RMassBank

eawag aquatic research

# RMassBank: Run-through the Principles and Workflow in R

**Emma Schymanski**

Michael Stravs, Heinz Singer & Juliane Hollender

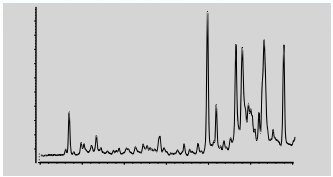
*Eawag, Dübendorf, Switzerland*

Steffen Neumann, Erik Müller: *IPB Halle, Germany*

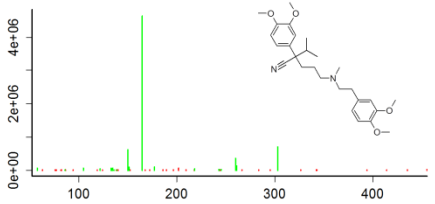
Tobias Schulze: *UFZ Leipzig, Germany*

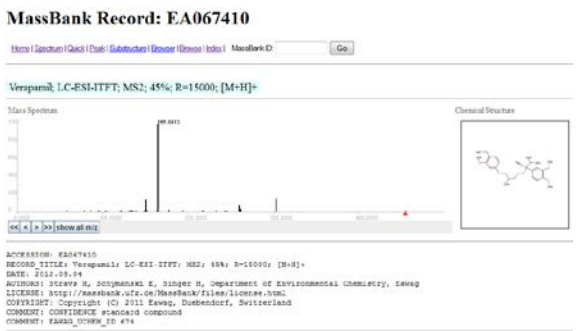
**RMassBank Questions: massbank@eawag.ch**

# Demonstration Overview



- Install (and update) RMassBank and associated programs
- Run RMassBank for trial data
- Get a basic understanding for the workflow
- Understand the manual checking required
  - “Fail peaks” – checking these in raw data
  - Automatic annotation with CTS – editing this data
- Generate records with RMassBank
- Upload these records to MassBank

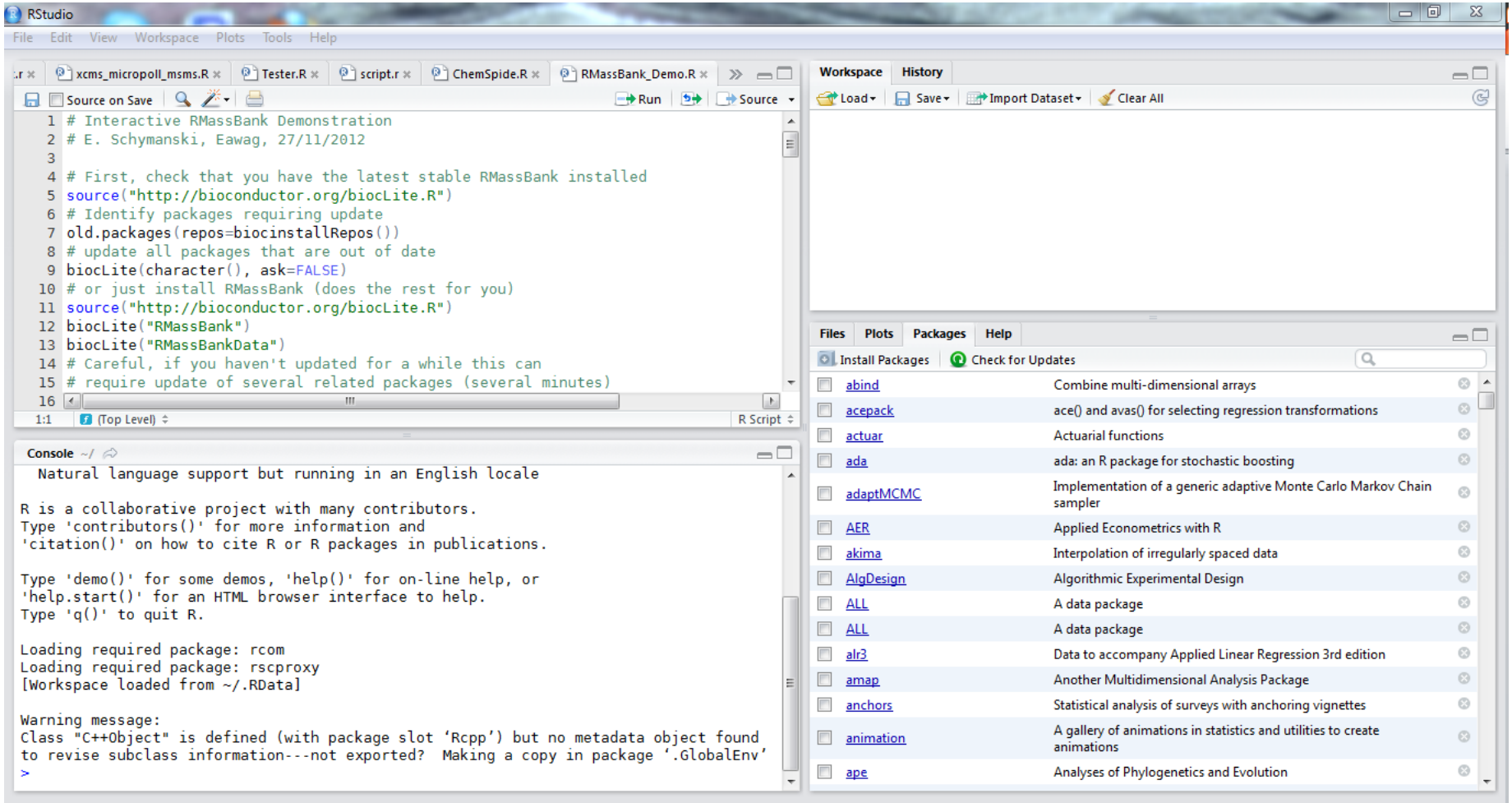


# Starting RMassBank

Open up RStudio (or R console)

- And open up “RMassBank_Demo.R”



# Installing / Updating RMassBank

To check if you need to update RMassBank:

Source on Save | Run | Source

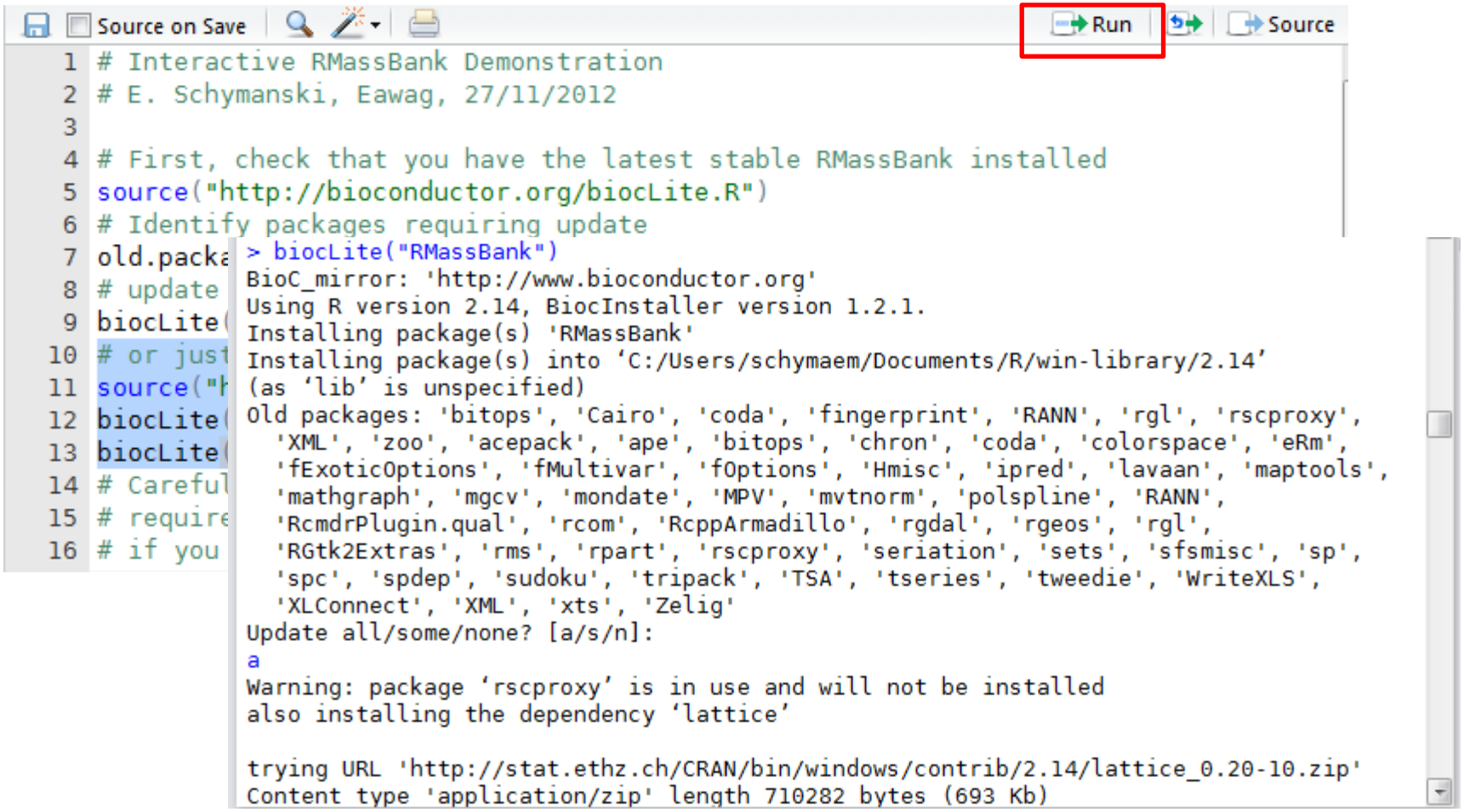
```
1  # Interactive RMassBank Demonstration
2  # E. Schymanski, Eawag, 27/11/2012
3
4  # First, check that you have the latest stable RMassBank installed
5  source("http://bioconductor.org/biocLite.R")
6  # Identify packages requiring update
7  old.packages(repos=biocinstallRepos())
8  # update all packages that are out of date
9  biocLite(character(), ask=FALSE)
10 # or just install RMassBank (does the rest for you)
11 source("http
12 biocLite("RM
13 biocLite("RM
14 # Careful, i
15 # require up
16 # if you get
```

Console ~/

```
> source("http://bioconductor.org/biocLite.R")
BiocInstaller version 1.2.1, ?biocLite for help
> old.packages(repos=biocinstallRepos())
                Package           LibPath
bitops          "bitops"          "C:/Users/schymaem/Documents/R/win-library/2.14"
Cairo           "Cairo"           "C:/Users/schymaem/Documents/R/win-library/2.14"
coda            "coda"            "C:/Users/schymaem/Documents/R/win-library/2.14"
fingerprint     "fingerprint"     "C:/Users/schymaem/Documents/R/win-library/2.14"
RANN            "RANN"            "C:/Users/schymaem/Documents/R/win-library/2.14"
rgl             "rgl"             "C:/Users/schymaem/Documents/R/win-library/2.14"
rscproxy        "rscproxy"        "C:/Users/schymaem/Documents/R/win-library/2.14"
XML             "XML"             "C:/Users/schymaem/Documents/R/win-library/2.14"
zoo             "zoo"             "C:/Users/schymaem/Documents/R/win-library/2.14"
acepack         "acepack"         "C:/Program Files/R/R-2.14.1/library"
ape             "ape"             "C:/Program Files/R/R-2.14.1/library"
```

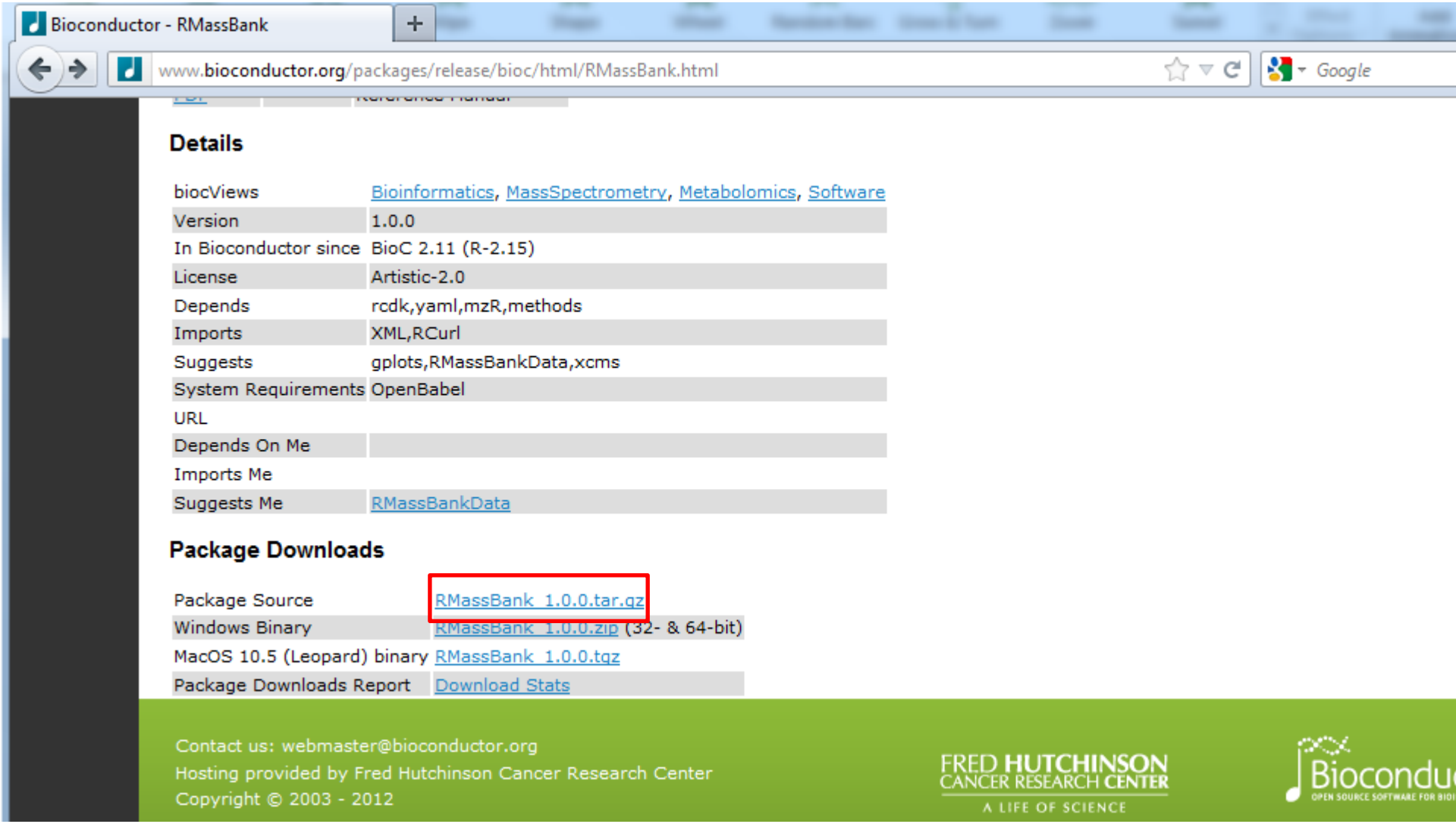
# Installing / Updating RMassBank

To install RMassBank:



# Installing from source – if all else fails

(this means hand-installing other packages, avoid if possible!)



