



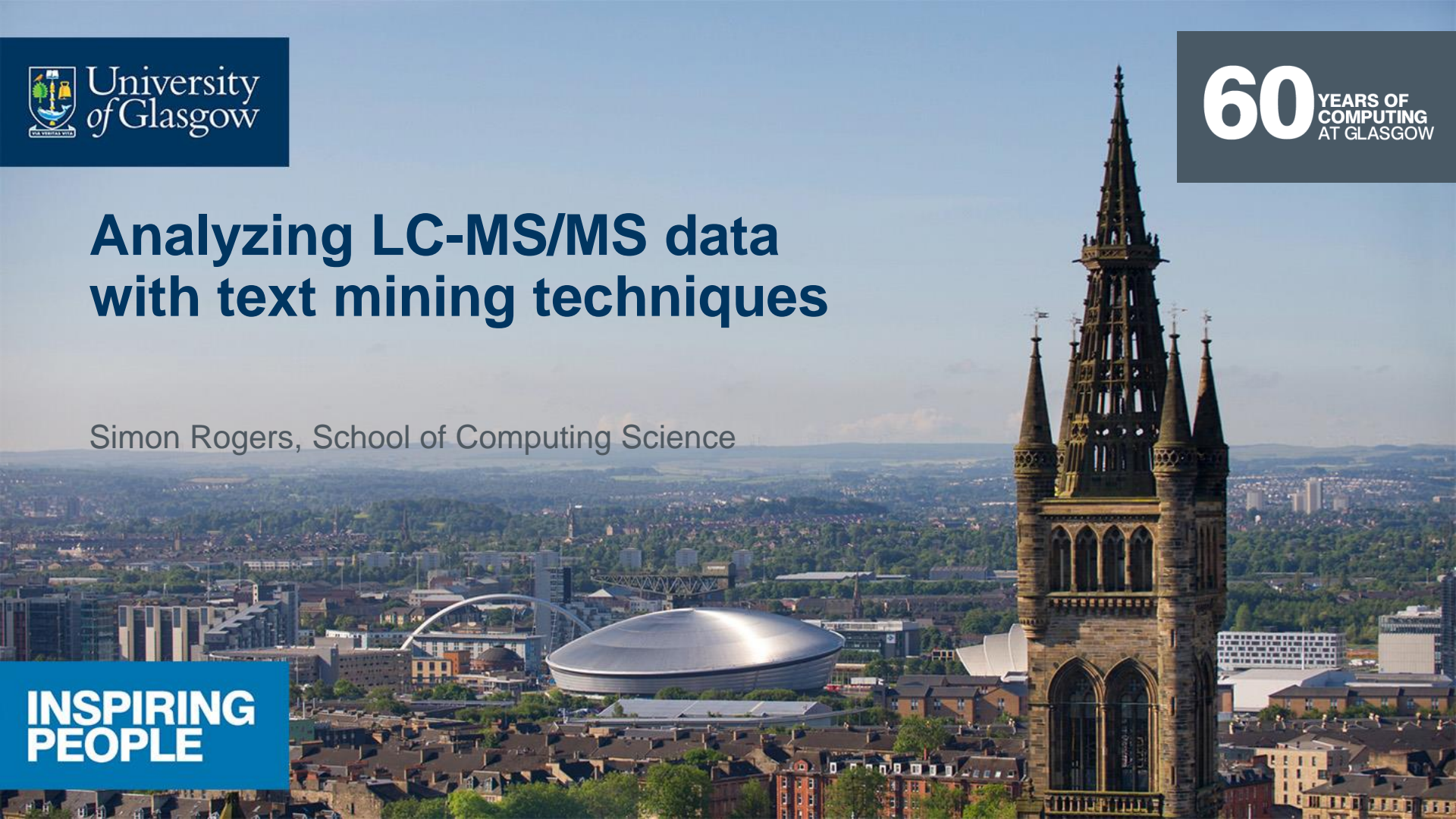
University
of Glasgow

60 YEARS OF
COMPUTING
AT GLASGOW

