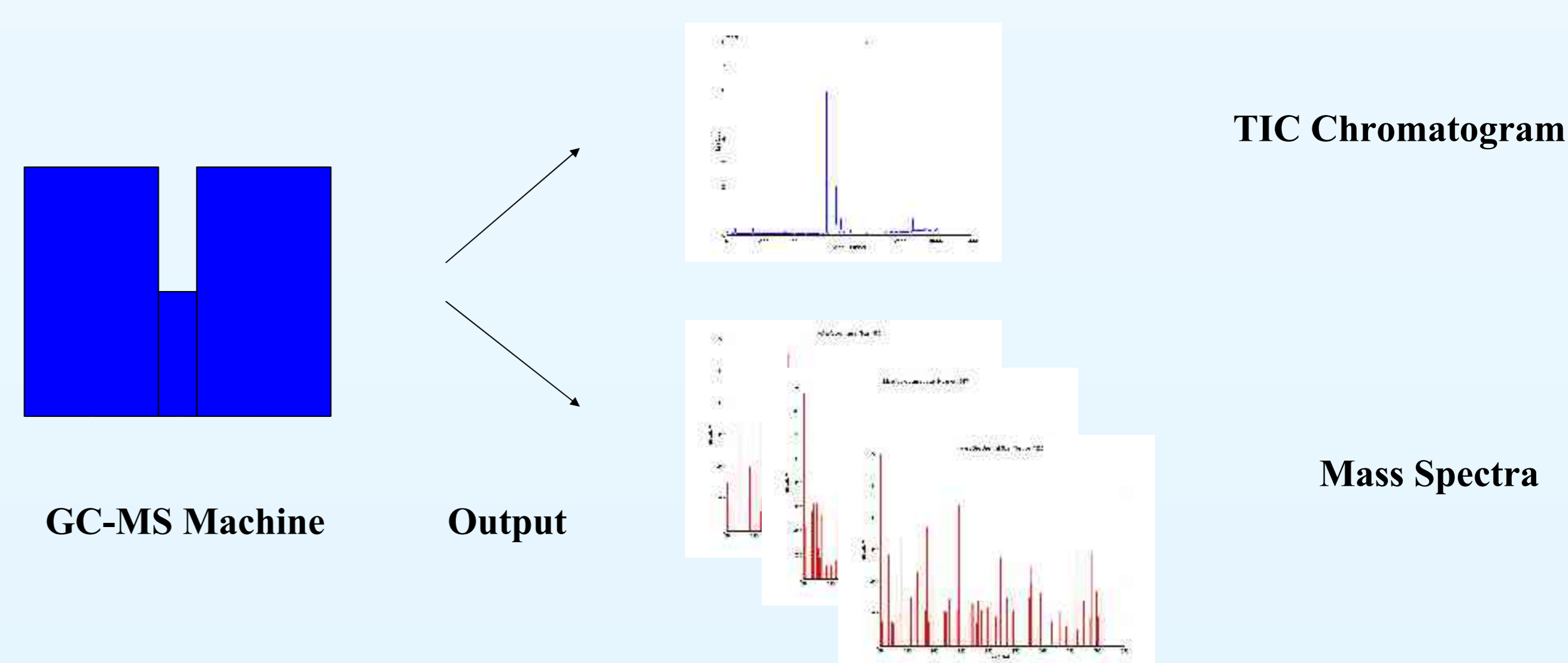
This poster describes work carried out to develop a sub-package for ArMet that describes the results of GC-MS analysis carried out upon the samples produced during a plant metabolomics experiment.

## 2. GC-MS Data

There are a variety of types of metabolomics analysis:

- **Fingerprinting:** Fingerprint analysis involves the development of complete metabolome descriptions for samples without knowledge of the chemical identities of the compounds that the samples contain. TIC (Total Ion Current) chromatograms are examples of such descriptions which may be used to globally compare the metabolomes of samples.
- **Targeted analysis:** Targeted analysis involves detection and precise quantification of a single or small set of target compounds within the metabolome of a sample.
- **Metabolite Profiling:** Profiling involves the detection, identification and approximate quantification of a large set of target compounds within the metabolome of a sample.
- **Metabolomics:** Metabolomics involves the detection, tentative identification and approximate quantification of as many of the compounds within the metabolome of a sample as possible.

