RMassBank

eawag aquatic research

RMassBank: Automatic Recalibration and Processing of Tandem HR-MS Spectra for MassBank

Eawag: Swiss Federal Institute of Aquatic Science and Technology

Michael Stravs, Emma Schymanski, Heinz Singer and Juliane Hollender

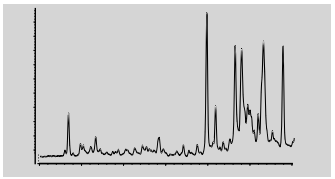
Eawag – Department of Environmental Chemistry

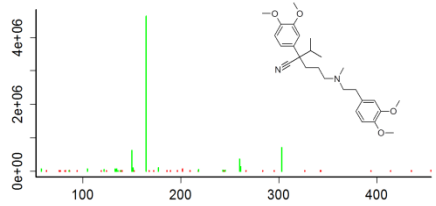
Dübendorf (Zurich), Switzerland

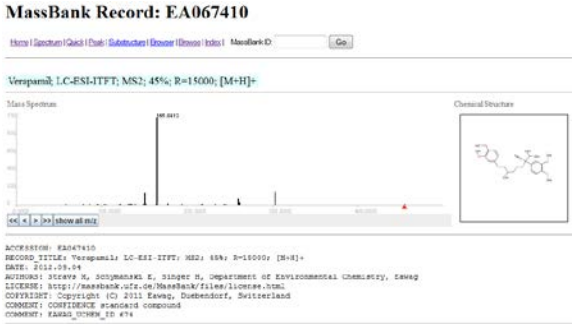
RMassBank Questions: massbank@eawag.ch

Overview

- Building a spectral library: challenges
  - clean-up
  - annotation
- RMassBank workflow
  - data processing
  - record creation
- Results and examples







Building a spectral library - Challenges

Processing spectra efficiently

- Manual entry
  - tedious
  - Available tools not suited for mass processing
- Metadata collection / annotation
- Reproducible procedure

Ensuring high quality

- Noise removal / clean-up
- Mass accuracy
- Curation

Building a spectral library

Previous approaches to «clean-up» and annotation*

- Simple noise cutoff – e.g. WA001201
  - Cut-off at 5 ‰ (MassBank reporting can go down to 1 ‰)
- Minimal or no processing – e.g. CE000143
  - Many low intensity noise peaks, including peaks above $[M+H]^+$
- Minimal annotation – e.g. JEL00007

The result

- Inconsistent and varying quality of spectra in MassBank

  → “spectral dump” **

* These examples are demonstrations only and no offense intended!

** Thanks to Oliver Fiehn for this apt description!

Spectral Clean-up

MassBank Record: WA001201

```
MS$DATA_PROCESSING: FIND_PEAK ignore rel.int. < 5

PK$NUM_PEAK: 19
PK$PEAK: m/z int. rel.int.
  102 63 63
  105 8 8
  130 133 133
  131 8 8
  139 20 20
  162 12 12
  163 12 12
  166 834 834
  167 59 59
  178 12 12
  180 63 63
  182 24 24
  184 999 999
  185 63 63
```

* These examples are demonstrations only and no offense intended!

Spectral Clean-up

MassBank Record: CE000143

```
MS$FOCUSED_ION: PRECURSOR_M/Z 268.10404
MS$FOCUSED_ION: PRECURSOR_TYPE [M+H]+

PK$NUM_PEAK: 20
PK$PEAK: m/z int. rel.int.
  76.423012 1164.474609 2
  79.144173 972.286438 2
  88.860359 1310.202271 2
  106.183556 1195.565674 2
  135.887161 3993.299072 7
  135.992874 3979.442627 7
  136.034454 2340.972656 4
  136.041092 1992.728271 4
  136.061371 568236.125 999
  136.08461 4257.030762 7
  136.130478 2980.912598 5
  136.228577 1192.598145 2
  136.234634 1166.915161 2
  197.613419 1193.385498 2
  203.694138 1189.134888 2
  219.079987 21971.183594 39
  237.09053 180033.921875 317
  292.21759 1154.94812 2
  293.092346 1334.223755 2
  293.587189 1281.828003 2
//
```

These examples are demonstrations only and no offense intended!

Spectral Clean-up – Orbitrap spectra

Garbage in = garbage out!

Measurement artefacts include

- Systematic increase in ppm error with low *m/z* values
  - 5 ppm accuracy => 15 ppm at *m/z* < 100
- Satellite or shoulder peaks – result of FT instrument processing
- Consistent electronic or measurement noise peaks
- Noise peaks at a ~fixed level hiding real peaks of lower intensity
  - A strict noise cut-off will result in a loss of information

High mass accuracy standard spectra have advantages:

- Meringer et al. (2011) showed a missing subformula indicates instrument noise or interfering peaks
  - Use subformula assignment to perform spectral clean-up!

Meringer et al. 2011, *MATCH Commun. Math. Comput. Chem*. 65, 259-290.

Annotation

Mandatory

- Name, structure, InChI, SMILES
- Minimal analytical information
  - MS instrument
  - MS polarity and spectrum type

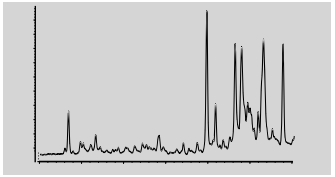
Optional and useful!