Analyzing LC-MS/MS data with text mining techniques

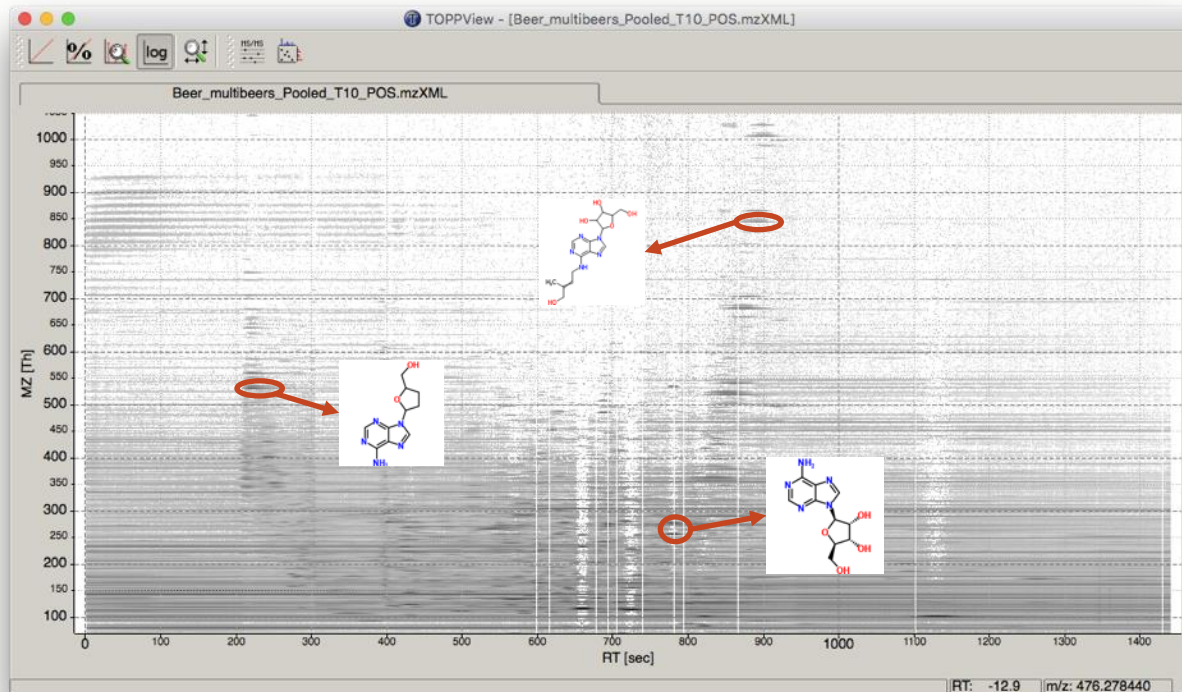
Simon Rogers, School of Computing Science