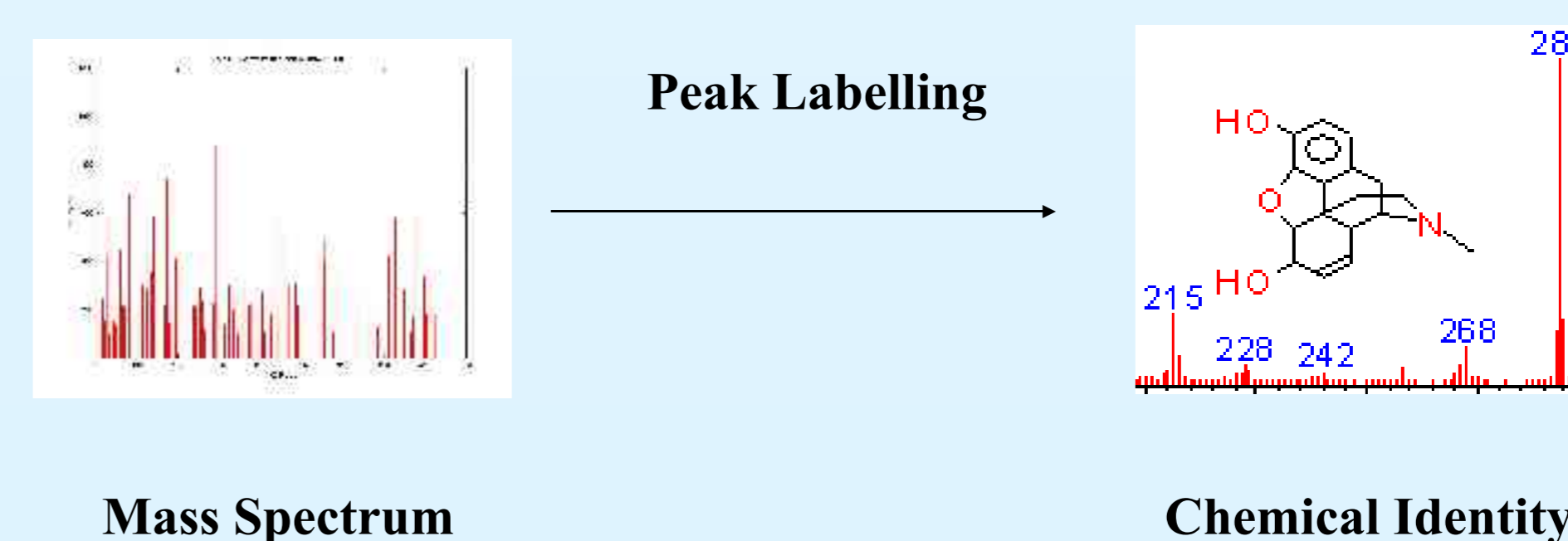
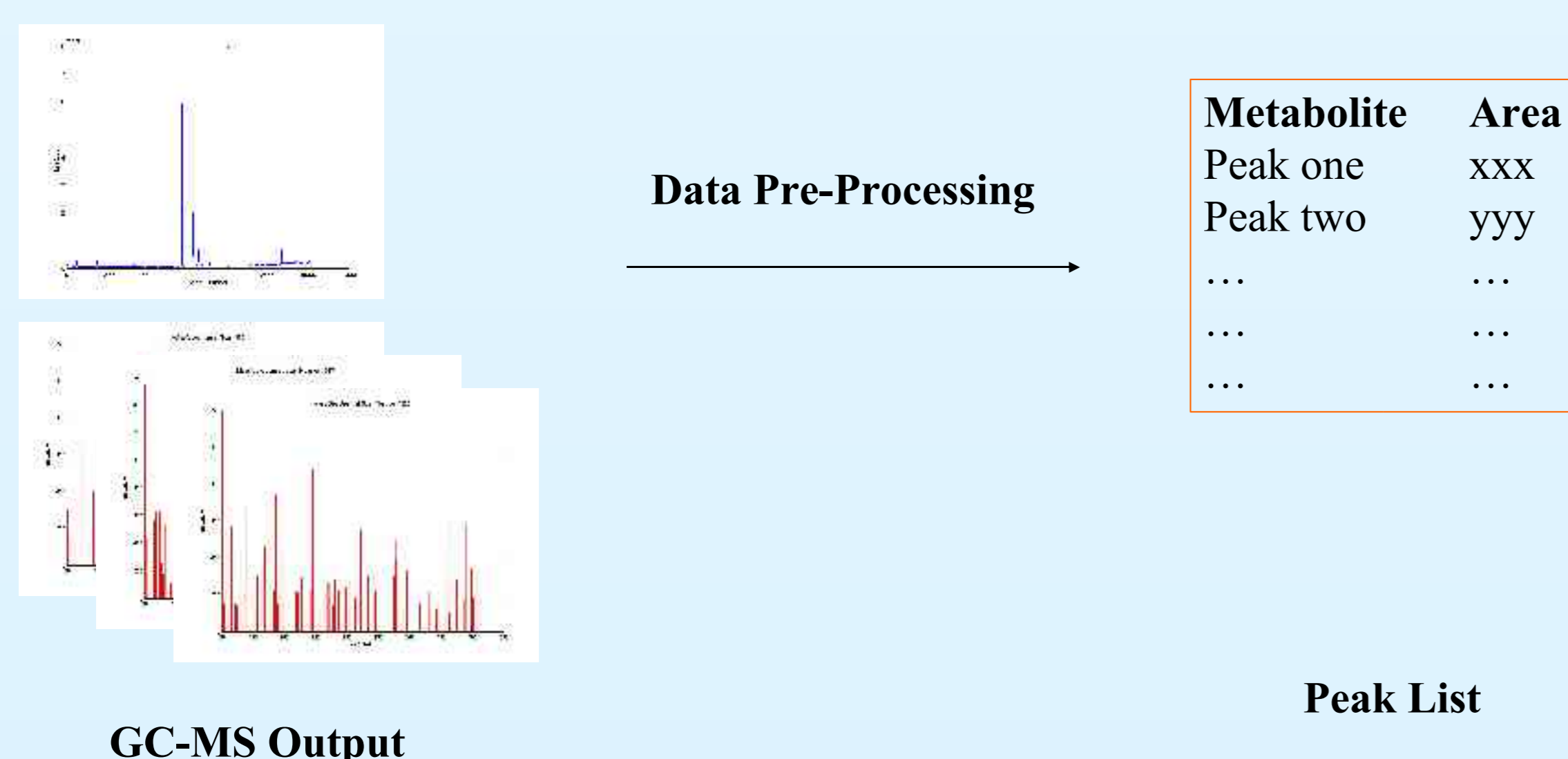
The ArMet GC-MS sub-package supports all four of these types of analysis. Fingerprinting datasets have a simple structure which represents the raw data from instrumental analysis, i.e. values from a TIC chromatogram and its associated mass spectra or a summed mass spectrum. However, the output from instrumental analysis must be processed to produce the quantified list of metabolites that comprise targeted analysis, metabolite profiling or metabolomics metabolome descriptions.