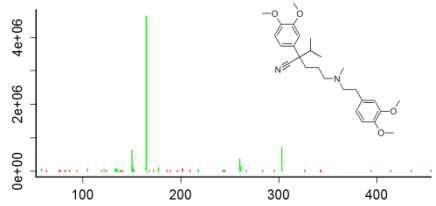
- Links to databases
  - CAS, KEGG, ChEBI, PubChem…
- Analytical conditions
  - Chromatography, RT […]
  - MS2: precursor, conditions […]
- Peak annotation
  - chemical formula, substructure

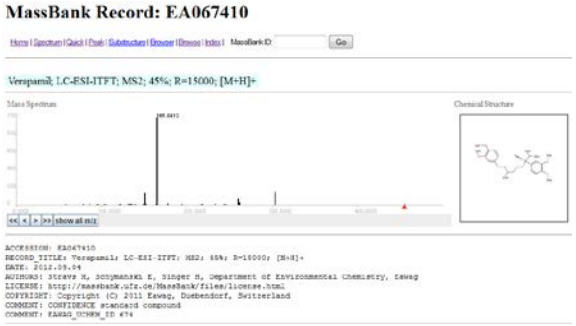
more is better, but manual work is tedious

Building a spectral library - RMassBank

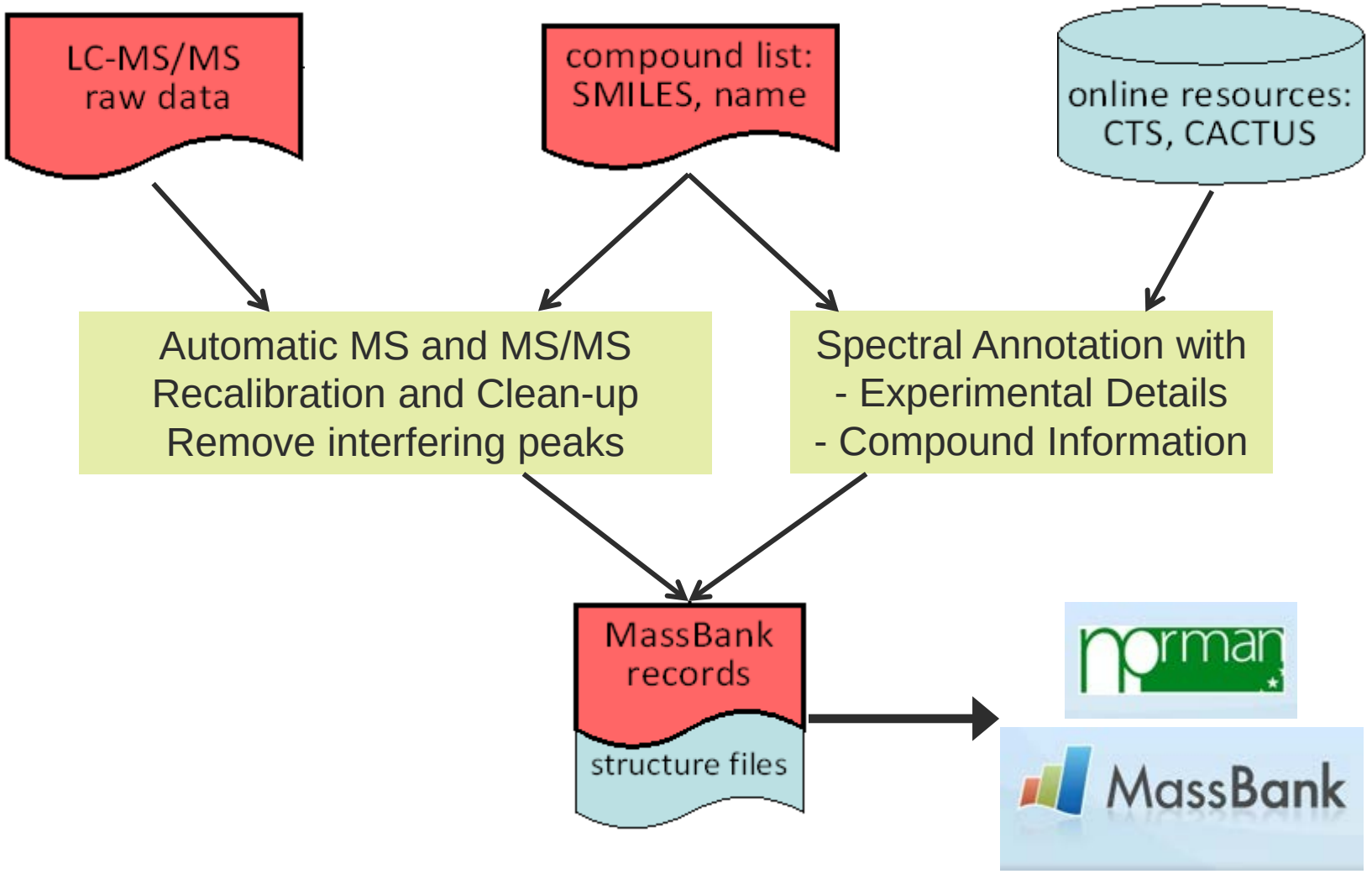
- **Automate what can be automated**
- **Ensure high quality**
- Workflow from raw LC-MS file to annotated spectrum
  - spectrum extraction
  - processing and clean-up
  - automated annotation using internet resources