**INSPIRING
PEOPLE**



Metabolomics is hard

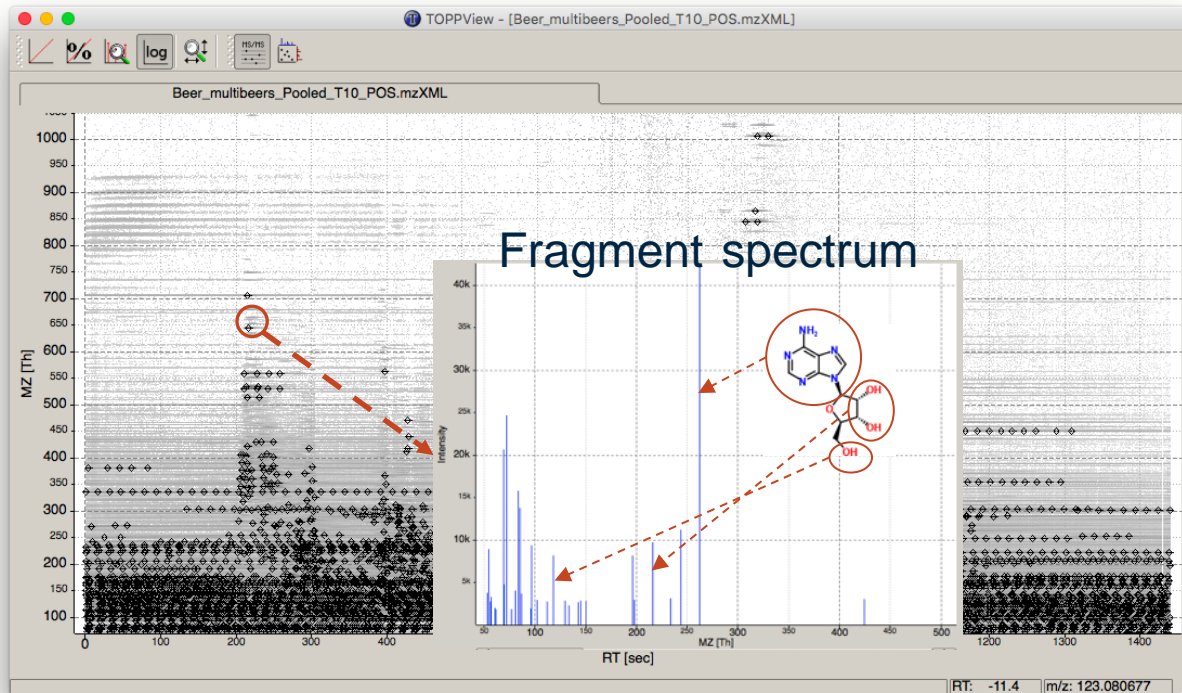
- Beer, run through a mass spectrometer
- Blobs are molecules
- Here there are ~4000 blobs
- *What are they?*
- *How do they differ across beers?*





Traditionally...

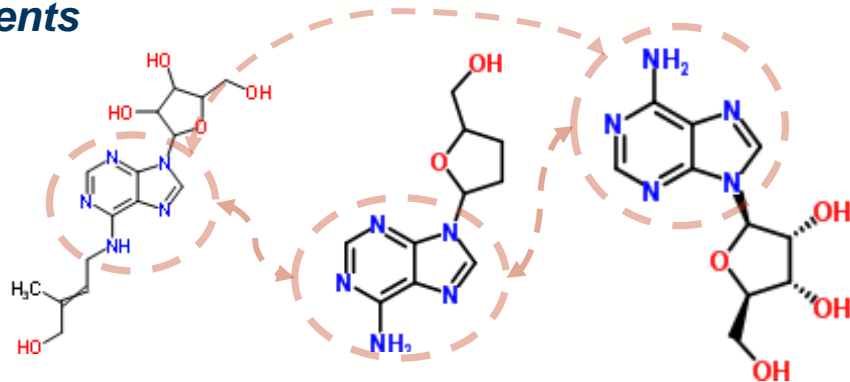
- Mass is not enough to identify
- So, we **fragment**
- ~2000 unknown molecules are fragmented (black dots)
- Query databases with fragment spectra
- ~**2% identified**





A data-centric approach

- The strategy of ID from spectral databases is fundamentally limited (esp. for discovering new molecules)
- The ~2000 fragment spectra in a dataset are **not independent**
- Molecules are built from a library of building blocks
- *Parallels with collections of text documents*
- *Lots of nice models exist for text*





Topic Modelling

Classic LDA for text

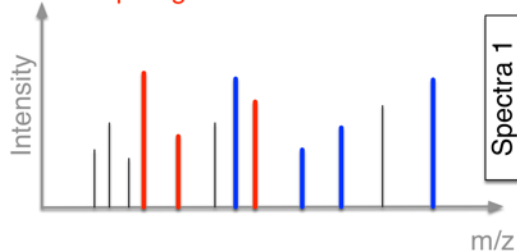
MS2LDA for fragments and losses

Football-related topic

Asparagine-related Mass2Motif

Document 1

Hereford **United**, the **club** formed in 1924 who have **played** continuously in the **Football** League lower **divisions** or in the senior **semi-professional game** for 90 years, has been put into **liquidation**. The **club company**, The **club lawyer's** argued that its owner, Andy Lonsdale, had proof of £1m **funding** to pay the **club's creditors**, but was stuck in traffic.