# Installing from source – if all else fails

(this means hand-installing other packages, avoid if possible!)

```
#If this doesn't work, try installing from source
# http://www.bioconductor.org/packages/release/bioc/html/RMassBank.html
# download Package
# http://www.biocon
install.packages("f
                 re
# http://www.biocon
# download package
# http://www.biocon
install.packages("f
                 re
```

```
> install.packages("C:/DATA/RMassBank/RMassBank_Versions/RMassBank_1.0.0.tar.gz",
repos=NULL, type="source")
Installing package(s) into 'C:/Users/schymaem/Documents/R/win-library/2.14'
(as 'lib' is unspecified)
* installing *source* package 'RMassBank' ...
** R
** inst
** preparing package for lazy loading
Warning: package 'rcdk' was built under R version 2.14.2
Warning: package 'rcdklibs' was built under R version 2.14.2
Warning: package 'fingerprint' was built under R version 2.14.2
Warning: package 'iterators' was built under R version 2.14.2
Warning: package 'yaml' was built under R version 2.14.2
Warning: package 'Rcpp' was built under R version 2.14.2
** help
*** installing help indices
** building package indices ...
*** tangling vignette sources ...
   'RMassBank.Rnw'
** testing if installed package can be loaded
Warning messages:
1: package 'rcdk' was built under R version 2.14.2
2: package 'rcdklibs' was built under R version 2.14.2
3: package 'fingerprint' was built under R version 2.14.2
4: package 'iterators' was built under R version 2.14.2
5: package 'yaml' was built under R version 2.14.2
6: package 'Rcpp' was built under R version 2.14.2

* DONE (RMassBank)
>
```

# Loading RMassBank

```
30 # Once you have got through all that, you need to load the packages:
31 library("RMassBank")
32 library("RMassBankData")
33
```

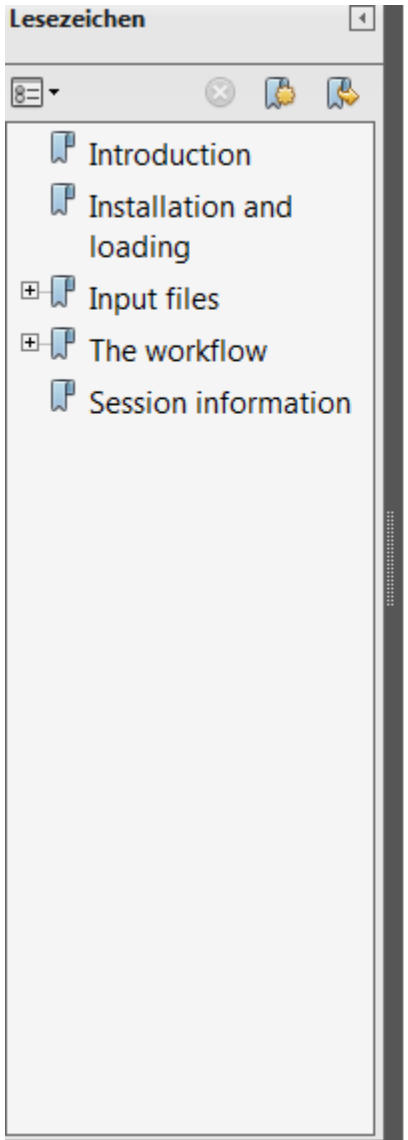
32:25 (Top Level) R Script

Console ~/

```
> # Once you have got through all that, you need to load the packages:
> library("RMassBank")
Loading required package: mzR
Loading required package: Rcpp
Loading required package: XML
Loading required package: RCurl
Loading required package: bitops
Loading required package: zoo

Attaching package: 'zoo'

The following object(s) are masked from 'package:base':

    as.Date, as.Date.numeric

Loading required package: iterators
Warning messages:
1: package 'Rcpp' was built under R version 2.14.2
2: package 'XML' was built under R version 2.14.2
3: package 'bitops' was built under R version 2.14.2
4: package 'zoo' was built under R version 2.14.2
5: package 'rcdk' was built under R version 2.14.2
6: package 'rcdklibs' was built under R version 2.14.2
7: package 'fingerprint' was built under R version 2.14.2
8: package 'iterators' was built under R version 2.14.2
> library("RMassBankData")
Warning message:
package 'RMassBankData' was built under R version 2.14.2
>
```

# Finding out more about RMassBank

vignette("RMassBank")




RMassBank: The workflow by example

Michael Stravs

October 2, 2012

We'll kind-of work through this today…

# Before getting started:

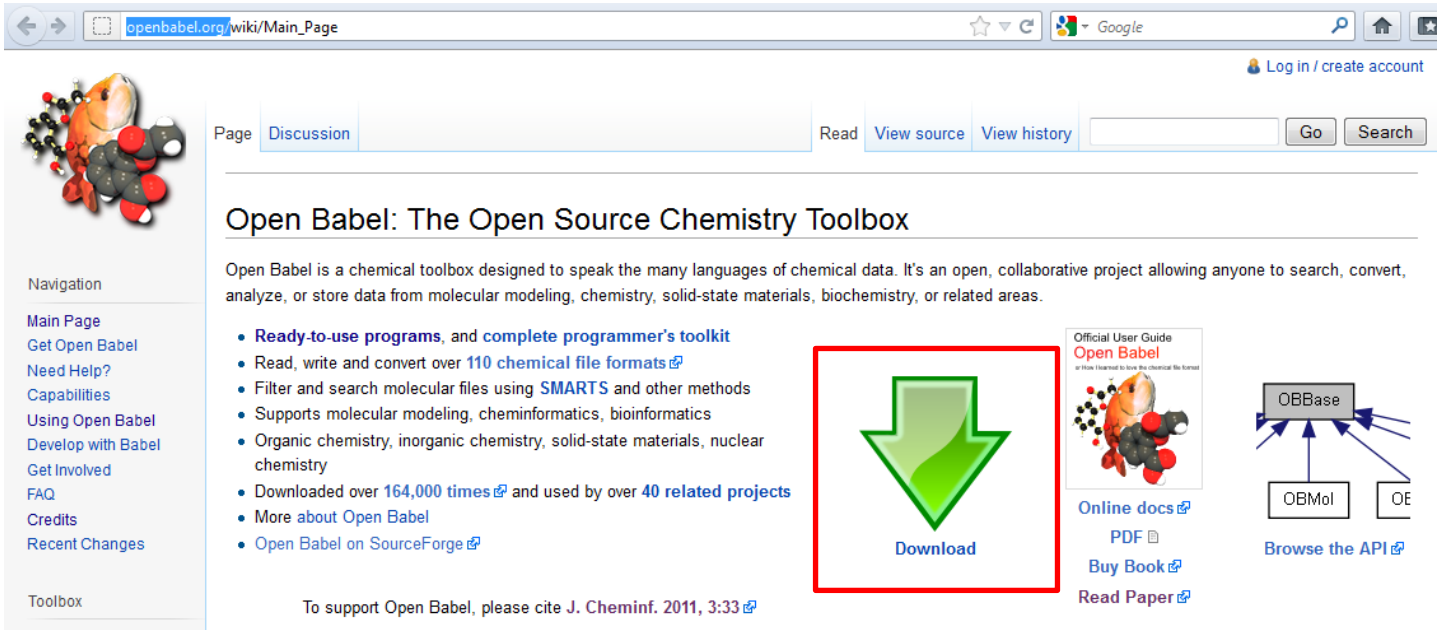
Install / Check OpenBabel

If you think you already have this, check e.g. (for Windows….)

- All Programs => OpenBabel 2.X.X
- C:\Program Files\OpenBabel-2.3.0

If you don't have this, please download (not *required* but it is *recommended*)

- http://openbabel.org/

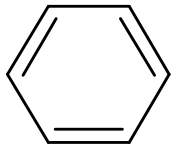


# Before getting started:

Install / Check OpenBabel

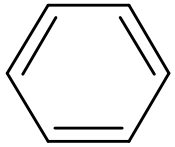
Why OpenBabel?

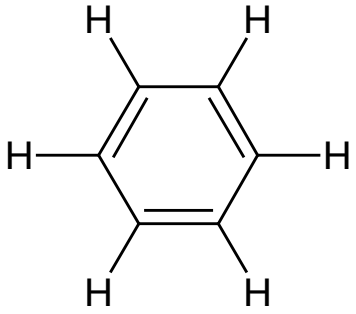
- Convert SMILES (c1ccccc1) to structure
- C:\Program Files\OpenBabel-2.3.0



NOTE:

- By default, RMassBank will work without OpenBabel and has a back-up solution using CACTUS
- But, structures are less visually appealing





# Before getting started:

Install / Check Proteowizard

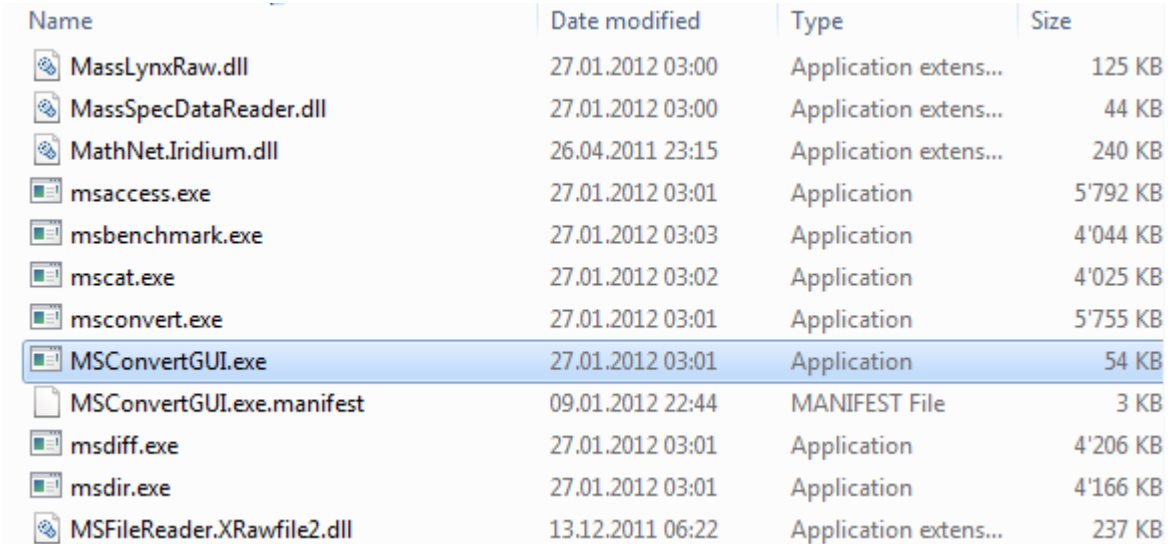
Why Proteowizard?

- Convert raw data files to an open format (we chose mzML)
- Is one of most established and supported converters

**NOTE: If you will only ever use RMassBank once (i.e. today)**
**Don't bother installing! The test data is already converted!**

Check to see if you have it

- E.g. for me: C:\DATA\Program_Info\pwiz
- You want "MSConvertGUI.exe"

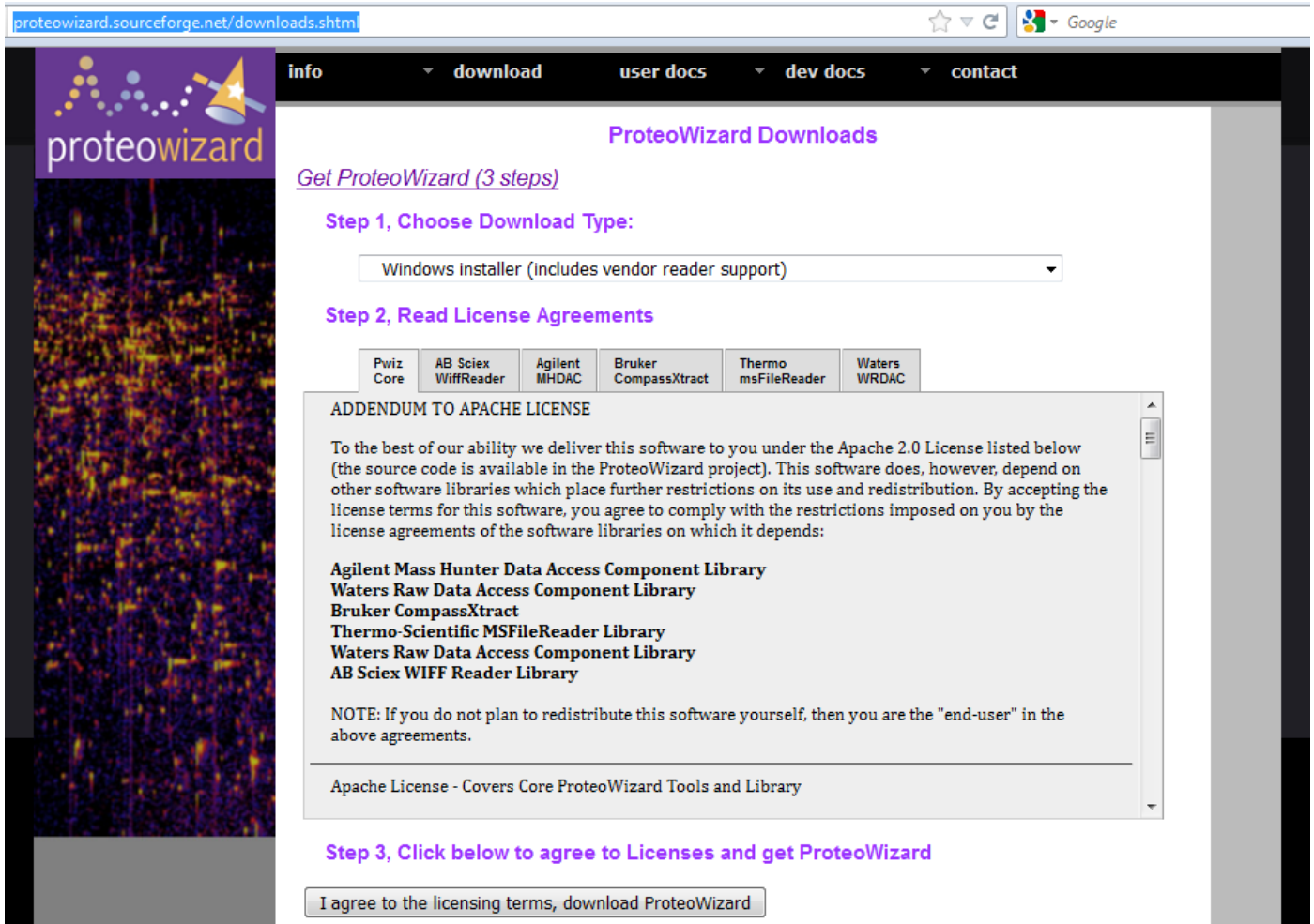


| Name | Date modified | Type | Size |
|---|---|---|---|
| MassLynxRaw.dll | 27.01.2012 03:00 | Application extens... | 125 KB |
| MassSpecDataReader.dll | 27.01.2012 03:00 | Application extens... | 44 KB |
| MathNet.Iridium.dll | 26.04.2011 23:15 | Application extens... | 240 KB |
| msaccess.exe | 27.01.2012 03:01 | Application | 5'792 KB |
| msbenchmark.exe | 27.01.2012 03:03 | Application | 4'044 KB |
| mscat.exe | 27.01.2012 03:02 | Application | 4'025 KB |
| msconvert.exe | 27.01.2012 03:01 | Application | 5'755 KB |
| MSConvertGUI.exe | 27.01.2012 03:01 | Application | 54 KB |
| MSConvertGUI.exe.manifest | 09.01.2012 22:44 | MANIFEST File | 3 KB |
| msdiff.exe | 27.01.2012 03:01 | Application | 4'206 KB |
| msdir.exe | 27.01.2012 03:01 | Application | 4'166 KB |
| MSFileReader.XRawfile2.dll | 13.12.2011 06:22 | Application extens... | 237 KB |