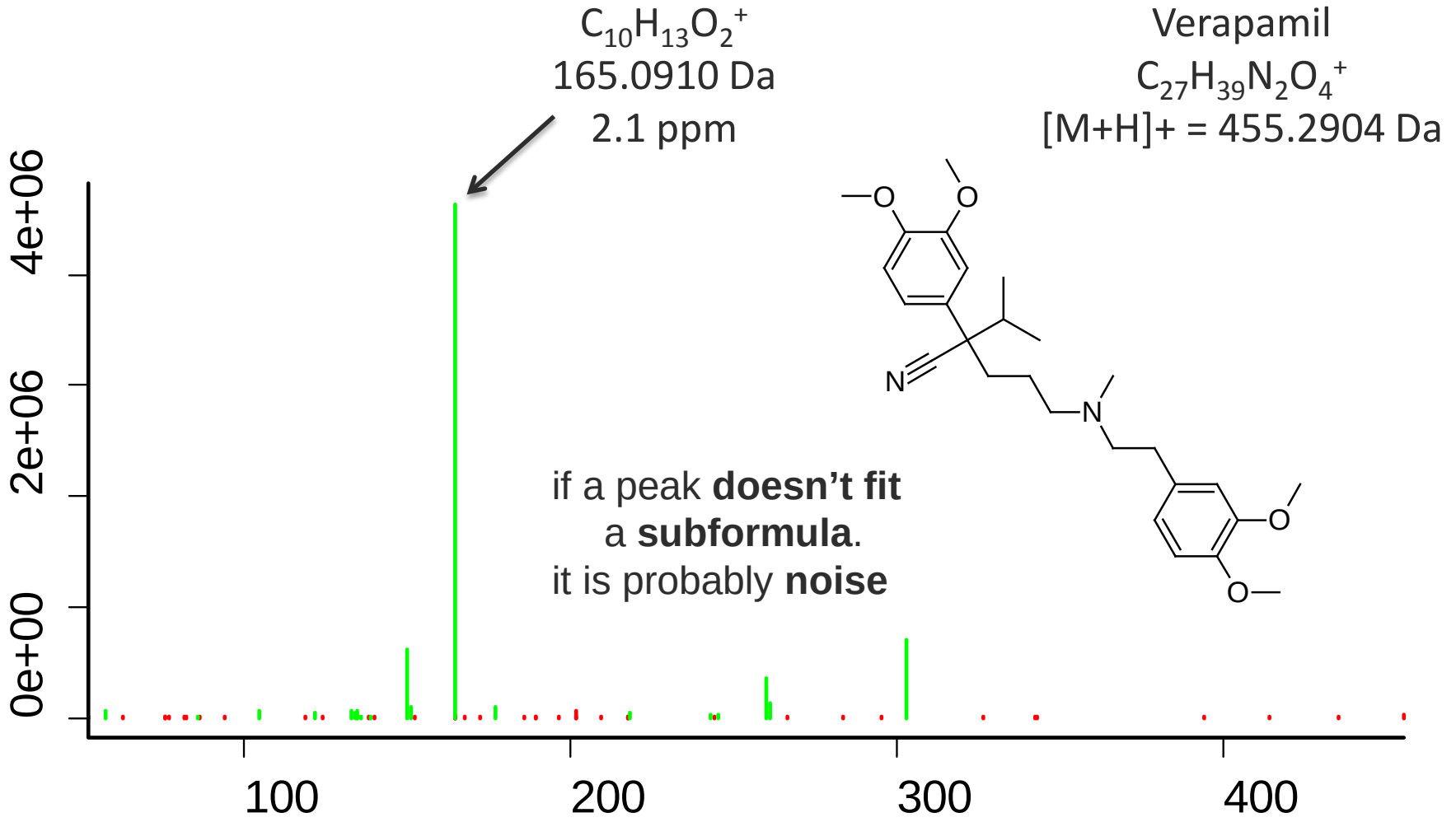


RMassBank Workflow – Simple Form



Stravs, Schymanski, Singer and Hollender, 2012,
Journal of Mass Spectrometry, 2013, 48, 89–99. DOI: 10.1002/jms.3131

RMassBank – Example Clean-up



Spectral Annotation

Matching Spectrum and Compound Information

User needs to contribute a bare minimum of information

- Only the user knows what compound has been measured
  - At least one form of unambiguous compound identifier is required
  - e.g. internal ID, name, SMILES and retention time
- Measurement parameters / methods / settings are relatively consistent
  - These can be added in batch form, not individually

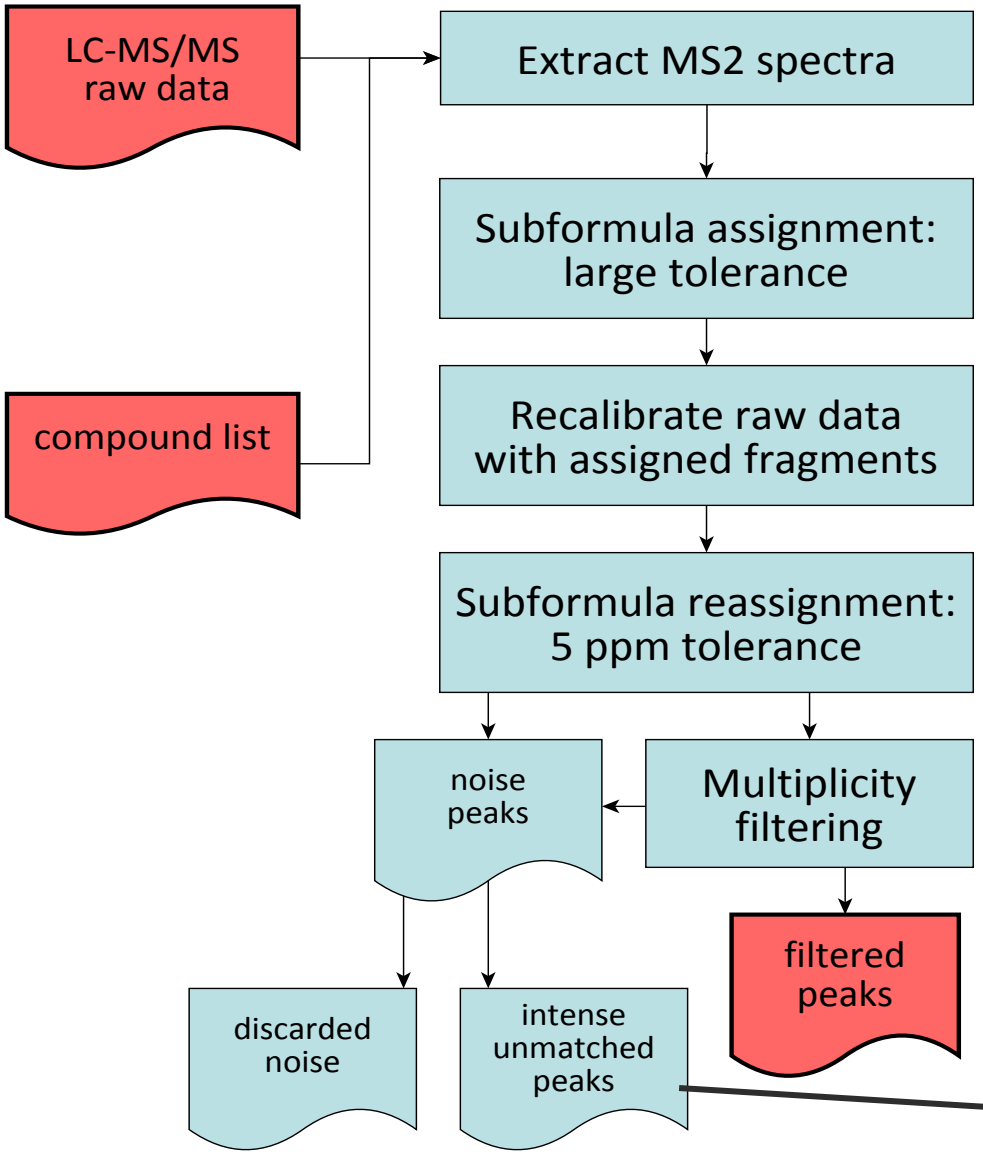
Internet Services: Let search engines do the work for you!