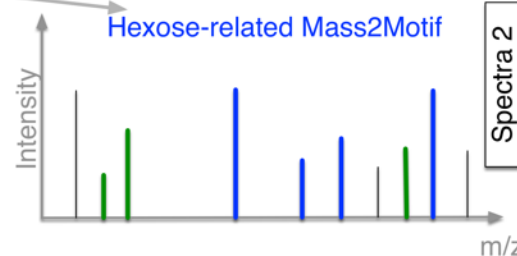


Business-related topic

Hexose-related Mass2Motif

Document 2

One of Britain's leading **solar entrepreneurs** is set to announce that his **business** has gone into **liquidation**, in the third high-profile casualty for the **sector** this month. [...] Howard Johns, the former **chairman** of the **Solar Trade** Association and an adviser on **renewable energy**...



Environment-related topic

Adenine-related Mass2Motif

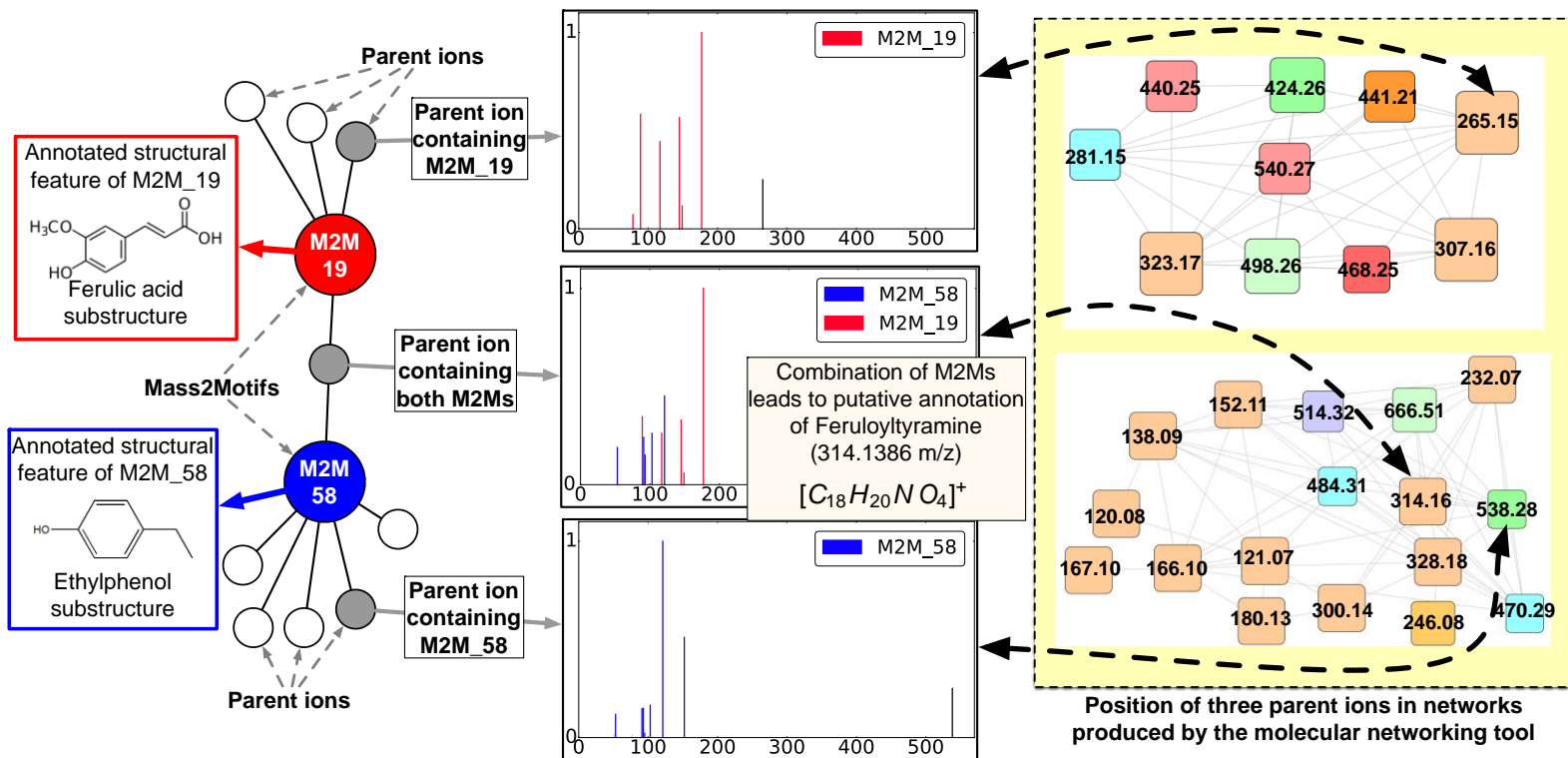
Can we use topic modelling to **decompose** fragment data into biochemically relevant topics?