# Before getting started:

Install / Check Proteowizard

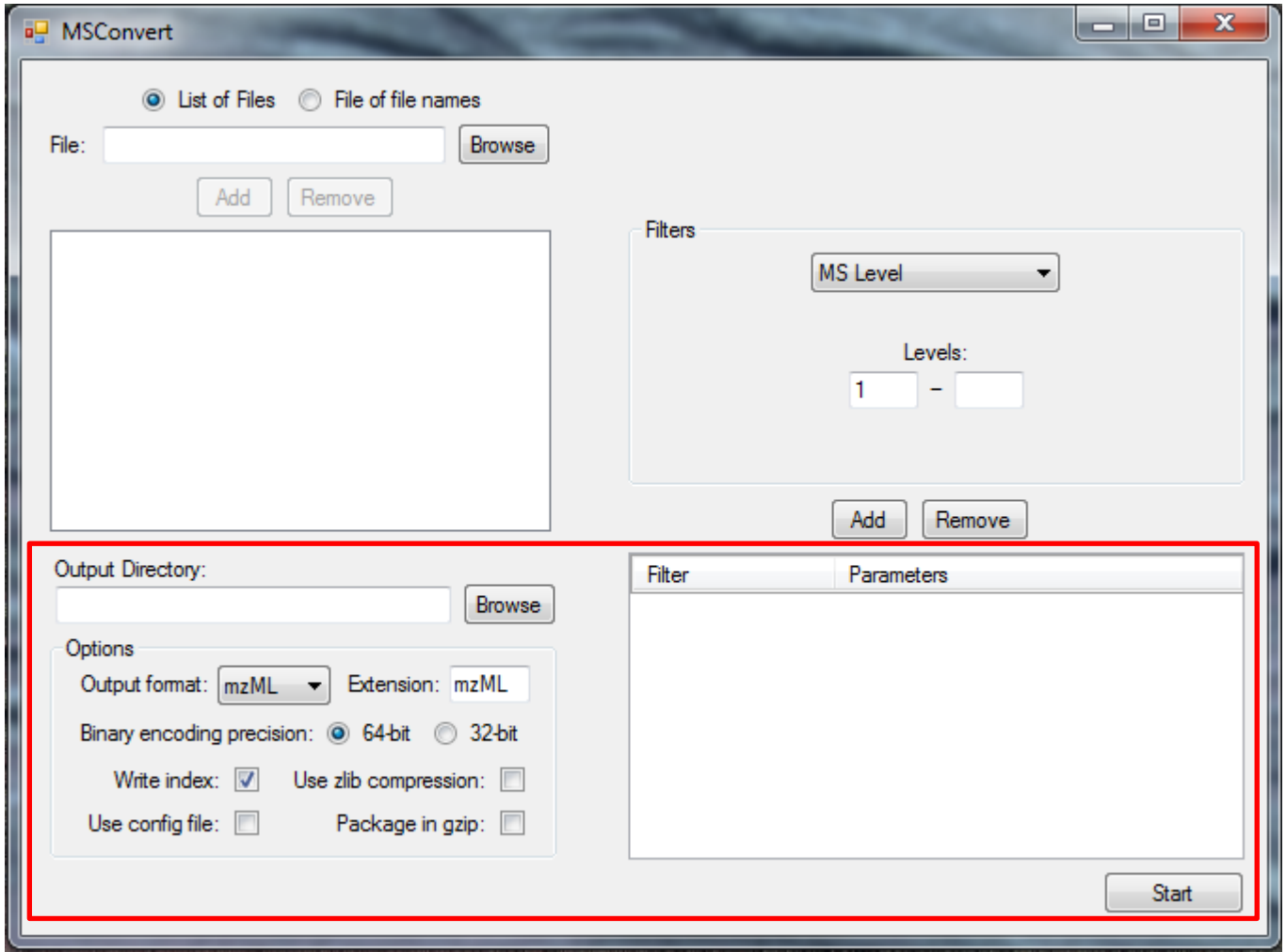
No Proteowizard (pwiz)?: http://proteowizard.sourceforge.net/downloads.shtml



# Before getting started:

Install / Check Proteowizard

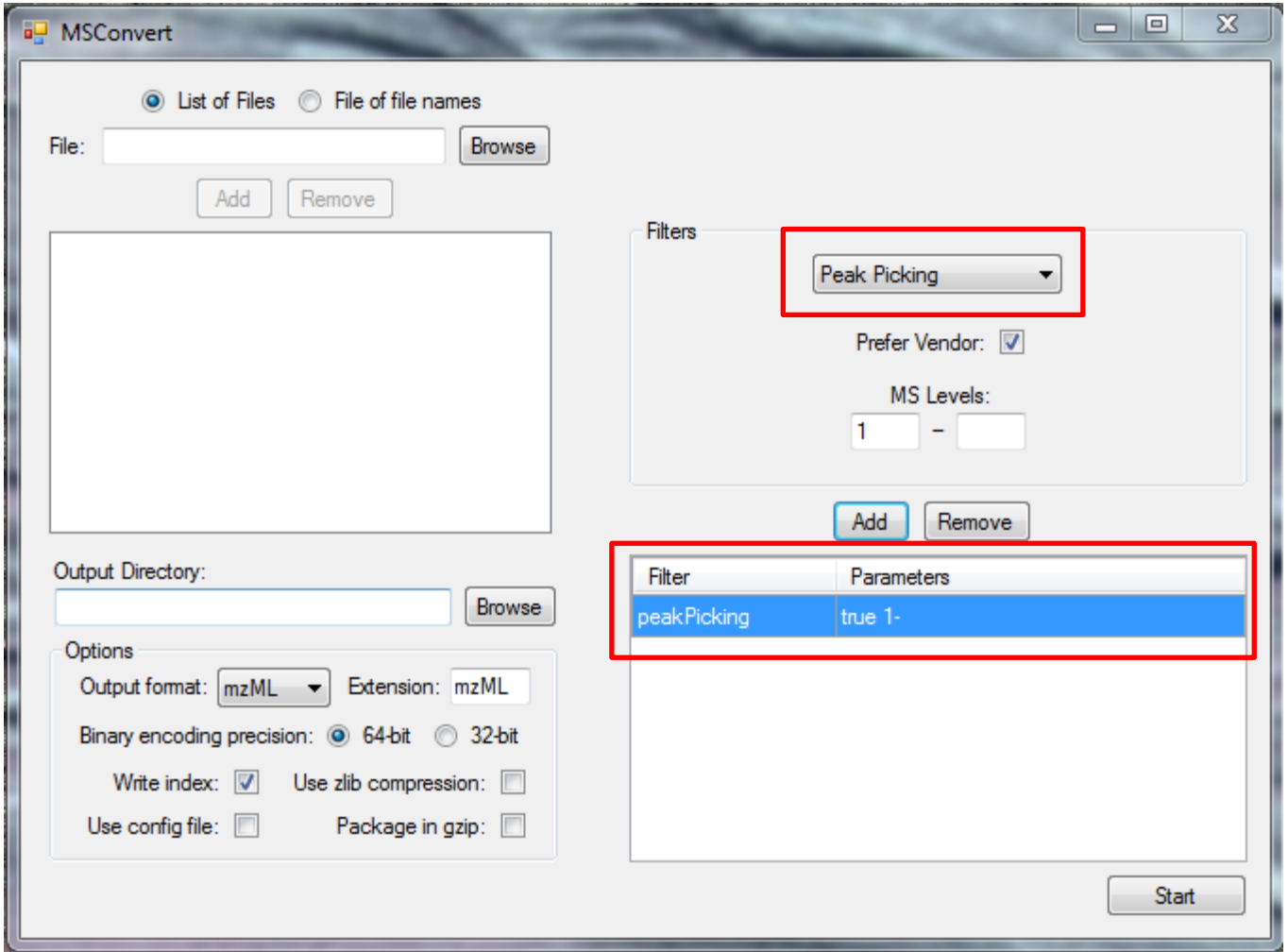
To use profile data (RMassBank deprofiler):



# Before getting started:

Install / Check Proteowizard

To use centroid data (Pwiz deprofiler – example data prepared this way):



# Before getting started

File names…

The files are used to identify a compound in the workflow

- Currently, we require a format like:

  xxxxxxxxxxxx_**1234**_xxx.**mzML**
- _1234_ is the compound ID which is also in the compound list
- mzML is the format currently read by RMassBank

- Advanced / alternative formats possible
  - Provided in the GUI; more details from Steffen

# Before getting started:

The compound list

Go back to the R / RStudio window

- Choose a working directory (change this!!! and create a new one if you like)
- Copy the compound list into this directory:

```
39 # Save a copy of the compound list
40 # First, set a working directory (Imake sure this exists)
41 setwd("C:/DATA/RMassBank/RMassBank_Demo")
42 # then copy the file there
43 file.copy(system.file("list/NarcoticsDataset.csv",
44                       package="RMassBankData"), "./Compoundlist.csv")
45
```

```
> setwd("C:/DATA/RMassBank/RMassBank_Demo")
> # then copy the file there
> file.copy(system.file("list/NarcoticsDataset.csv",
+                       package="RMassBankData"), "./Compoundlist.csv")
[1] TRUE
>
```

# Before getting started:

The compound list

Open up the compound list (Excel, OpenOffice):

- Ignore any error messages Excel gives you (it is CSV, not SYLK!!!)
- Red fields are required: ID for file name, SMILES,
- Optional: Name (SMILES back-up) and RT for MS and MS/MS retrieval window

| | A | B | C | D | E | F | G | H |
|---|---|---|---|---|---|---|---|---|
| 1 | ID | Name_de | Name_En | Name | SMILES | RT | CAS | |
| 2 | 2817 | Kokain | Cocaine | Cocaine | [C@H]1([C@@H]2N([C@@H](C[C@@H]1OC( | 5.15 | 50-36-2 | |
| 3 | 2818 | 1-(3-Chlor | 1-(3-Chlor | 1-(3-Chlorophenyl)pipera | c1c(Cl)cccc1N1CCNCC1 | 5.25 | 6640-24-0 | |
| 4 | 2819 | 1-(3-Triflu | 1-(3-Triflu | 1-(3-Trifluoromethylphen | c1c(C(F)(F)F)cccc1N1CCNCC1 | 6.05 | 15532-75-9 | |
| 5 | 2820 | 1-Benzylp | 1-Benzylp | 1-Benzylpiperazine | C1CN(CCN1)Cc1ccccc1 | 1.64 | 2759-28-6 | |
| 6 | 2821 | Amitriptyl | Amitriptyl | Amitriptyline | C1(\c2c(CCc3c1cccc3)cccc2)=C\CCN(C)C | 8.65 | 50-48-6 | |
| 7 | 2822 | Amphetar | Amphetar | Amphetamine | c1(ccccc1)CC(N)C | 3.7 | 300-62-9 | |
| 8 | 2823 | Benzoylec | Benzoylec | Benzoylecgonine | O(C(=O)c1ccccc1)[C@@H]1[C@@H]([C@@H] | 4.7 | 519-09-5 | |
| 9 | 2824 | Dextrome | Dextrome | Dextromethorphan | c12[C@]34[C@@H]([C@@H]([N@@](C)CC3) | 6.7 | 125-71-3 | |
| 10 | 2825 | EDDP (2-E | EDDP (2-E | EDDP (2-Ethylidene-1,5-di | C1[C@@H](C)N(C)C(\C1(c1ccccc1)c1ccccc1)=( | 6.65 | 30223-73-5 | |
| 11 | 2826 | Ketamin | Ketamine | Ketamine | c1([C@@]2(C(CCCC2)=O)NC)c(cccc1)Cl | 4.7 | 6740-88-1 | |
| 12 | 2827 | Mephedrc | Mephedrc | Mephedrone (4-Methylm | Cc1ccc(cc1)C(=O)C(C)NC | 4.45 | 1189805-46-6 | |
| 13 | 2828 | Methador | Methador | Methadone | CCC(=O)C(CC(C)N(C)C)(c1ccccc1)c2ccccc2 | 8.35 | 76-99-3 | |
| 14 | 2829 | Methamp | Methamp | Methamphetamine | c1cccc(c1)C[C@H](C)NC | 3.85 | 537-46-2 | |
| 15 | 2830 | Naltrexon | Naltrexon | Naltrexone | c12[C@]34[C@@]5([C@H]([N@](CC6CC6)CC4 | 3.25 | 16590-41-3 | |
| 16 | 2758 | Ephedrin | Ephedrine | Ephedrine | c1([C@H]([C@@H](NC)C)O)ccccc1 | 3.03 | 299-42-3 | |
| 17 | | | | | | | | |

# Before getting started:

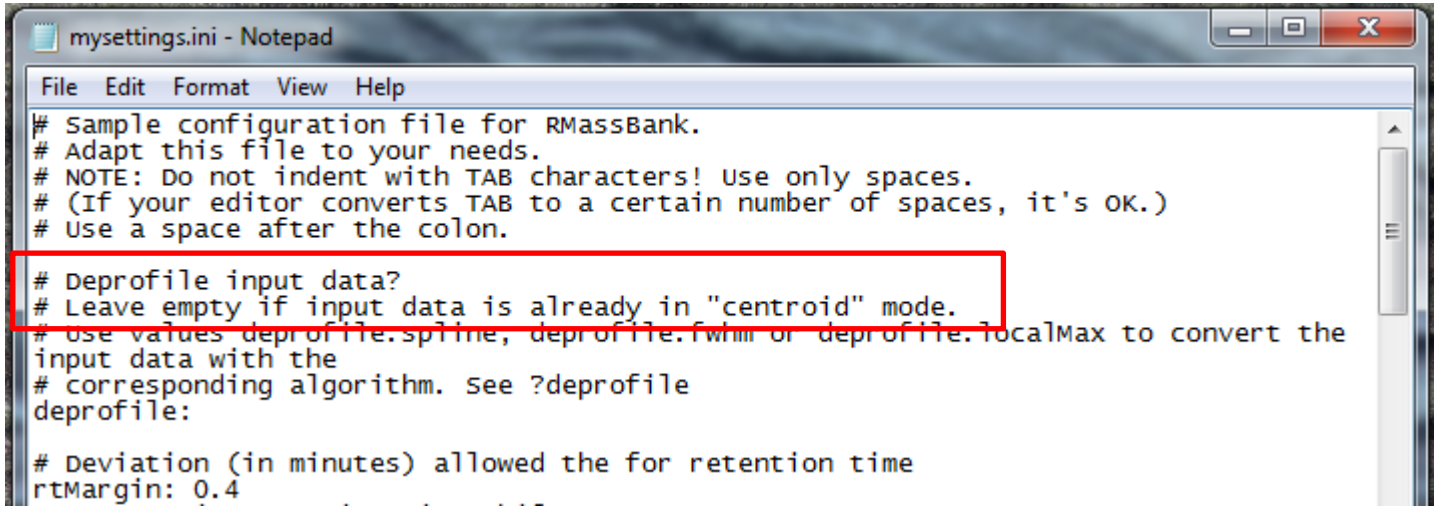
The settings file (lucky last bit!!)

A number of settings are necessary for successful generation:

- Data processing
- Correct annotation of spectra

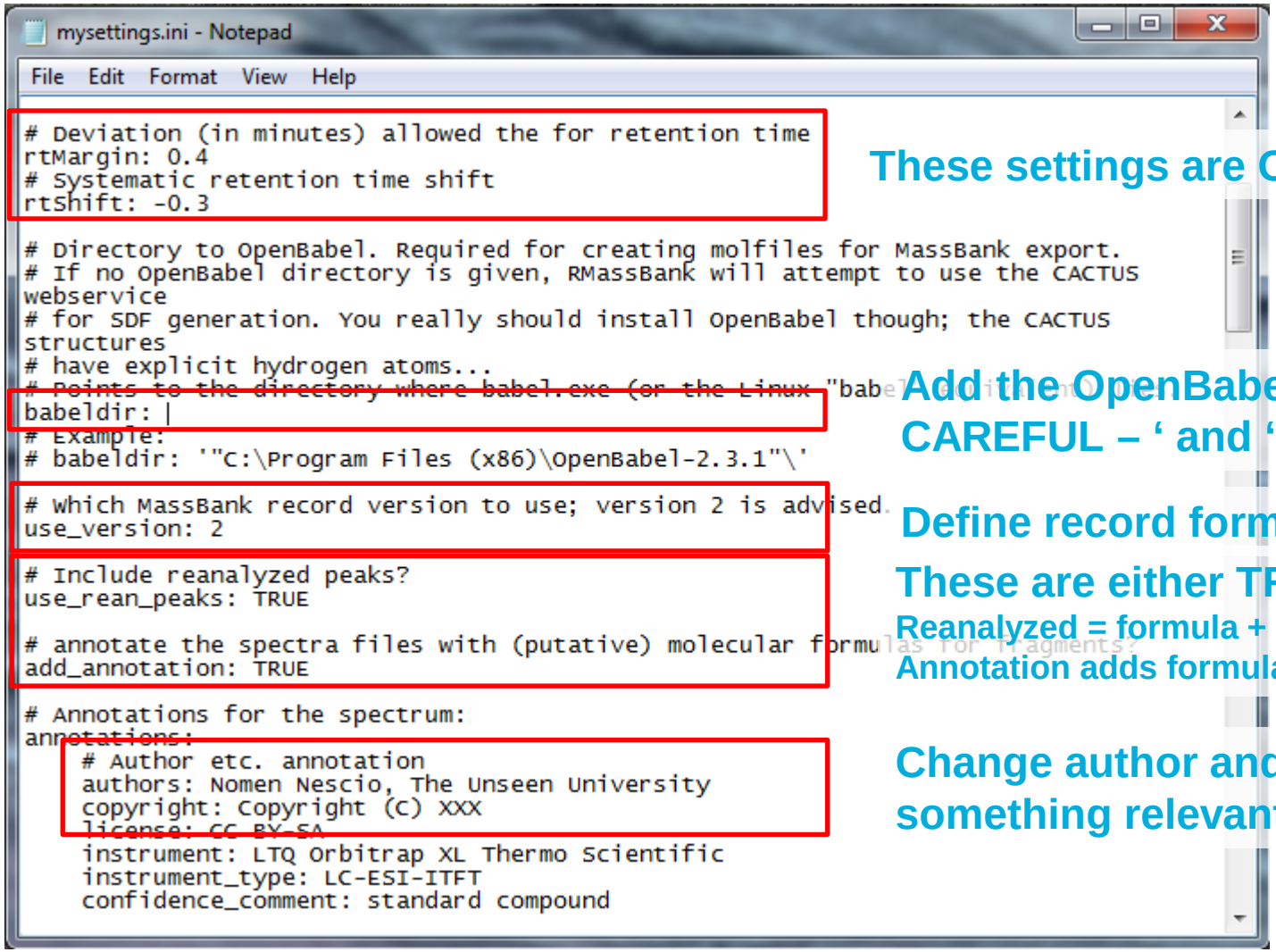
```
# Now, generate a settings file:
RmbSettingsTemplate("mysettings.ini")
```

- Open up in any text editor and edit…



# Before getting started:

The settings file (lucky last bit!!)