- CACTUS Chemical Identifier Resolver[1]
  - SMILES (c1ccccc1) to InChI Key (UHOVQNZJYSORNB-UHFFFAOYSA-N)
- Chemical Translation Service (CTS)[2] to do the rest
  - Names, CAS #, InChI and Identifiers (IDs, if available): PubChem CID, ChemSpider, ChEBI, HMDB, KEGG, LipidMaps

[1] http://cactus.nci.nih.gov/chemical/structure

[2] http://cts.fiehnlab.ucdavis.edu/

RMassBank – Clean-up and Recalibration



Automated extraction by accurate m/z and RT.
FT-satellite removal (± 0.5 Da; I < 5 % of peak)

Elements of molecular formula ± adduct
ppm ≤ 15 (m/z < 120) or ≤ 10 (m/z > 120)

Use only peaks with unique formula calculated

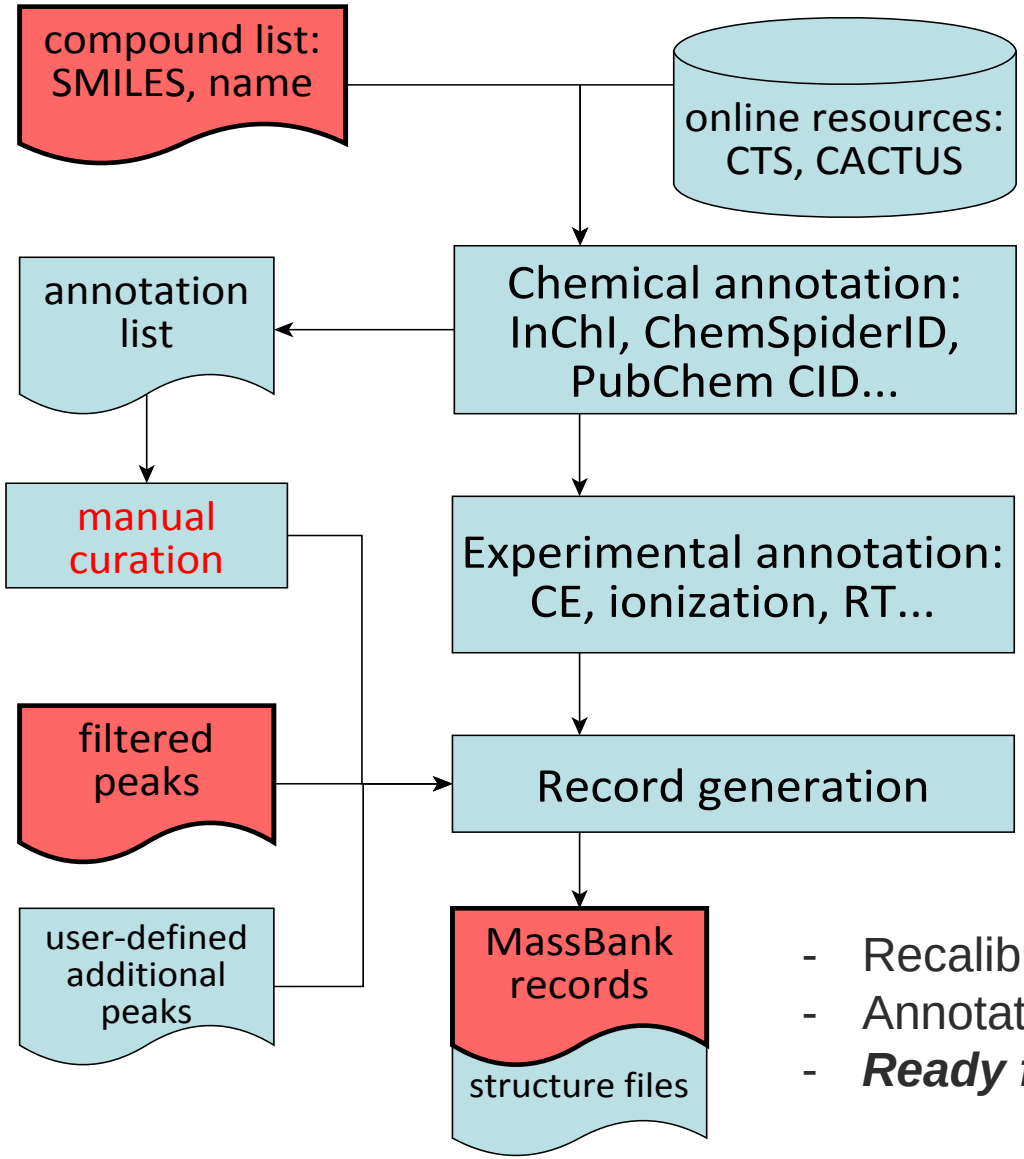
Formulas recalculated post calibration.
For "Fail Peaks", **recalculate with +2N & O**