Yes...but why?

- We know we cannot **identify** many molecules.
- But we can identify many of the substructures.
- An **unknown molecule** including a **known substructure** provides us with some useful information (better than nothing).
- *Spoiler*: we found that by identifying ~30 substructures we knew something about **70%** of the beer molecules.
- C.f. documents: if we can identify a topic as 'football' we can annotate all documents including it as 'football related'



Results taster



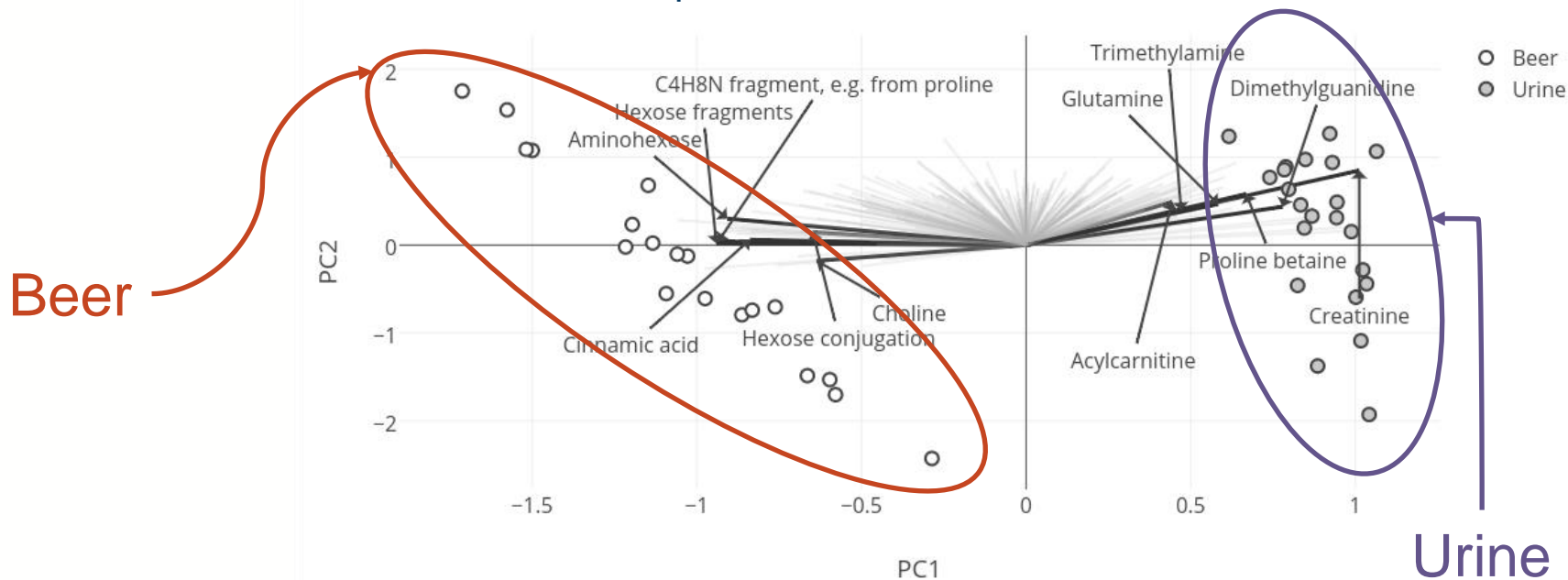
What's next?