**These settings are OK for test data**

**Add the OpenBabel location here**
**CAREFUL – ‘ and “ as in example**

**Define record format (keep 2)**

**These are either TRUE or FALSE**
**Reanalyzed = formula + $N_2O$**
**Annotation adds formula to records**

**Change author and copyright to something relevant for you…**

# Before getting started:

The settings file (lucky last bit!!)

mysettings.ini - Notepad

File Edit Format View Help

```
annotations:
    # Author etc. annotation
    authors: Nomen Nescio, The Unseen University
    copyright: Copyright (c) XXX
    license: CC BY-SA
    instrument: LTQ Orbitrap XL Thermo Scientific
    instrument_type: LC-ESI-ITFT
    confidence_comment: standard compound
    compound_class: N/A; Environmental Standard
    internal_id_fieldname: INTERNAL_ID
    #
    # HPLC annotations:
    #
    # example: lc_gradient: 90/10 at 0 min, 50/50 at 4 min,
25 min, 90/10 at 25.1 min, 90/10 at 30 min
    lc_gradient:
    # example: lc_flow: 200 uL/min
    lc_flow:
    # example: lc_solvent_a: water with 0.1% formic acid
    lc_solvent_a:
    lc_solvent_b:
    # example: lc_column: XBridge C18 3.5um, 2.1x50mm, Wate
    lc_column:
    # Prefix for MassBank accession IDs
    entry_prefix: XX
    ms_type: MS2
    ionization: ESI
    ms_dataprocessing:
        RECALIBRATE: loess on assigned fragments and MS1

# List of data-dependent scans in their order (relative to the parent scan), for
annotation of the MassBank records
# For every data-dependent scan event, specify an element with:
# mode: fragmentation mode, e.g. CID
# ces: "short" format collision energy (for record title)
```

**License: change for own records?**

**These settings are OK for test data**

**These are currently empty**

**You can use the examples for the example data. Solvent B was methanol!**

**These values will need adjusting for own chromatographic program**

**Choose your record letters!**

**E.g. Eawag = EA**

**These are OK for test data**

# Before getting started:

The settings file (lucky last bit!!)

mysettings.ini - Notepad

File Edit Format View Help

```
# List of data-dependent scans in their order (relative to the parent scan), for
annotation of the MassBank records
# For every data-dependent scan event, specify an element with:
# mode: fragmentation mode, e.g. CID
# ces: "short" format collision energy (for record title)
# ce: "long" format collision energy (for annotation field)
# res: FT resolution
spectraList:
 # First scan: CID 35% NCE, resolution 7500
- mode: CID
  ces: 35%
  ce: 35 % (nominal)
  res: 7500
 # Second scan: HCD 15% NCE, resolution 7500
- mode: HCD
  ces: 15%
  ce: 15 % (nominal)
  res: 7500
 # Third scan, etc.
- mode: HCD
  ces: 30%
  ce: 30 % (nominal)
  res: 7500
- mode: HCD
  ces: 45%
  ce: 45 % (nominal)
  res: 7500
- mode: HCD
  ces: 60%
  ce: 60 % (nominal)
  res: 7500
- mode: HCD
  ces: 75%
  ce: 75 % (nominal)
```

**These settings are OK for test data**

**However:**
**This will need adjustment for own experiments!**

# Before getting started:

The settings file (lucky last bit!!)

mysettings.ini - Notepad

File Edit Format View Help

```
# Shifts of the starting points for RMassBank accession numbers.
# Change these if you measure different adducts
accessionNumberShifts:
    pH: 0 # [M+H]+: Accession numbers 1-14
    pM: 16 # [M]+: 17-30
    pNa: 32 # [M+Na]+: 33-46
    mH: 50 # [M-H]-: 51-64
    mFA: 66 # [M+FA]-: 67-80

# A list of known electronic noise peaks
electronicNoise:
- 189.825
- 201.725
- 196.875
# Exclusion width of electronic noise peaks (from unmatched peaks, or during
# reanalysis)
electronicNoiseWidth: 0.3

# recalibration settings:
# recalibrate by: dppm or dmz
recalibrateBy: dppm

# recalibrate MS1:
# separately (separate)
# with common curve (common)
# do not recalibrate (none)
recalibrateMS1: common

# Custom recalibration function: You can overwrite the recal function by
# making any function which takes rcdata$recalfield ~ rcdata$mzFound.
# The settings define which recal function is used.
# Note: if recalibrateMS1 is "common", the setting "recalibrator: MS1: ..." is
meaningless
# because the MS1 points will be recalibrated together with the MS2 points with
# the MS2 recalibration function.
recalibrator:
    MS1: recalibrate.loess
    MS2: recalibrate.loess
```

**Fixed accession number: XXYYYYZZ;**
**XX = letters, YYYY = compound ID**
**ZZ = "shifts" – for us, these adducts**

**Remove consistent electronic noise**
**Add the middle mass of the noise (often broad peaks, or slightly varying)**
**Then add exclusion width –not too generous!**

**Recalibrate settings:**
**Mass accuracy unit**
**MS & MS/MS**

**These are OK for test data**
**Changes necessary if recalibration needs adjusting**

# Before getting started:

The settings file (lucky last bit!!)

**Finally: Save this before you go on!!!**

mysettings_ES.ini - Notepad

File Edit Format View Help

```
# Define the multiplicity filtering level
# Default is 2 (peak occurs at least twice)
# Set this to 1 if you want to turn this option off.
# Set this to anything > 2 if you want harder filtering
multiplicityFilter: 2
```

**Multiplicity Filtering settings e.g. require Peak to occur twice or more.**

```
# Define the title format.
# You can use all entries from MassBank records as tokens
# plus the additional token RECORD_TITLE_CE, which is a shortened
# version of the collision energy specifically for use in the title.
# Every line is one entry and must have one token in curly brackets
# e.g. {CH$NAME} or {AC$MASS_SPECTROMETRY: MS_TYPE} plus optionally
# additional text in front or behind e.g.
# R={AC$MASS_SPECTROMETRY: RESOLUTION}
# If this is not specified, it defaults to a title of the format
# "Dinotefuran; LC-ESI-QFT; MS2; CE: 35%; R=35000; [M+H]+"
# Note how everything must be in "" here because otherwise the : are ge
titleFormat:
- "{CH$NAME}"
- "{AC$INSTRUMENT_TYPE}"
- "{AC$MASS_SPECTROMETRY: MS_TYPE}"
- "CE: {RECORD_TITLE_CE}"
- "R={AC$MASS_SPECTROMETRY: RESOLUTION}"
- "{MS$FOCUSED_ION: PRECURSOR_TYPE}"
```

**Change the title format**

**If you wish to display different Information in the title**

```
# Define filter settings.
# For Orbitrap, settings of 15 ppm in low mass range, 10 ppm in high
# mass range, m/z = 120 as mass range division and 5 ppm for recalibrated
# data overall are recommended.
filterSettings:
    ppmHighMass: 10
    ppmLowMass: 15
    massRangeDivision: 120
    ppmFine: 5
    prelimCut: 1e4
    prelimCutRatio: 0
    fineCut: 0
    fineCutRatio: 0
    specOkLimit: 1e4
    dbeMinLimit: -0.5
```

**These are all advanced settings and do not need to be changed**

# Before getting started:

The settings file (lucky last bit!!)

**Finally: Save this before you go on!!!**

mysettings_ES.ini - Notepad

File Edit Format View Help

```
# Note how everything must be in "" here because otherwise the : are getting mangled!
titleFormat:
- "{CH$NAME}"
- "{AC$INSTRUMENT_TYPE}"
- "{AC$MASS_SPECTROMETRY: MS_TYPE}"
- "CE: {RECORD_TITLE_CE}"
- "R={AC$MASS_SPECTROMETRY: RESOLUTION}"
- "{MS$FOCUSED_ION: PRECURSOR_TYPE}"

# Define filter settings.
# For Orbitrap, settings of 15 ppm in low mass range, 10 ppm in high
# mass range, m/z = 120 as mass range division and 5 ppm for recalibrated
# data overall are recommended.
filterSettings:
    ppmHighMass: 10
    ppmLowMass: 15
    massRangeDivision: 120
    ppmFine: 5
    prelimCut: 1e4
    prelimCutRatio: 0
    fineCut: 0
    fineCutRatio: 0
    specOkLimit: 1e4
    dbeMinLimit: -0.5
    satelliteMzLimit: 0.5
    satelliteIntLimit: 0.05

 # Define raw MS retrieval settings.
findMsMsRawSettings:
    ppmFine: 10
    mzCoarse: 0.5
    # fillPrecursorScan is FALSE for "good" mzML files which have all the info needed.
    # However, for example AB Sciex files will have missing precursor scan information,
    # in which case fillPrecursorScan = TRUE is needed. Try it out.
    fillPrecursorScan: FALSE
```

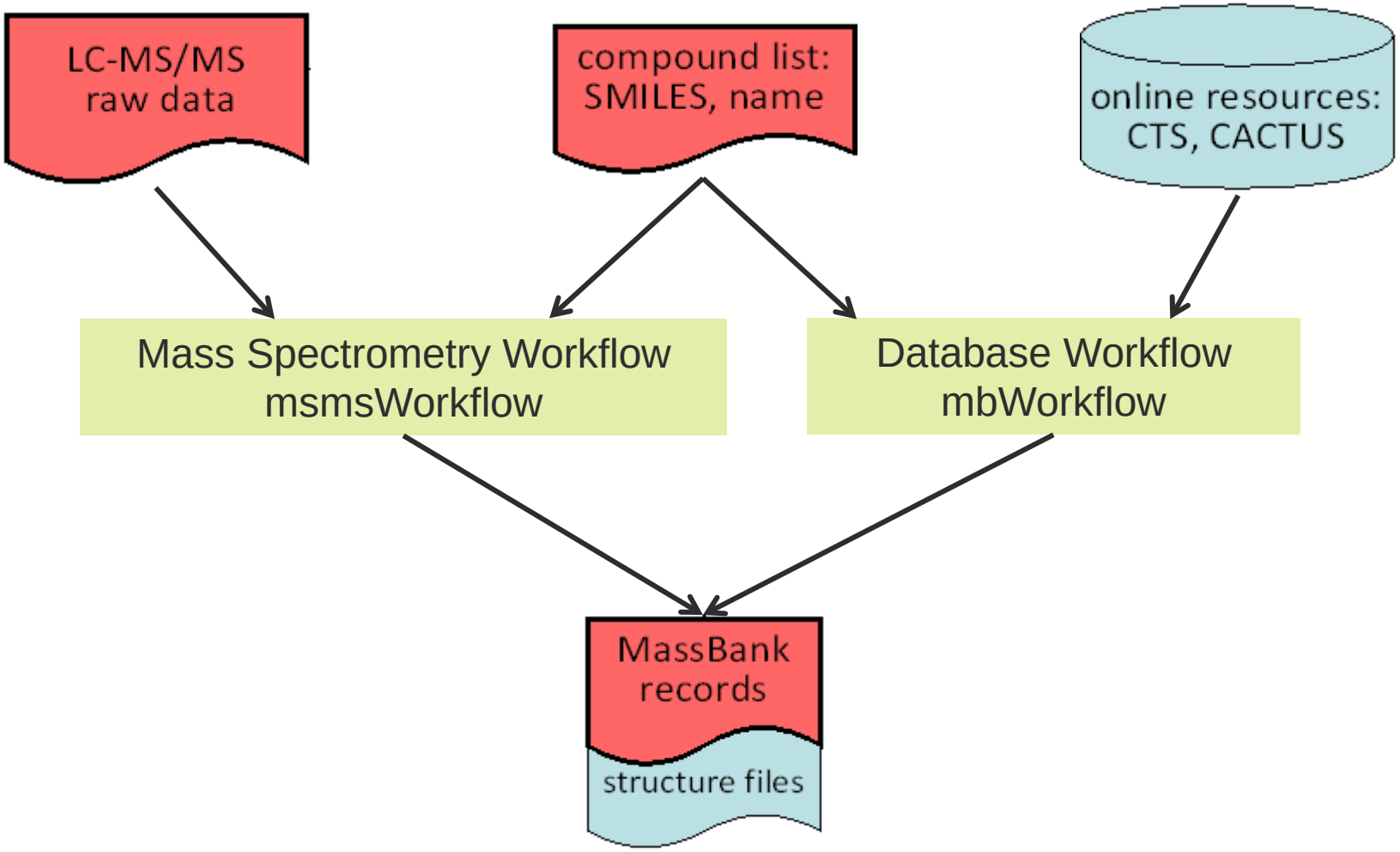
**These are all advanced settings and do not need to be changed**

**Adjust the boundaries for recalibration**

**Change the MS retrieval**

**May be needed for different Instruments than the default.**

# RMassBank Workflow – Simple Form

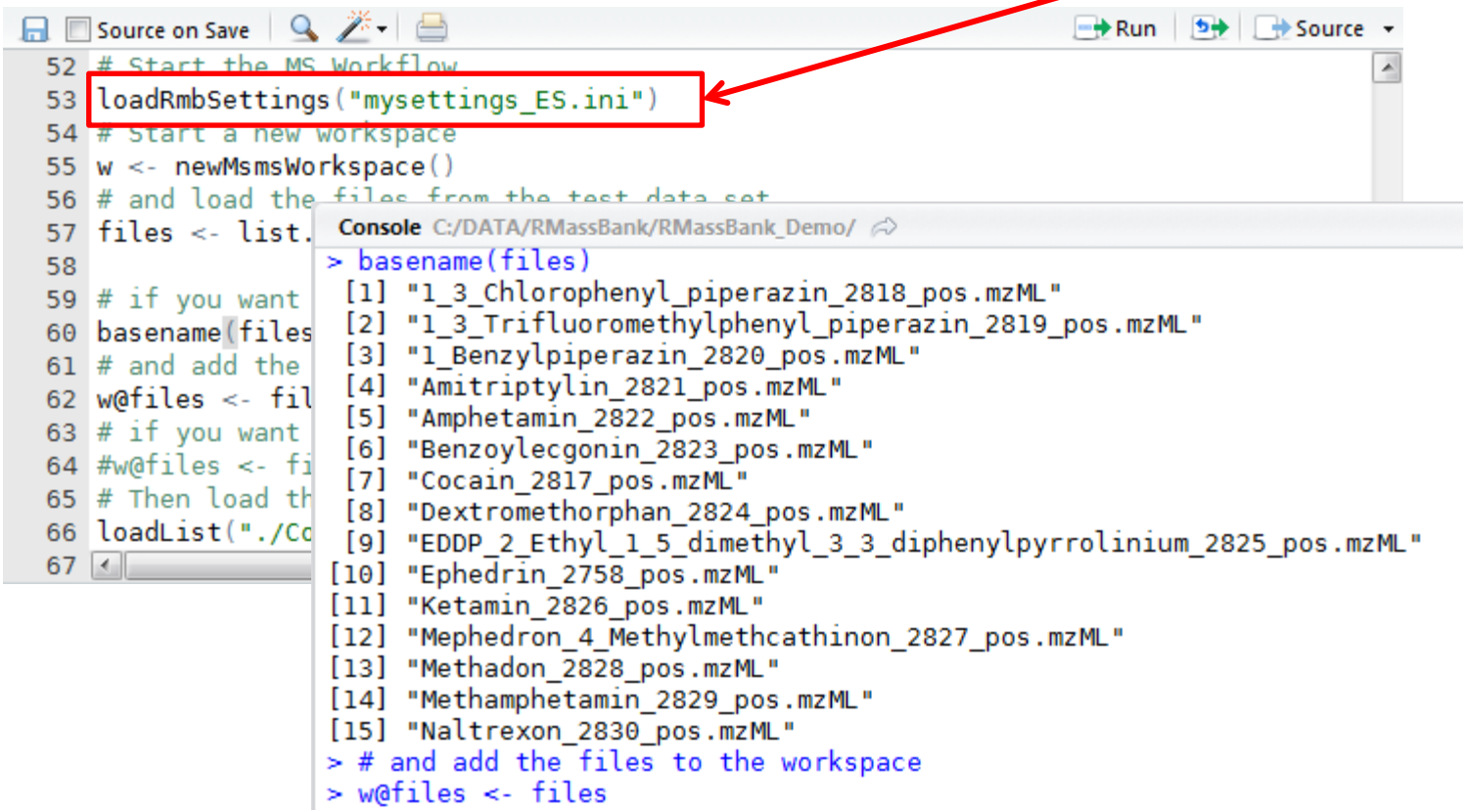


# Mass Spectrometry Workflow

Now we can get started!