Peaks are retained if they occur ≥2x per compound (in e.g. 14 recorded spectra)

Filtered peaks go through to Records

Intense unmatched peaks for manual check

RMassBank – Spectrum Annotation



- Recalibrated, cleaned-up
- Annotated
- ***Ready for upload!***

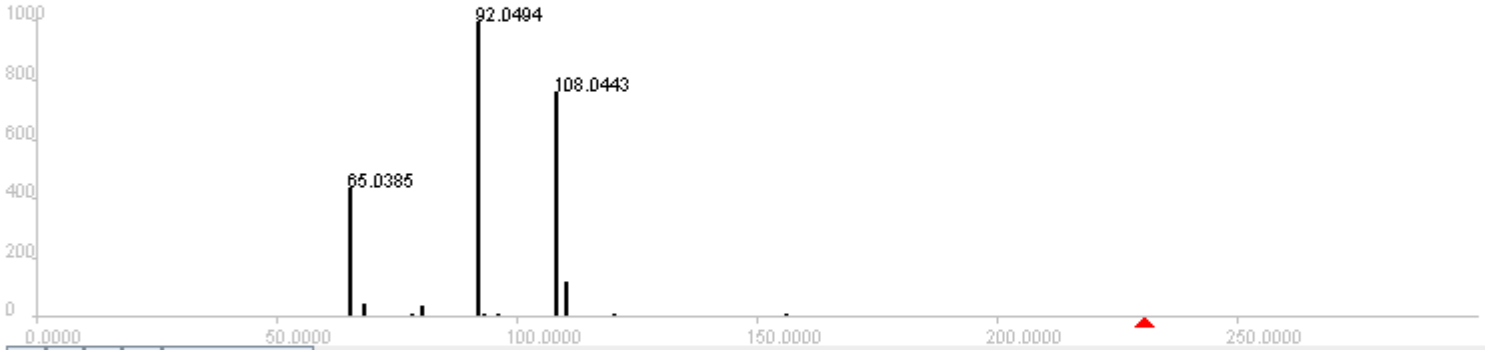
Example MassBank Record

MassBank Record: EA015612

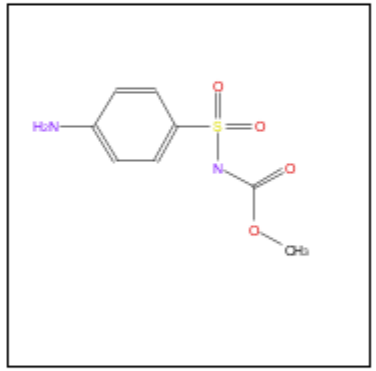
Home | Spectrum | Quick | Peak | Substructure | Browser | Browse | Index | MassBank ID: Go

Asulam; LC-ESI-ITFT; MS2; 75%; R=15000; [M+H]+

Mass Spectrum



Chemical Structure

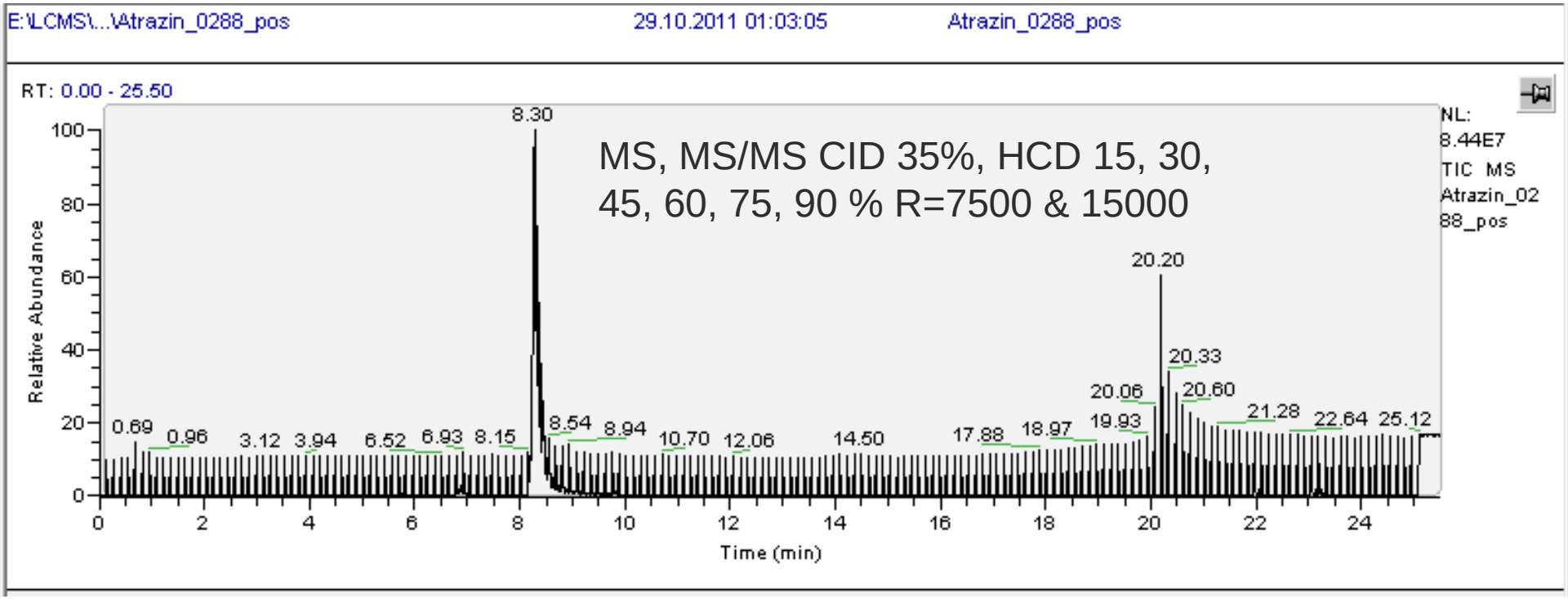



```
ACCESSION: EA015612
RECORD_TITLE: Asulam; LC-ESI-ITFT; MS2; 75%; R=15000; [M+H]+
DATE: 2012.03.16
AUTHORS: Stravs M, Schymanski E, Singer H, Department of Environmental Chemistry, Eawag
LICENSE: http://massbank.ufz.de/MassBank/files/license.html
COPYRIGHT: Copyright (C) 2011 Eawag, Duebendorf, Switzerland
COMMENT: CONFIDENCE standard compound
COMMENT: EAWAG_UCHEM_ID 156
```