## 3. Data Pre-Processing

We identified the stages of pre-processing as follows:

- **Noise removal:** Thresholding mass intensities and reconstructing the chromatogram accordingly.
- **Peak detection:** Analysis to decide the start and end of peaks in the TIC chromatogram.
- **Peak deconvolution:** Analysis of mass spectra to identify separate and co-eluting compounds.
- **Peak quantification:** Determining the area under the peaks in order to quantify the individual metabolites either with respect to one another or to an internal standard.



## 4. Operating Procedure Factors that Affect Dataset Comparability

Each of the stages of pre-processing was examined and characterised on the assumption that this will give data analysts the greatest opportunity for meaningful comparison of data collected under different sets of procedures.

This characterisation led to descriptions of the stages in the following terms:

- The software and algorithms used to perform the different stages of pre-processing.
- The parameters to the software and algorithms employed.
- Any procedures for manual adjustment or correction of the automatic output of each stage.

In the absence of standardised operating procedures in this area, it was decided that the ArMet GC-MS module should annotate the datasets that it maintains with these descriptions of the stages of pre-processing.

## 5. Peak Labelling

Peak identification for targeted analysis or metabolite profiling involves the comparison of the peaks for a sample with external standards or a reference list of target compounds. These datasets should, therefore, be annotated with these details.

Peak identification for metabolomics involves the use of spectral libraries to label GC-MS peaks with their chemical identities. Characterisation of this process resulted in the following description:

- The software and algorithms used to perform spectral library look-up.
- The parameters to the software and algorithms used.
- The spectral library used.

As the use of two different spectral libraries can result in zero, one or more different chemical identities being attributed to a peak it was decided that within the ArMet GC-MS sub-package the primary label for each metabolomics peak would be the 20 most abundant m/z values, and associated relative intensities, from the representative mass spectrum for the peak. In addition each peak may be given candidate chemical identities by a number of different methods each with an associated confidence value.

## 6. Results

Following our investigations we modelled the data required to represent the results of GC-MS plant metabolomics profiling experiments.

Our model has the following features:

- Support for all four types of metabolomics analysis.
- Metabolomics datasets containing peaks described by retention and area information and labelled with the 20 most abundant m/z values, and associated relative intensities, from the mass spectrum that best describes each one. Each peak also has a set of zero or more candidate chemical identities annotated with provenance and confidence.
- Targeted analysis and metabolite profiling datasets annotated with information about the external standards or target compounds that they contain.
- Fingerprinting datasets as either TIC chromatograms with or without associated mass spectra or summed mass spectra.
- Descriptions of the stages of pre-processing.
- A reference to the archive of the raw data

We have represented our model using the Unified Modelling Language (UML) and translated this representation into a database implementation which is undergoing evaluation.