Go back to your R session and “RMassBank_Demo.R”

- Select and run this whole snippet – but change the file name first!



# Mass Spectrometry Workflow

msmsWorkflow: 8 Steps in summary

- Step 1: Search all `files` for MS/MS spectra of compound
- Step 2: Molecular formula fitting for all peaks (large tolerance)
- Step 3: Analyzed spectra aggregated into a list
- Step 4: Recalibration curve is calculated; all spectra are recalibrated
- Step 5: Recalibrated spectra are reanalyzed (new formula fit post calibration)
- Step 6: Reanalyzed spectra aggregated. Electronic noise removed.
  - Back-up saved as `archiveName.RData`
- Step 7: Fail peaks are reanalyzed (formula + 2N, O)
  - Back-up saved as `archiveName_RA.RData`
- Step 8: Multiplicity Filtering: peaks occurring once only are removed.
  - Back-up saved as `archiveName_RF.RData`
  - High Intensity Fail Peaks: `archiveName_Failpeaks.csv`

# Mass Spectrometry Workflow

8 steps in RMassBank code…

- Steps 1:4 – up to recalibration

```
# Start the workflow with steps 1 to 4:
w <- msmsWorkflow(w, mode="pH", steps=c(1:4), archivename = "pH_narcotics")
```

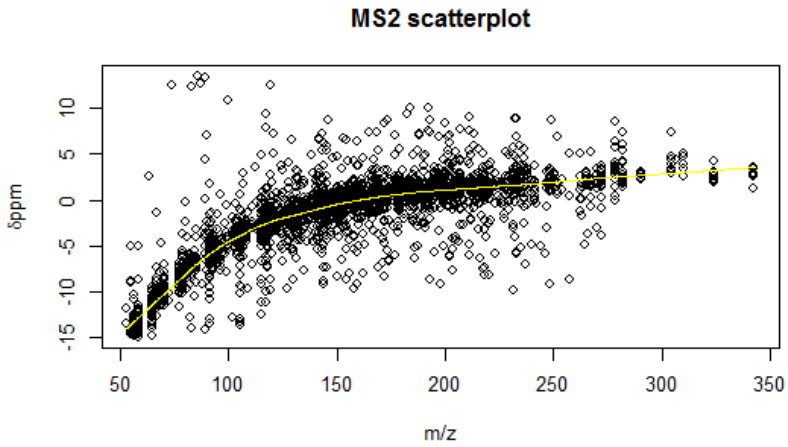
  - Warning: this can take a while….
  - Progress:

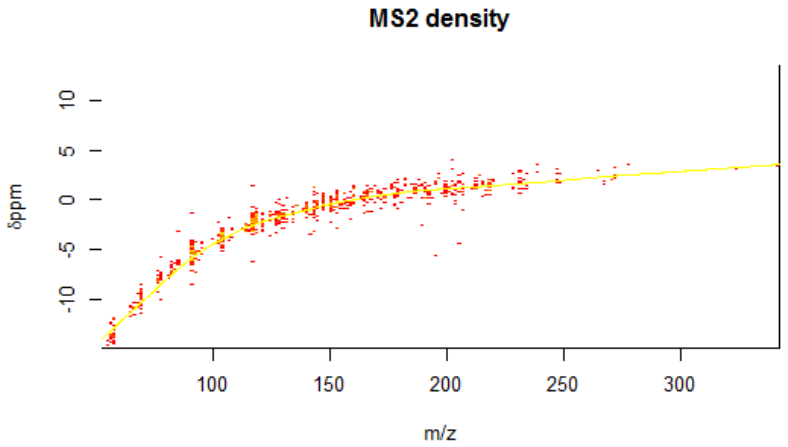
```
> # Start the workflow with steps 1 to 4:
> w <- msmsWorkflow(w, mode="pH", steps=c(1:4), archivename = "pH_narcotics")
msmsWorkflow: Step 1
  |
|   0%3: >C:/Users/schymaem/Documents/R/win-
library/2.14/RMassBankData/spectra/1_3_Chlorophenyl_piperazin_2818_pos.mzML
  |
|   7%3: >C:/Users/schymaem/Documents/R/win-
library/2.14/RMassBankData/spectra/1_3_Trifluoromethylphenyl_piperazin_2819_pos.mzML
  |
|  13%3: >C:/Users/schymaem/Documents/R/win-
library/2.14/RMassBankData/spectra/1_Benzylpiperazin_2820_pos.mzML
  |
|===============                                                  |  20%3:
>C:/Users/schymaem/Documents/R/win-library/2.14/RMassBankData/spectra/Amitriptylin_2821
  |
```

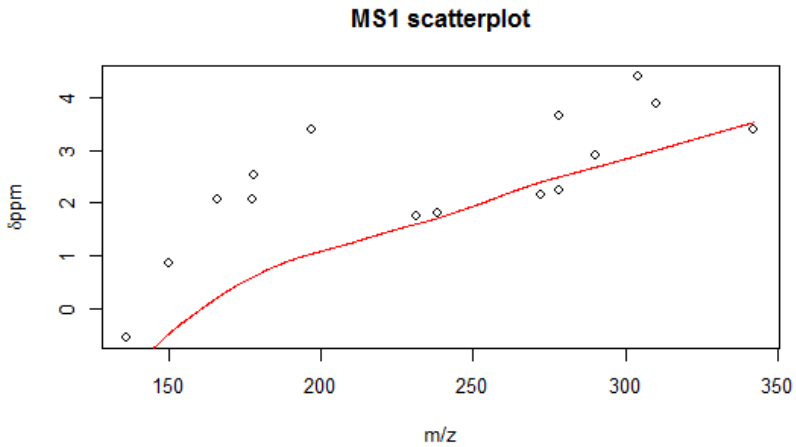
# Mass Spectrometry Workflow

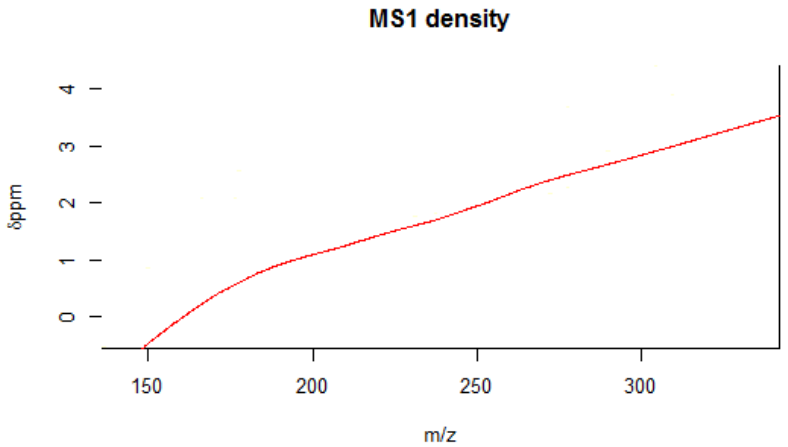
Steps 1:4 of RMassBank code…

- Recalibration plot at the end



# Mass Spectrometry Workflow

Steps 5:8 in RMassBank code…

- Steps 5:8 – complete the MS workflow

```
# Keep going with steps 5-8:
w <- msmsWorkflow(w, mode="pH", steps=c(5:8), archivename = "pH_narcotics")
```

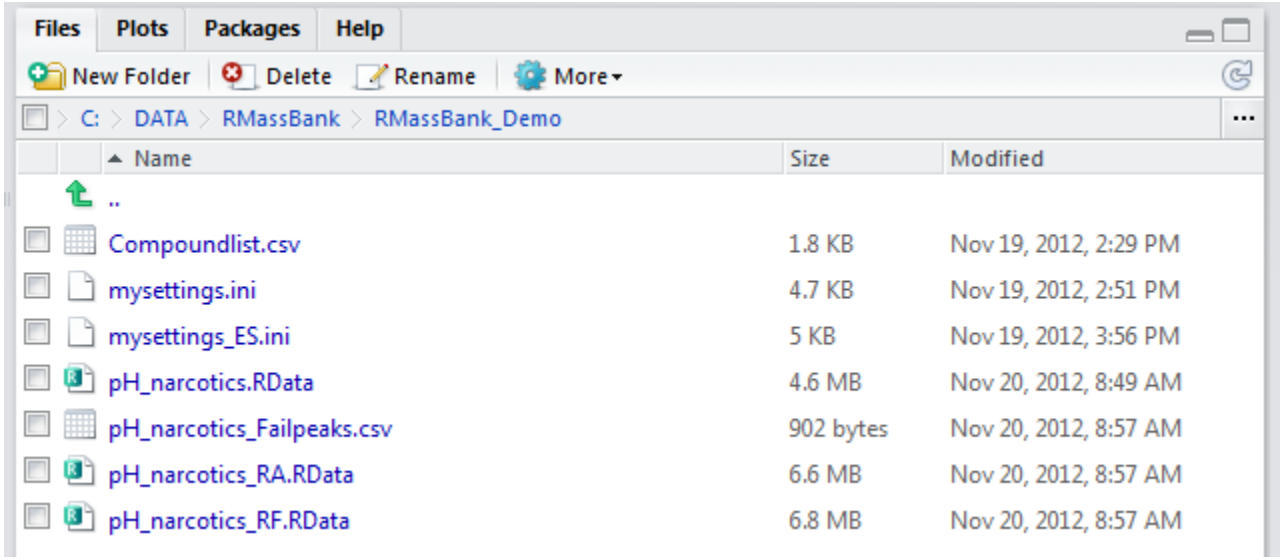
  - Warning: this can also take a while….
  - Progress:

```
> # Keep going with steps 5-8:
> w <- msmsWorkflow(w, mode="pH", steps=c(5:8), archivename = "pH_narcotics")
msmsWorkflow: Step 5
  |
|   0%  |
|   7%  |
|==========                                                       |  13%  |
|===============                                                  |  20%  |
|====================                                             |  27%  |
|=========================                                        |  33%  |
|==============================                                   |  40%  |
|===================================                              |  47%  |
|========================================                         |  53%  |
|=============================================                    |  60%
```

  - Now we need to look at the “Fail Peaks”

# Mass Spectrometry Workflow

File listing should now look something like:




- Open up the fail peaks (outside R) and have a look

| | A | B | C | D | E | F | G | H | I | J | K | L | M | N | O |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | OK | name | cpdID | scan | mzFound | formula | mzCalc | dppm | dbe | mz | int | formulaCc | parentSca | aMax | mzCenter |
| 2 | | 407 | 2817 | 558 | 182.1191 | NA | NA | NA | NA | 182.1193 | 117786.4 | 0 | 556 | 9902212 | 304.1543 |
| 3 | | 2231 | 2821 | 963 | 233.1351 | NA | NA | NA | NA | 233.1355 | 255822.6 | 0 | 961 | 16383558 | 278.1903 |
| 4 | | 3491 | 2824 | 723 | 215.1468 | NA | NA | NA | NA | 215.1471 | 242559.4 | 0 | 721 | 15875343 | 272.2009 |
| 5 | | 3988 | 2825 | 725 | 249.1548 | NA | NA | NA | NA | 249.1553 | 516037.1 | 0 | 723 | 23457888 | 278.1903 |
| 6 | | 4110 | 2825 | 730 | 193.0993 | NA | NA | NA | NA | 193.0995 | 138038.3 | 0 | 723 | 5673618 | 278.1903 |
| 7 | | 5543 | 2828 | 938 | 265.1612 | NA | NA | NA | NA | 265.1618 | 146493.5 | 0 | 933 | 4158767 | 310.2165 |
| 8 | | 6503 | 2830 | 352 | 211.0738 | C12H9N3O | 211.074 | -0.9633 | 13 | 211.0741 | 24055.95 | 1 | 346 | 195149.4 | 342.17 |

# Mass Spectrometry Workflow

## Fail Peaks

Compound ID

MS/MS Scan

m/z & I of MS/MS peak

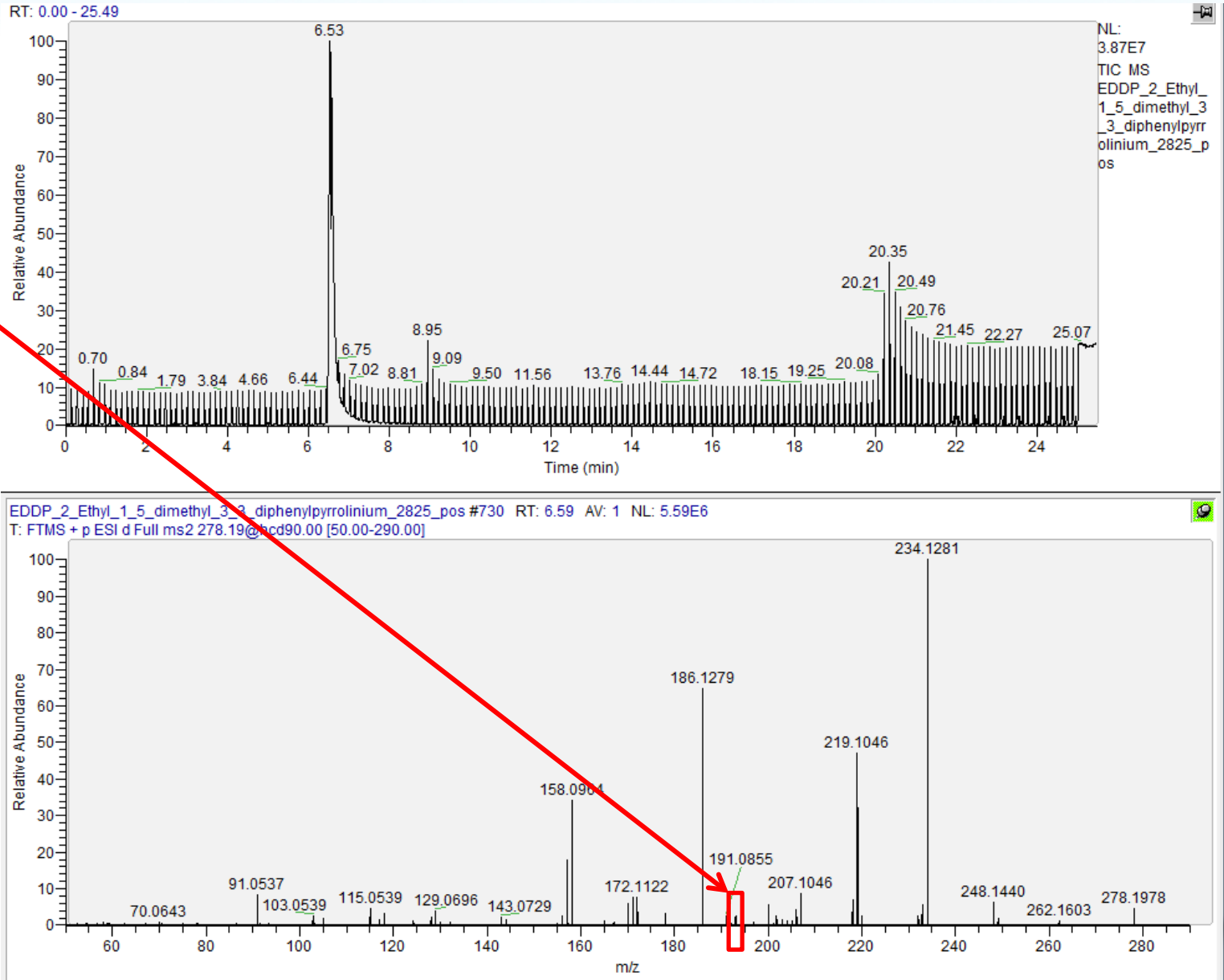
I, m/z of precursor

| | A | B | C | D | E | F | G | H | I | J | K | L | M | N | O |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | OK | name | cpdID | scan | mzFound | formula | mzCalc | dppm | dbe | mz | int | formulaCo | parentSca | aMax | mzCenter |
| 2 | | 407 | 2817 | 558 | 182.1191 | NA | NA | NA | NA | 182.1193 | 117786.4 | 0 | 556 | 9902212 | 304.1543 |
| 3 | | 2231 | 2821 | 963 | 233.1351 | NA | NA | NA | NA | 233.1355 | 255822.6 | 0 | 961 | 16383558 | 278.1903 |
| 4 | | 3491 | 2824 | 723 | 215.1468 | NA | NA | NA | NA | 215.1471 | 242559.4 | 0 | 721 | 15875343 | 272.2009 |
| 5 | | 3988 | 2825 | 725 | 249.1548 | NA | NA | NA | NA | 249.1553 | 516037.1 | 0 | 723 | 23457888 | 278.1903 |
| 6 | | 4110 | 2825 | 730 | 193.0993 | NA | NA | NA | NA | 193.0995 | 138038.3 | 0 | 723 | 5673618 | 278.1903 |
| 7 | | 5543 | 2828 | 938 | 265.1612 | NA | NA | NA | NA | 265.1618 | 146493.5 | 0 | 933 | 4158767 | 310.2165 |
| 8 | | 6503 | 2830 | 352 | 211.0738 | C12H9N3O | 211.074 | -0.9633 | 13 | 211.0741 | 24055.95 | 1 | 346 | 195149.4 | 342.17 |