RMassBank with 70 Eawag Pesticide Spectra

Experimental

- Individual Injection; routine chromatography; Orbitrap XL with ESI + / -
- Daily vendor-recommended calibration



Screened for $[M+H]^+$ precursor within RT ± 0.3 min
MS/MS retrieved from MS with highest intensity

RMassBank with 70 Eawag Pesticide Spectra

Processing: The Numbers

- 68 of 70 pesticides with sufficient $[M+H]^+$ for processing
- 55,594 peaks present following satellite removal
  - 14,699 with at least one subformula post-calibration
  - 13,305 of these present in two or more spectra per compound
- 454 peaks with subformula assigned
  - 256 of these occurred at least twice per compound => *reproducible*
- Only 44 peaks remained for “manual inspection”
- No difference observed between spectra with different resolutions

~76 % of peaks are noise!

Formation of N_2 and H_2O adducts is relevant in MS/MS!

Additional modes (results not shown here)

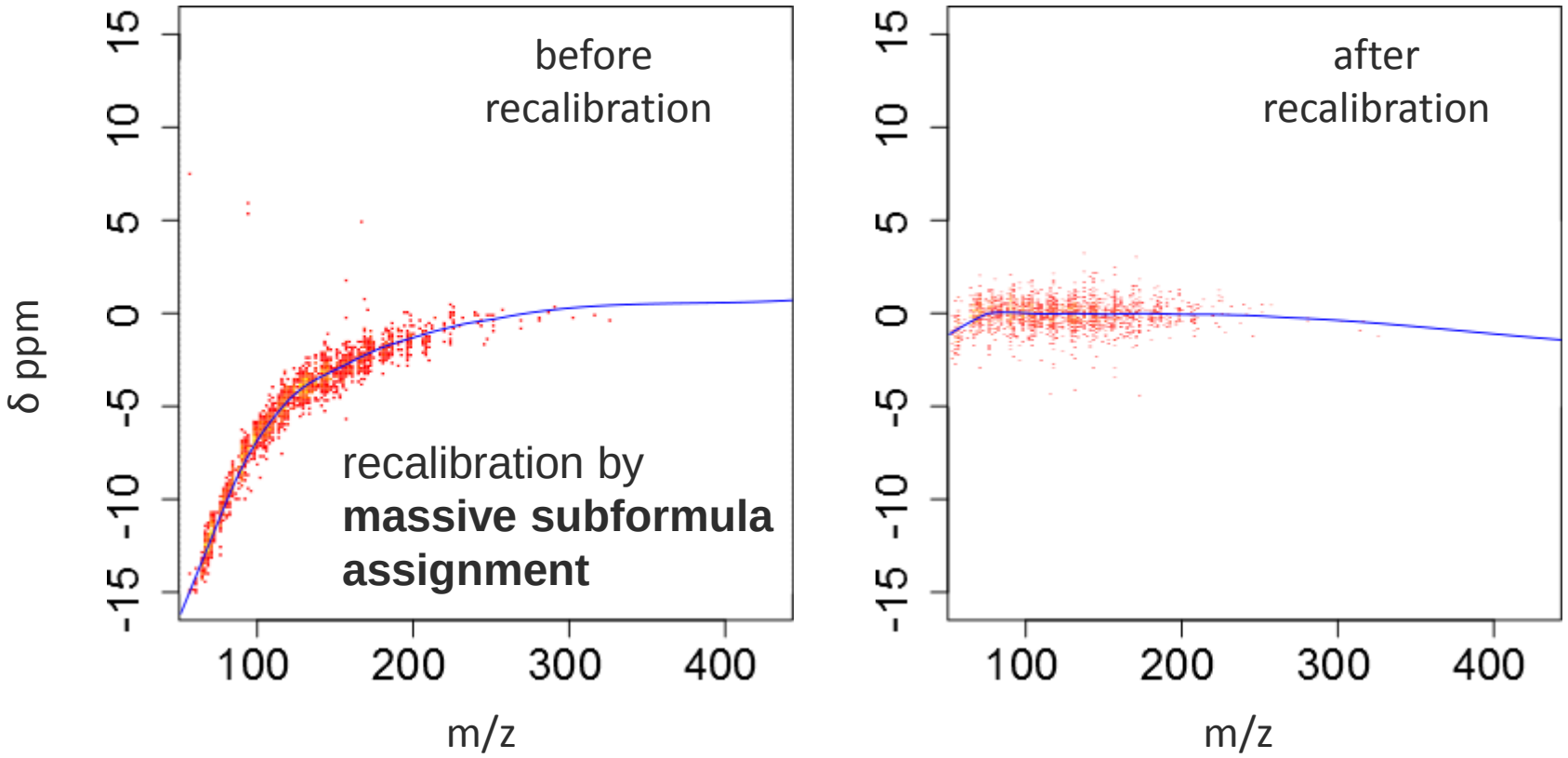
- M^+, $[M+Na]^+$, $[M-H]^-$, M^-, $[M+FA]^-$,