- Most researchers are interested in how the metabolome **changes** (e.g. case v control; drug 1 v drug 2; etc)
- Traditionally done at the molecule level (for small subset that can be identified)
- We can now do this at the substructure level by decomposing multiple samples with the same set of building blocks.
- ***How does the prevalence of a particular building block vary...***
- C.f. Text: techniques have been developed to model how topics appear, disappear, and change over time (e.g. 'Brexit'; 'Fake News'; etc)



Beer v urine

- ~20 beers and ~20 urine samples analysed together
- Perform PCA on the substructure prevalence





Conclusions

- Translating techniques from 'rich' areas (e.g. text) to 'poor' areas (e.g. metabolomics) has high potential.
- IMO text modelling is particularly good for borrowing from as it is a) quite mature and b) used to fairly large datasets (i.e. they've found a happy medium between model complexity and computational feasibility).
- It still takes a long time (18 months from start to publication)
- Metabolomics is still hard but text approaches are helping...

Online visualisation tool

Topic modeling for untargeted substructure exploration in metabolomics

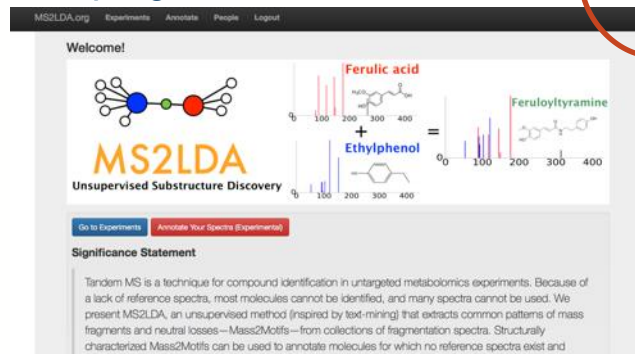
Justin Johan Jozias van der Hooft^{a,b}, Joe Wandy^{a,c}, Michael P. Barrett^{a,d}, Karl E. V. Burgess^a, and Simon Rogers^{a,c,1}

^aGlasgow Polyomics, University of Glasgow, Glasgow G61 1QH, United Kingdom; ^bInstitute of Infection, Immunity, and Inflammation, College of Medical, Veterinary, and Life Sciences, University of Glasgow, Glasgow G12 8TA, United Kingdom; ^cSchool of Computing Science, University of Glasgow, Glasgow G12 8RZ, United Kingdom; and ^dWellcome Trust Centre for Molecular Parasitology, Institute of Infection, Immunity and Inflammation, University of Glasgow, Glasgow G12 8TA, United Kingdom

Edited by Jerrold Meinwald, Cornell University, Ithaca, NY, and approved October 12, 2016 (received for review May 20, 2016)

Acknowledgements

Justin van der Hooft (Glasgow Polyomics) **Joe Wandy** (SoCS; Glasgow Polyomics), **Karl Burgess** (Glasgow Polyomics), **Mike Barrett** (Glasgow Polyomics)





Network view

Mass2Motifs