No formula calculated; not replicated (only present once) and low I compared with parent => “true” fail peaks
Look at raw data of “2825”

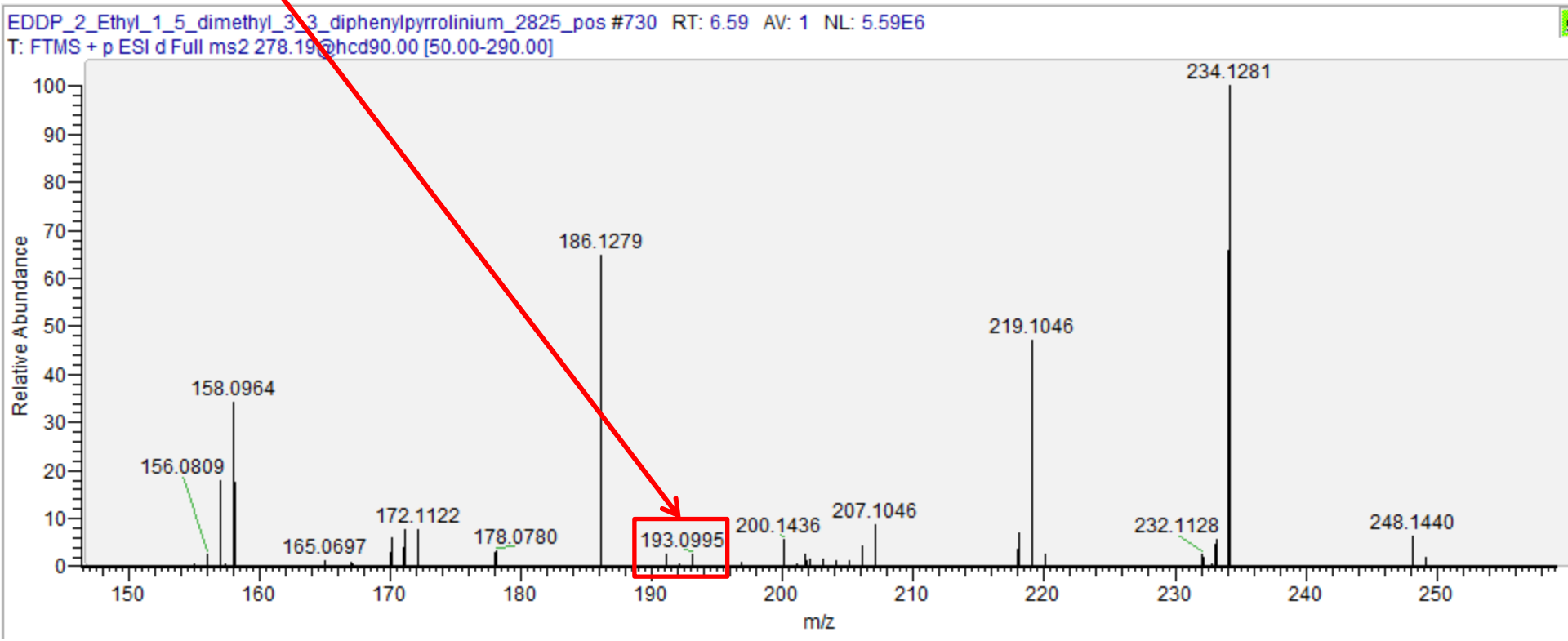
# Mass Spectrometry Workflow

Fail Peak



# Mass Spectrometry Workflow

Fail Peaks




"Fail peak" is very minor part of spectrum, already within noise levels

# Mass Spectrometry Workflow

## Fail Peaks

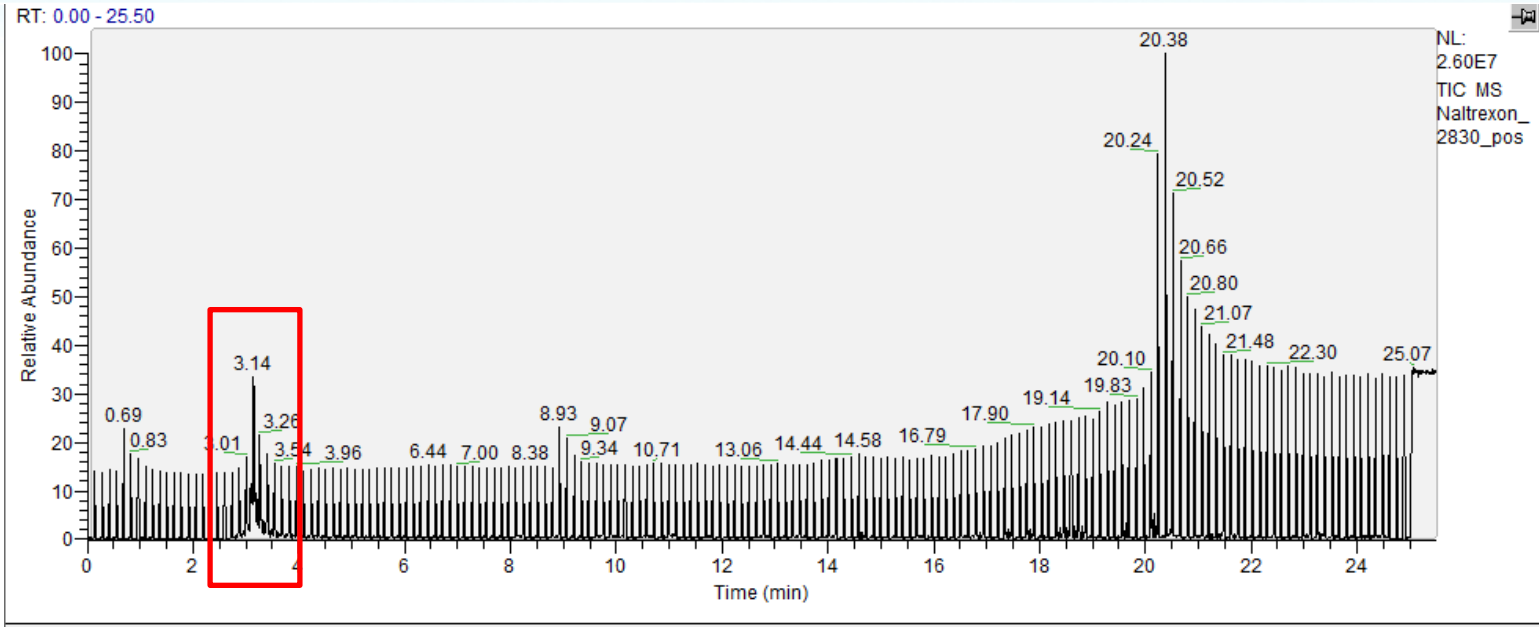
| | A | B | C | D | E | F | G | H | I | J | K | L | M | N | O |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | OK | name | cpdID | scan | mzFound | formula | mzCalc | dppm | dbe | mz | int | formulaC | parentSca | aMax | mzCenter |
| 2 | | 407 | 2817 | 558 | 182.1191 | NA | NA | NA | NA | 182.1193 | 117786.4 | 0 | 556 | 9902212 | 304.1543 |
| 3 | | 2231 | 2821 | 963 | 233.1351 | NA | NA | NA | NA | 233.1355 | 255822.6 | 0 | 961 | 16383558 | 278.1903 |
| 4 | | 3491 | 2824 | 723 | 215.1468 | NA | NA | NA | NA | 215.1471 | 242559.4 | 0 | 721 | 15875343 | 272.2009 |
| 5 | | 3988 | 2825 | 725 | 249.1548 | NA | NA | NA | NA | 249.1553 | 516037.1 | 0 | 723 | 23457888 | 278.1903 |
| 6 | | 4110 | 2825 | 730 | 193.0993 | NA | NA | NA | NA | 193.0995 | 138038.3 | 0 | 723 | 5673618 | 278.1903 |
| 7 | | 5543 | 2828 | 938 | 265.1612 | NA | NA | NA | NA | 265.1618 | 146493.5 | 0 | 933 | 4158767 | 310.2165 |
| 8 | | 6503 | 2830 | 352 | 211.0738 | C12H9N3O | 211.074 | -0.9633 | 13 | 211.0741 | 24055.95 | 1 | 346 | 195149.4 | 342.17 |

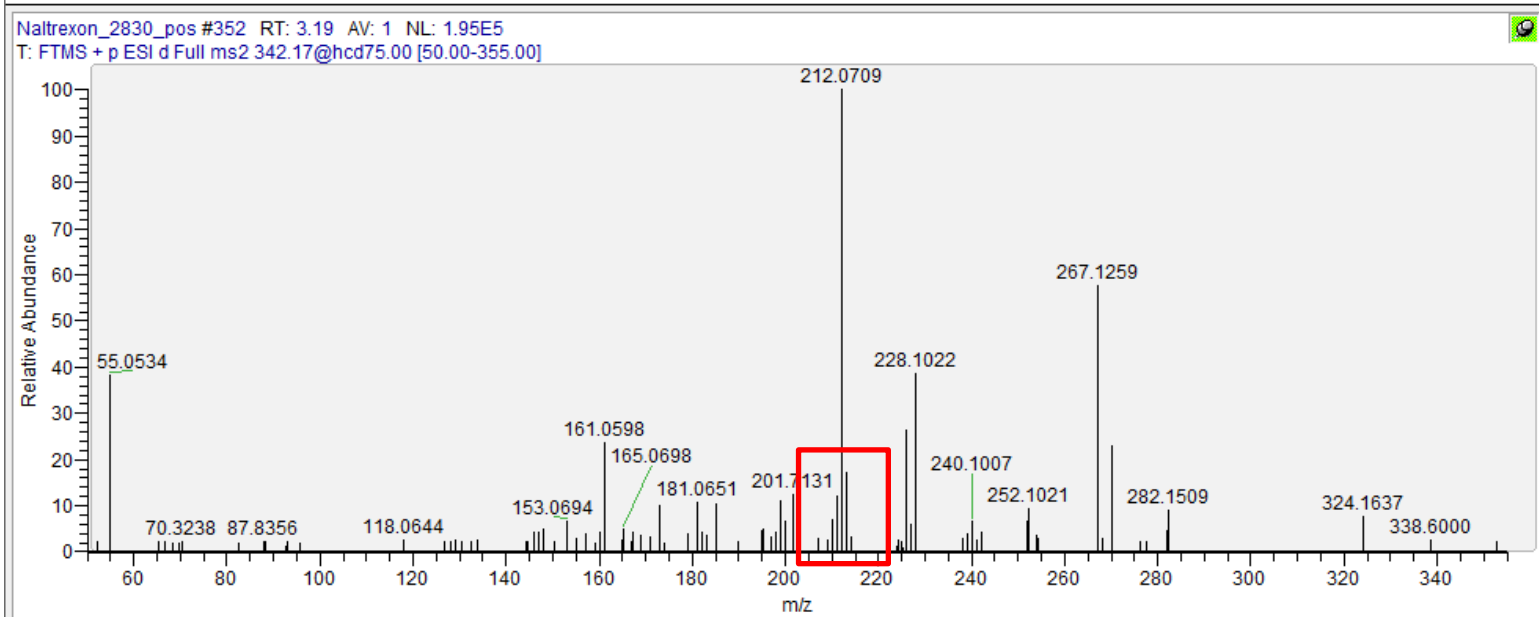
Formula calculated, but also low intensity compared with precursor
Go to raw data…

# Mass Spectrometry Workflow

Fail Peak

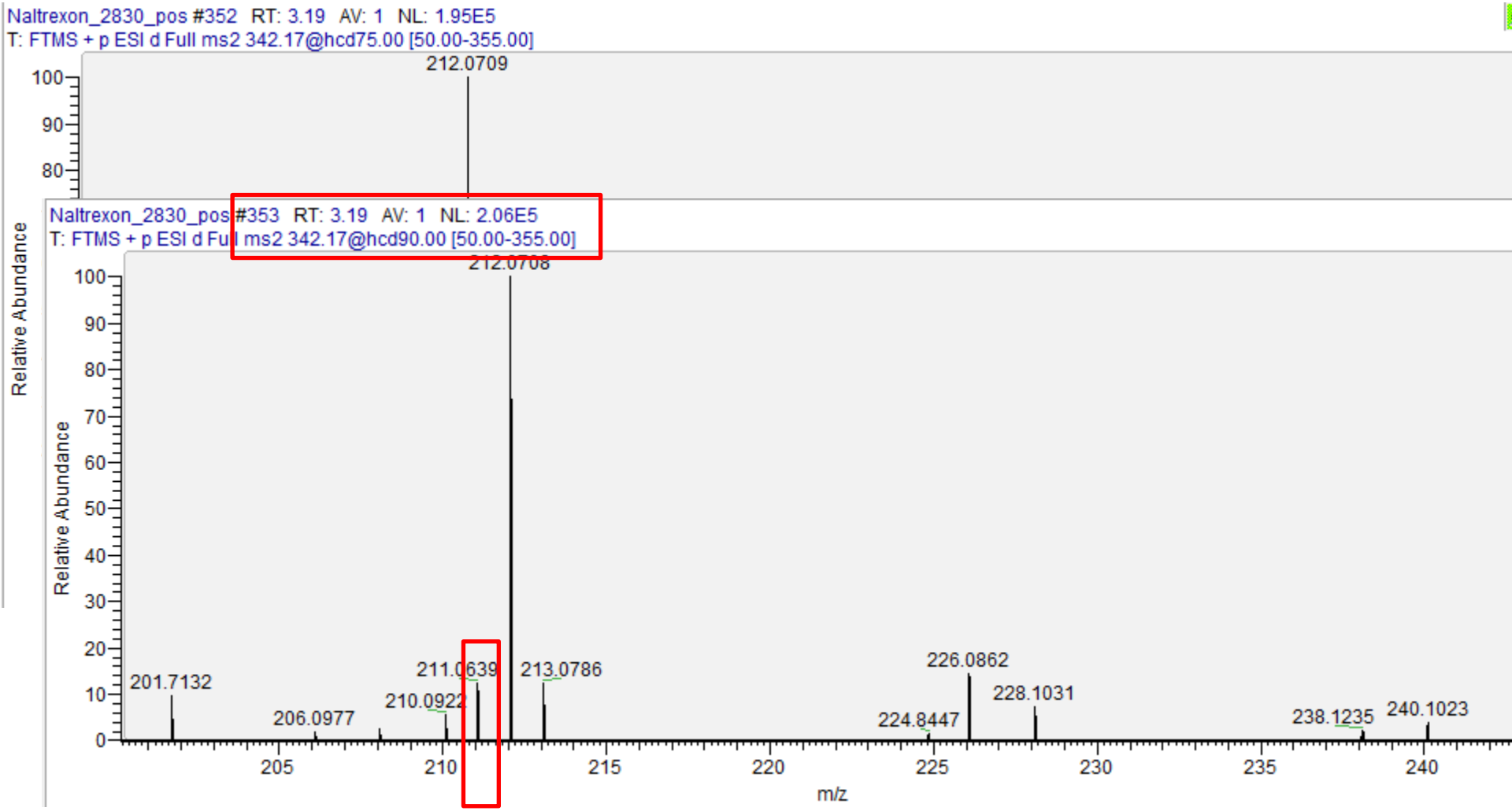
(with formula)