Effect of Recalibration

- Shown in the next few slides…

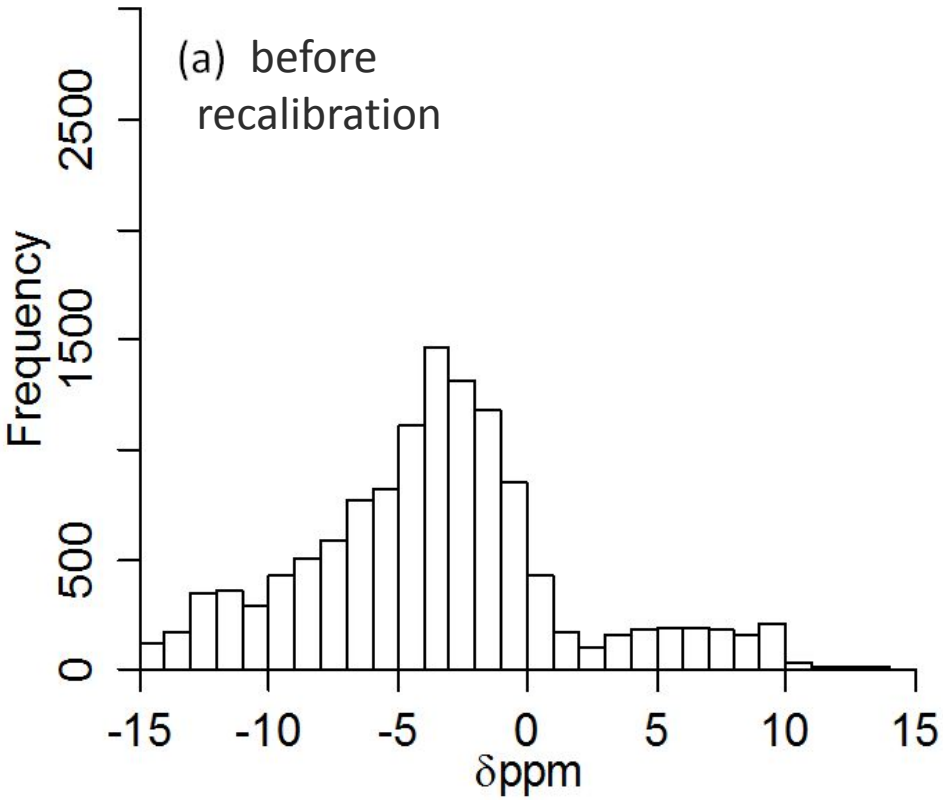
RMassBank with 70 Eawag Pesticide Spectra

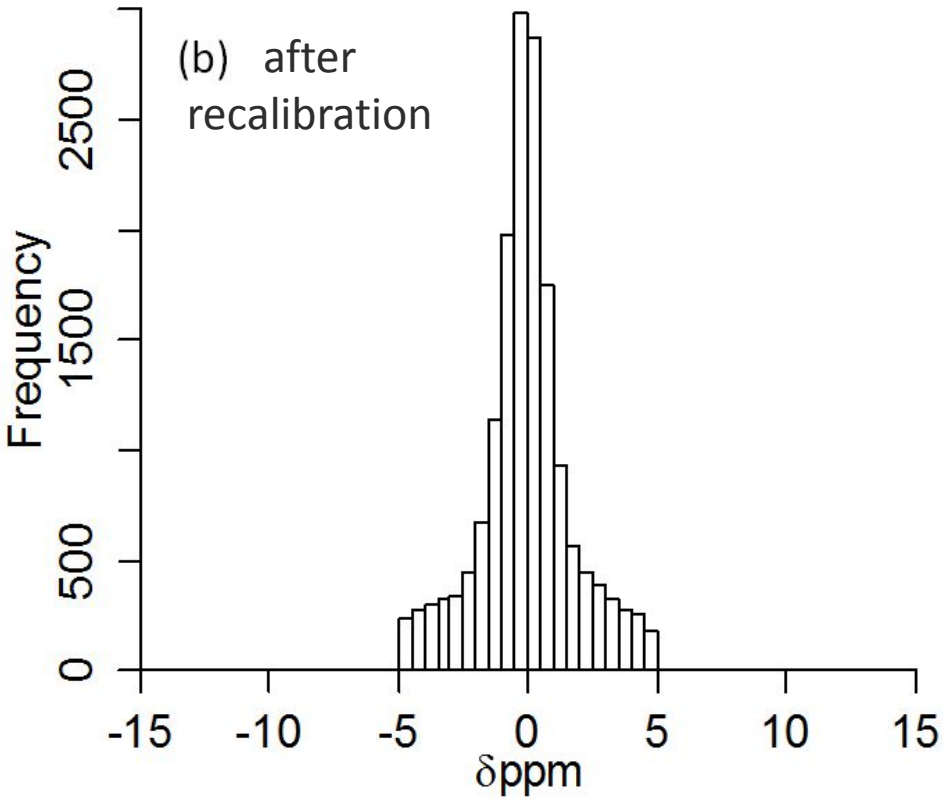
Recalibration Curve: Relative mass deviation over *m/z*