# Mass Spectrometry Workflow

## Fail peak (with formula)



# Mass Spectrometry Workflow

## Fail Peaks

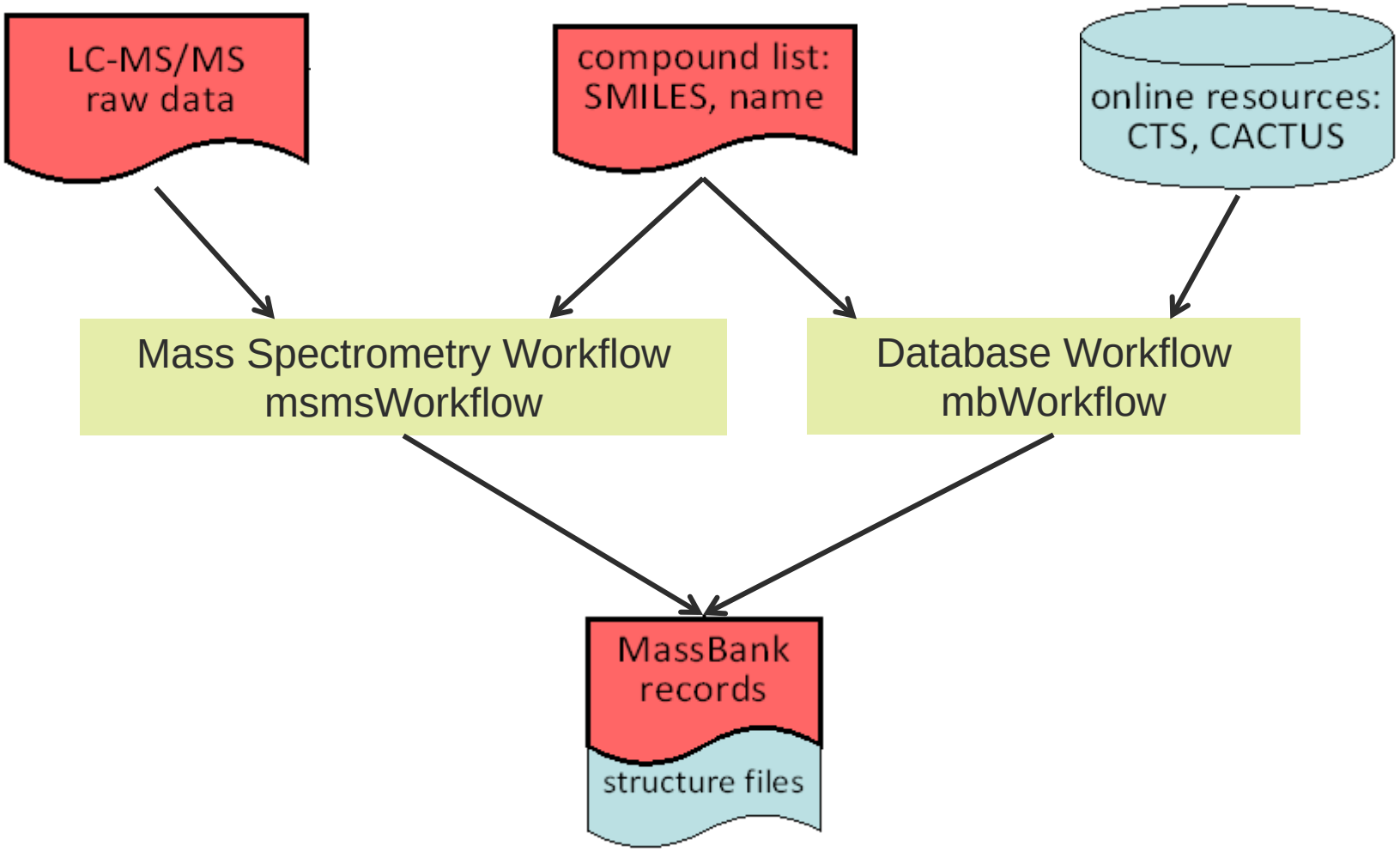
Include this “fail peak” by entering a 1 in the OK column

- Save under a new name…
  - e.g. “pH_narcotics_Failpeaks_wOKs.csv”

| | A | B | C | D | E | F | G | H | I | J | K | L | M | N | O |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | OK | name | cpdID | scan | mzFound | formula | mzCalc | dppm | dbe | mz | int | formulaC | parentSca | aMax | mzCe |
| 2 | 0 | 407 | 2817 | 558 | 182.1191 | NA | NA | NA | NA | 182.1193 | 117786.4 | 0 | 556 | 9902212 | 304. |
| 3 | 0 | 2231 | 2821 | 963 | 233.1351 | NA | NA | NA | NA | 233.1355 | 255822.6 | 0 | 961 | 16383558 | 278. |
| 4 | 0 | 3491 | 2824 | 723 | 215.1468 | NA | NA | NA | NA | 215.1471 | 242559.4 | 0 | 721 | 15875343 | 272. |
| 5 | 0 | 3988 | 2825 | 725 | 249.1548 | NA | NA | NA | NA | 249.1553 | 516037.1 | 0 | 723 | 23457888 | 278. |
| 6 | 0 | 4110 | 2825 | 730 | 193.0993 | NA | NA | NA | NA | 193.0995 | 138038.3 | 0 | 723 | 5673618 | 278. |
| 7 | 0 | 5543 | 2828 | 938 | 265.1612 | NA | NA | NA | NA | 265.1618 | 146493.5 | 0 | 933 | 4158767 | 310. |
| 8 | 1 | 6503 | 2830 | 352 | 211.0738 | C12H9N3C | 211.074 | -0.9633 | 13 | 211.0741 | 24055.95 | 1 | 346 | 195149.4 | 34 |
| 9 | | | | | | | | | | | | | | | |

We are now ready to go onto the Record Creation part of the workflow…

# RMassBank Workflow – Simple Form



# MassBank Record Workflow

Start a new MB workspace and reset the “Infolists” (compound information)

- Reset and load “Infolists” (compound information) – we will use a part-filled one

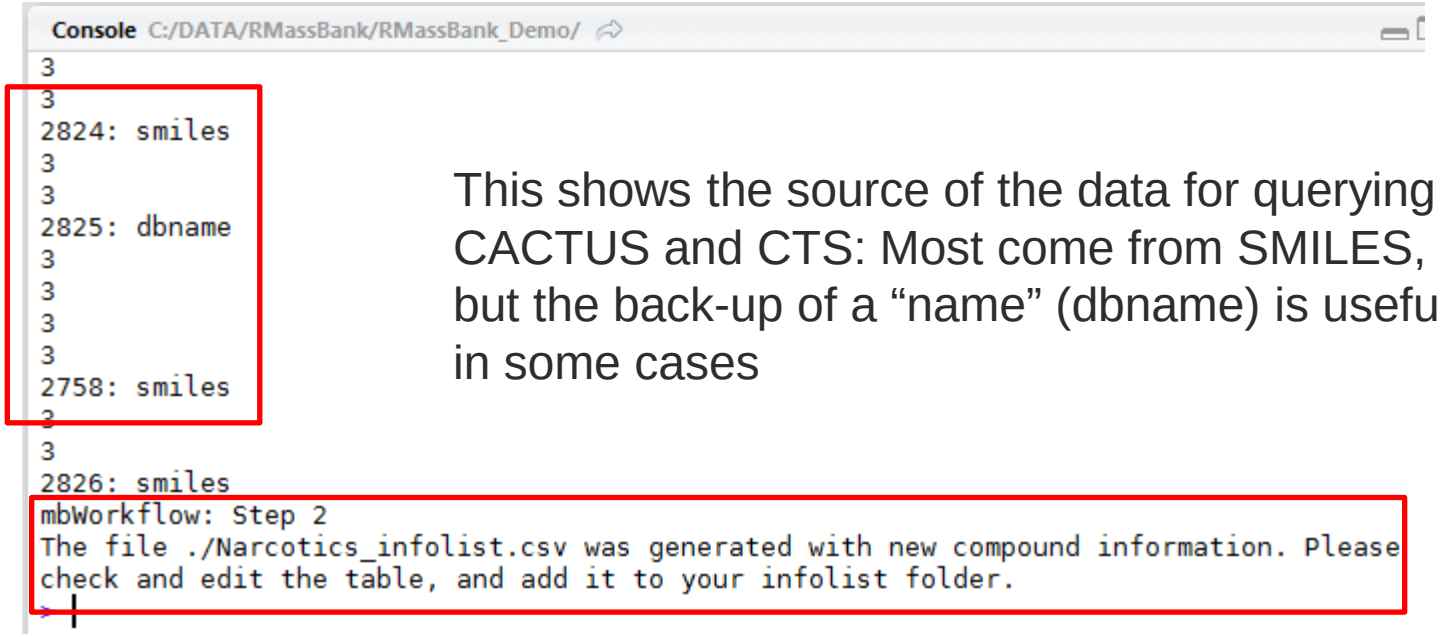
```
79
80 # Next, start the MassBank Record Workflow:
81 mb <- newMbWorkspace(w)
82 mb <- resetInfolists(mb)
83 # To speed up this example, we have provided a partially complete list
84 mb <- loadInfolists(mb, system.file("infolists_incomplete",
85                                     package="RMassBankData"))
```

- Add fail peaks, if applicable, and start the workflow:

```
90 # now add the fail peaks from above, if you have included any
91 mb <- addPeaks(mb, "pH_narcotics_Failpeaks_wOKs.csv")
92
93 # Start the record generation workflow
94 mb <- mbWorkflow(mb, infolist_path="./Narcotics_infolist.csv")
95
```

# MassBank Record Workflow

Progress:




This shows the source of the data for querying CACTUS and CTS: Most come from SMILES, but the back-up of a “name” (dbname) is useful in some cases

There is new information to check manually
Please open the table in Excel/OpenOffice…

# MassBank Record Workflow

## Infolist Editing (1) - Names

| | A | B | C | D | E | F | G | H | I | J | K | L | M |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | | id | dbcas | dbname | dataused | COMMEN | COMMEN | CH$NAME | CH$NAME | CH$NAME | CH$COMP | CH$FORM | CH$EXACT |
| 2 | 1 | 2819 | 15532-75-9 | 1-(3-Triflu | smiles | standard c | 2819 | 1-[3-(triflu | 1-(3-(Trifluorometh | | N/A; Envi | C11H13F3I | 230.1031 |
| 3 | 2 | 2821 | 50-48-6 | Amitriptyl | smiles | standard c | 2821 | | | | N/A; Envi | C20H23N | 277.183 |
| 4 | 3 | 2822 | 300-62-9 | Amphetar | smiles | standard c | 2822 | Amphetar | Amfetami | (1-methyl | N/A; Envi | C9H13N | 135.1048 |
| 5 | 4 | 2823 | 519-09-5 | Benzoylec | dbname | standard c | 2823 | Benzoylec | None | 3-benzoy | N/A; Envi | C16H19NC | 289.1314 |
| 6 | 5 | 2817 | 50-36-2 | Cocaine | dbname | standard c | 2817 | Cocaine | Unknown | (1S,3S,4R, | N/A; Envi | C17H21NC | 303.1471 |
| 7 | 6 | 2824 | 125-71-3 | Dextrome | smiles | standard c | 2824 | | | | N/A; Envi | C18H25NC | 271.1936 |
| 8 | 7 | 2825 | 30223-73-5 | EDDP (2-E | dbname | standard c | 2825 | | | | N/A; Envi | C20H23N | 277.183 |
| 9 | 8 | 2758 | 299-42-3 | Ephedrine | smiles | standard c | 2758 | Ephedrine | (1R,2S)-2-(methylam | | N/A; Envi | C10H15NC | 165.1154 |
| 10 | 9 | 2826 | 6740-88-1 | Ketamine | smiles | standard c | 2826 | | | | N/A; Envi | C13H16ClN | 237.092 |
| 11 | | | | | | | | | | | | | |

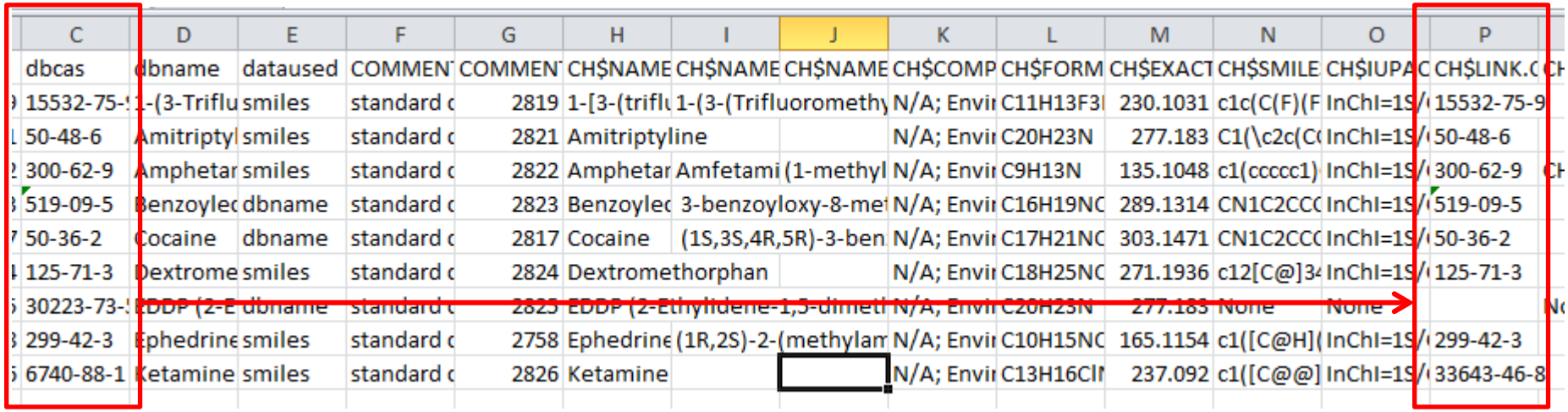
Fix up the names here

- “None”, “Unknown” unnecessary
- At least one name should be present (copy from dbname)

| | H | I | J | K |
|---|---|---|---|---|
| 1 | CH$NAME | CH$NAME | CH$NAME | CH$COMP |
| 2 | 1-[3-(triflu | 1-(3-(Trifluorometh | | N/A; Envi |
| 3 | Amitriptyline | | | N/A; Envi |
| 4 | Amphetar | Amfetami | (1-methyl | N/A; Envi |
| 5 | Benzoylec | 3-benzoyloxy-8-met | | N/A; Envi |
| 6 | Cocaine | (1S,3S,4R,5R)-3-ben | | N/A; Envi |
| 7 | Dextromethorphan | | | N/A; Envi |
| 8 | EDDP (2-Ethylidene-1,5-dimeth | | | N/A; Envi |
| 9 | Ephedrine | (1R,2S)-2-(methylam | | N/A; Envi |
| 10 | Ketamine | | | N/A; Envi |