RMassBank with 70 Eawag Pesticide Spectra

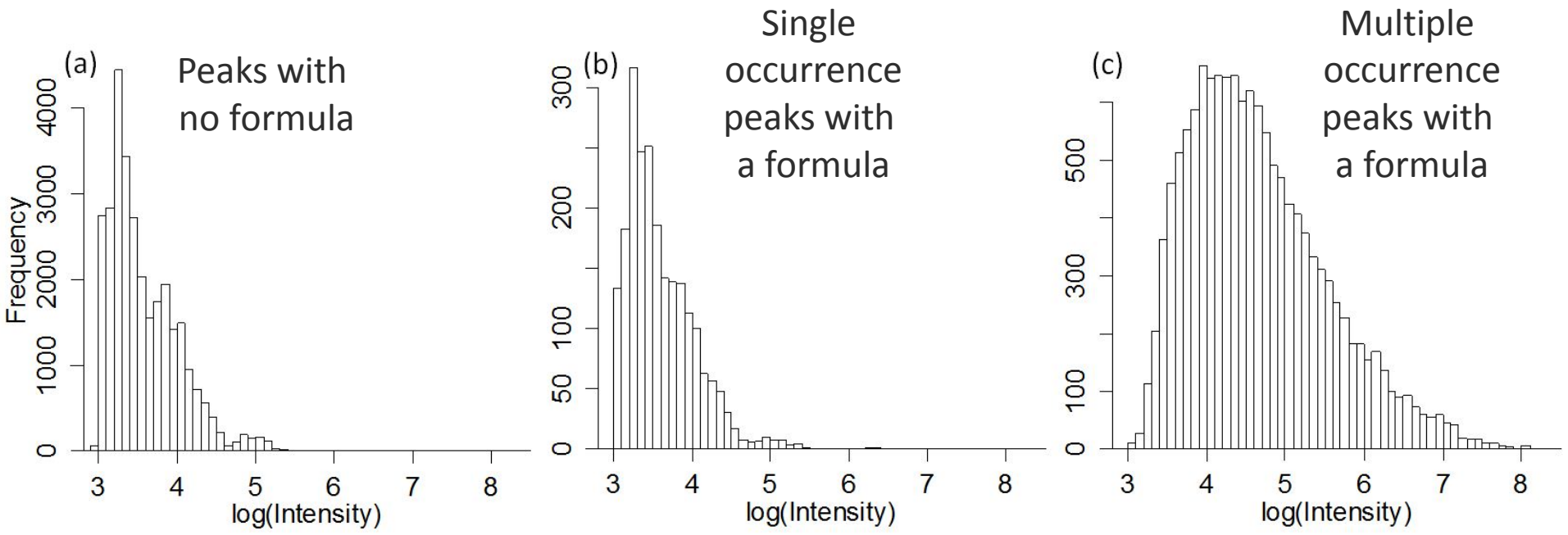
Recalibration: Relative Mass Deviation Distribution





RMassBank with 70 Eawag Pesticide Spectra

Frequency of Occurrence of Peaks by Intensity: Multiplicity Filtering



if a peak **doesn't occur at least twice**
(in 2x7 spectra) it is probably **noise**

RMassBank with 70 Eawag Pesticide Spectra

944 MS/MS spectra from $[M+H]^+$ of 70 pesticides

- www.massbank.jp
- www.massbank.eu

Total Number of “RMassBank Spectra”

RMassBank records in NORMAN MassBank:

- 6,106 records (364 compounds) from Eawag Orbitrap XL
- 1,030 records (216 compounds) from UFZ Orbitrap XL
- Q-Exactive spectra (not yet on NORMAN MassBank)

RMassBank «advanced usage»

For R workflow developers (xcms, nontarget, CAMERA…)

- dd-MS2 processing
- recalibration, filtering
- Interface to CTS, CACTUS
- Chemical formula calculations (C6H5 + H2O = C6H7O)
- fragment formula assignment
- (database search)

… if you are fluent in R ☺

Conclusions: RMassBank

The RMassBank Workflow

- Reduces much manual work associated with bulk creation of many records
- Creates high quality MS/MS spectra
- Annotation with formula adds value to the spectra
- Works very well for the spectra it was developed on (Orbitrap)
- BUT: Every mass spectrometer is different:
  - Processing and measurement steps will probably need adjusting

Benefit for Contributors

- We have learnt a lot about our spectra and compounds (e.g. MS/MS adducts!)
- MassBank is being used within our department
- If you want to know more about what recalibration can do for your data:
  - Stravs et al. 2012, *J. Mass Spectrom.*, DOI: 10.1002/jms.3131

Acknowledgements

- Coauthors: Emma Schymanski, Heinz Singer, Juliane Hollender
- R & MassBank help: Steffen Neumann, Michael Gerlich, Carsten Kuhl, (IPB)
- Discussions: Matthias Ruff, Martin Loos (Eawag); Tobias Schulze, Martin Krauss, Werner Brack (UFZ)
- Programs: Markus Meringer (MOLGEN-MS/MS), Florian Rasche (Sirius), Miguel Rojas-Cherto & Egon Willighagen (MEF)
- MassBank & Naming Rights: Prof. Takaaki Nishioka
- NORMAN Association

RMassBank

eawag aquatic research

Any Questions?

Michael Stravs, Emma Schymanski, Heinz Singer and Juliane Hollender

massbank@eawag.ch (ES, MS)
stravsmi@eawag.ch
emma.schymanski@eawag.ch

DOI: 10.1002/jms.3131
All Details Contained Within!