# MassBank Record Workflow

Infolist Editing (2) – double-check CAS Numbers (if they were in original csv)



| C | D | E | F | G | H | I | J | K | L | M | N | O | P |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| dbcas | dbname | dataused | COMMEN | COMMEN | CH$NAME | CH$NAME | CH$NAME | CH$COMP | CH$FORM | CH$EXACT | CH$SMILE | CH$IUPA | CH$LINK.C |
| 15532-75-9 | 1-(3-Triflu | smiles | standard c | 2819 | 1-[3-(triflu | 1-(3-(Trifluorometh | | N/A; Envi | C11H13F3 | 230.1031 | c1c(C(F)(F | InChI=1S/ | 15532-75-9 |
| 50-48-6 | Amitriptyl | smiles | standard c | 2821 | Amitriptyline | | | N/A; Envi | C20H23N | 277.183 | C1(\c2c(C | InChI=1S/ | 50-48-6 |
| 300-62-9 | Amphetar | smiles | standard c | 2822 | Amphetar | Amfetami | (1-methyl | N/A; Envi | C9H13N | 135.1048 | c1(ccccc1) | InChI=1S/ | 300-62-9 |
| 519-09-5 | Benzoylec | dbname | standard c | 2823 | Benzoylec | 3-benzoyloxy-8-me | | N/A; Envi | C16H19NC | 289.1314 | CN1C2CCC | InChI=1S/ | 519-09-5 |
| 50-36-2 | Cocaine | dbname | standard c | 2817 | Cocaine | (1S,3S,4R,5R)-3-ben | | N/A; Envi | C17H21NC | 303.1471 | CN1C2CCC | InChI=1S/ | 50-36-2 |
| 125-71-3 | Dextrome | smiles | standard c | 2824 | Dextromethorphan | | | N/A; Envi | C18H25NC | 271.1936 | c12[C@]3 | InChI=1S/ | 125-71-3 |
| 30223-73-5 | EDDP (2-E | dbname | standard c | 2825 | EDDP (2-Ethylidene-1,5-dimeth | | | N/A; Envi | C20H23N | 277.183 | None | None | |
| 299-42-3 | Ephedrine | smiles | standard c | 2758 | Ephedrine | (1R,2S)-2-(methylam | | N/A; Envi | C10H15NC | 165.1154 | c1([C@H]( | InChI=1S/ | 299-42-3 |
| 6740-88-1 | Ketamine | smiles | standard c | 2826 | Ketamine | | | N/A; Envi | C13H16ClN | 237.092 | c1([C@@] | InChI=1S/ | 33643-46-8 |

# MassBank Record Workflow

Infolist Editing (3) – reality-checking entries & delete “rubbish”

| M | N | O | P | Q | R | S | T | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| CH$EXACT | CH$SMILE | CH$IUPAC | CH$LINK.C | CH$LINK.C | CH$LINK.H | CH$LINK.K | CH$LI | | | DER |
| 230.1031 | c1c(C(F)(F | InChI=1S/( | 15532-75-9 | | | | | | | |
| 277.183 | C1(\c2c(C( | InChI=1S/( | 50-48-6 | 2666 | | D07448 | | CID:2160 | KRMDCWI | 2075 |
| 135.1048 | c1(ccccc1) | InChI=1S/( | 300-62-9 | CHEBI | | | | CID:3007 | KWTSXDU | 13852819 |
| 289.1314 | CN1C2CCC | InChI=1S/( | 519-09-5 | | | | | CID:2340 | GVGYEFKI | 2250 |
| 303.1471 | CN1C2CCC | InChI=1S/( | 50-36-2 | 27958 | | C01416 | | CID:44622 | ZPUCINDJ | 10194104 |
| 271.1936 | c12[C@]34 | InChI=1S/( | 125-71-3 | | | D03742 | | CID:69161 | MKXZASY/ | 13109865 |
| 277.183 | None | None | | None | None | None | None | CID:None | None | |
| 165.1154 | c1([C@H]( | InChI=1S/( | 299-42-3 | 15407 | | C01575 | | CID:9294 | KWGRBVC | 8935 |
| 237.092 | c1([C@@] | InChI=1S/( | 33643-46-8 | | | | | | YQEZLKZA | 158414 |

Delete this

Check this - ChemSpider & PubChem ID usually similar

# MassBank Workflow

## Infolist Editing (4) – checking “Amphetamine” in ChemSpider


Retrieved ChemSpider ID is correct, so no changes necessary

# MassBank Record Workflow

Infolist Editing (5) – fill in missing entries for EDDP - search by CAS

| M | N | O | P | Q | R | S | T | U | V | W | X |
|---|---|---|---|---|---|---|---|---|---|---|---|
| CH$EXACT | CH$SMILE | CH$IUPAC | CH$LINK.C | CH$LINK.C | CH$LINK.H | CH$LINK.K | CH$LINK.L | CH$LINK.P | CH$LINK.I | CH$LINK.CHEMSPIDER | |
| 230.1031 | c1c(C(F)(F | InChI=1S/( | 15532-75-9 | | | | | CID:4296 | KKIMDKM | 4145 | |
| 277.183 | C1(\c2c(C( | InChI=1S/( | 50-48-6 | 2666 | | D07448 | | CID:2160 | KRMDCWI | 2075 | |
| 135.1048 | c | | | | | D07445 | | CID:3007 | KWTSXDU | 13852819 | |
| 289.1314 | C | InChI=1S/( | 519-09-5 | | | | | CID:2340 | GVGYEFKI | 2250 | |
| 303.1471 | C | InChI=1S/( | 50-36-2 | 27958 | | C01416 | | CID:44622 | ZPUCINDJ | 10194104 | |
| 271.1936 | c12[C@]3 | InChI=1S/( | 125-71-3 | | | D03742 | | CID:69161 | MKXZASY | 13109865 | |
| 277.183 | None | None | | None | None | None | None | CID:None | None | | |
| 165.1154 | c1([C@H]( | InChI=1S/( | 299-42-3 | 15407 | | C01575 | | CID:9294 | KWGRBVC | 8935 | |
| 237.092 | c1([C@@] | InChI=1S/( | 33643-46-8 | | | | | | YQEZLKZA | 158414 | |

Search ChemSpider and PubChem with CAS 30223-73-5 (in csv)

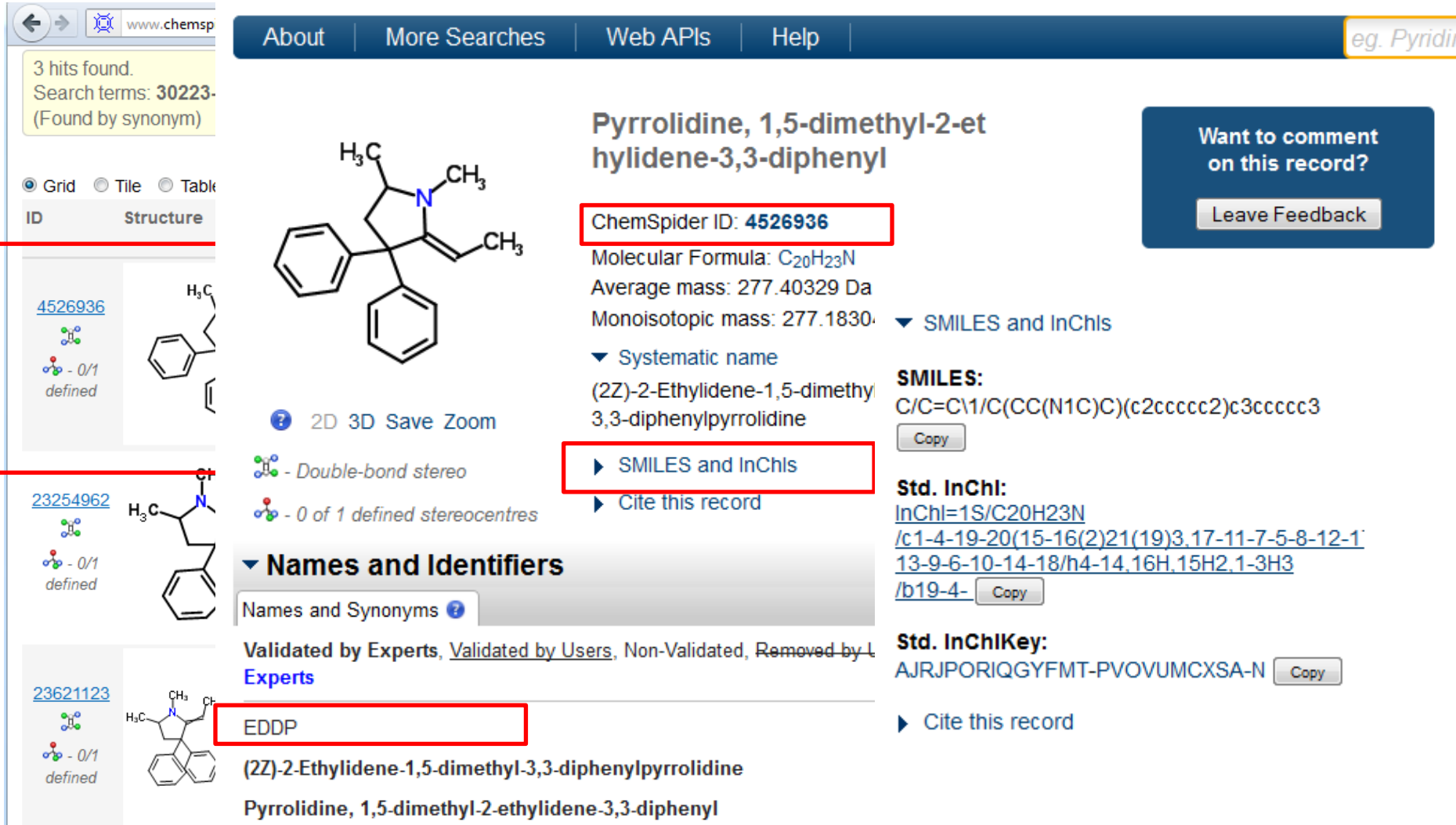
ChemSpider
The free chemical database

RSC | Advancing the Chemical Sciences

About | More Searches | Web APIs | Help | 30223-73-5 | Search

# MassBank Record Workflow

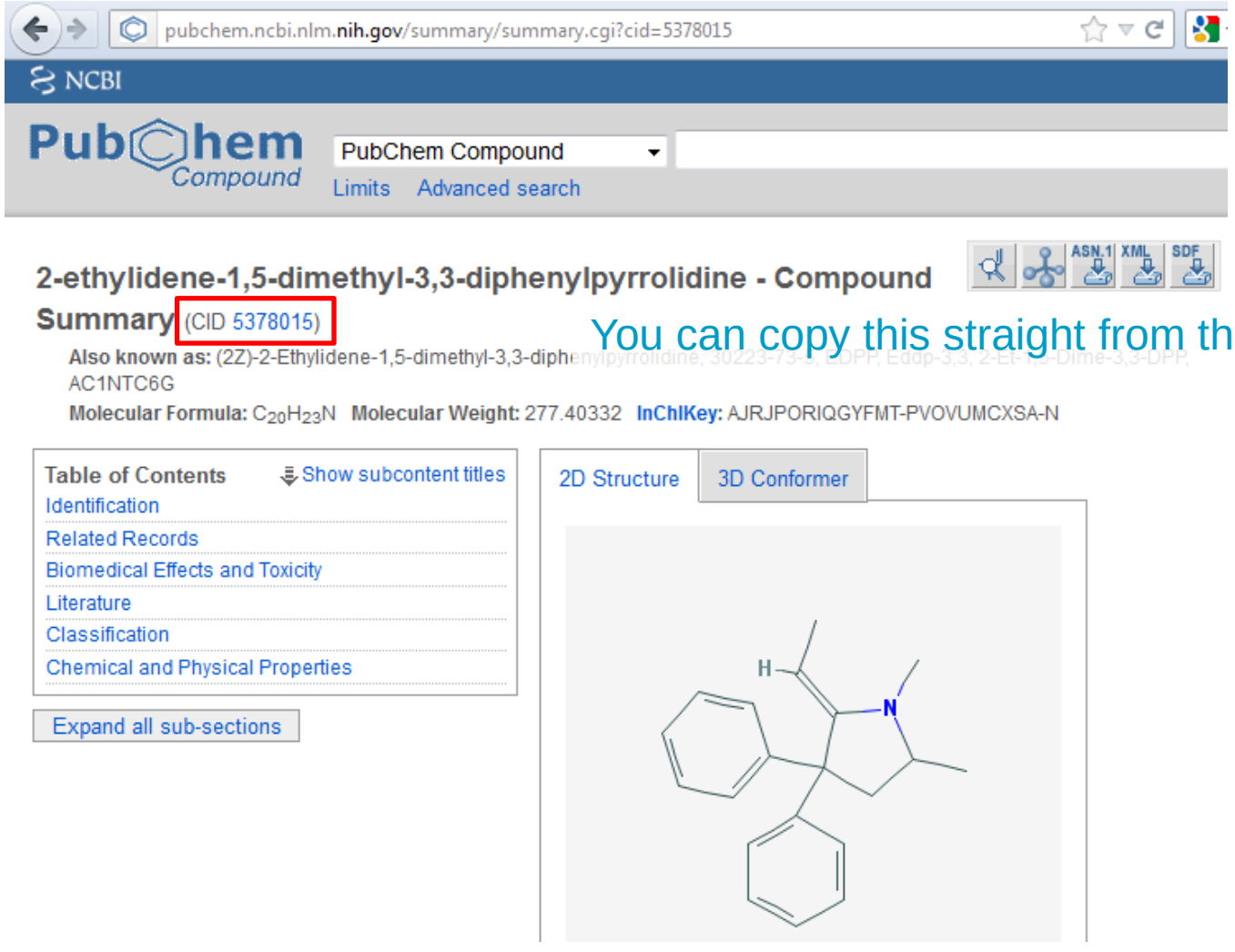
## Infolist Editing (5) – EDDP - search ChemSpider by CAS: 30223-73-5



# MassBank Record Workflow

=> http://pubchem.ncbi.nlm.nih.gov/search/search.cgi#

Infolist Editing (5) – EDDP - search PubChem by CAS: 30223-73-5



# MassBank Record Workflow

Infolist Editing (5) – fill in missing entries for EDDP – now filled in?

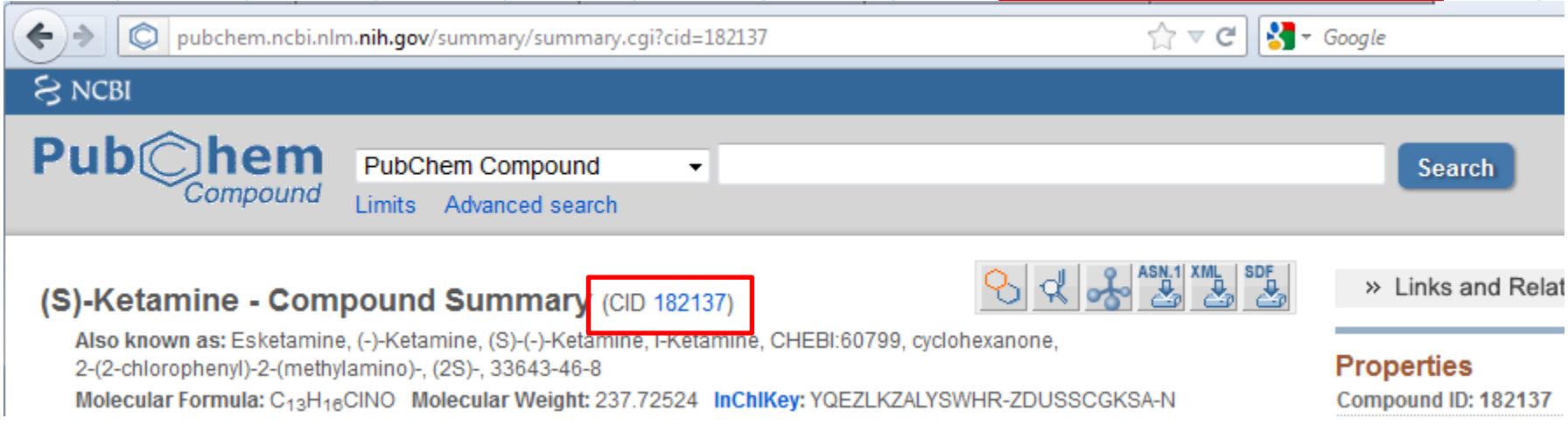
| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 135.1048 | c1(ccccc1) | InChI=1S/( | 300-62-9 | CHEBI | | D07445 | | CID:3007 | KWTSXDU | 13852819 | |
| 289.1314 | CN1C2CCC | InChI=1S/( | 519-09-5 | | | | | CID:2340 | GVGYEFKI | 2250 | |
| 303.1471 | CN1C2CCC | InChI=1S/( | 50-36-2 | 27958 | | C01416 | | CID:44622 | ZPUCINDJ | 10194104 | |
| 271.1936 | c12[C@]3 | InChI=1S/( | 125-71-3 | | | D03742 | | CID:69161 | MKXZASY | 13109865 | |
| 277.183 | None | None | | None | None | None | None | CID:None | None | | |
| 165.1154 | c1([C@H]( | InChI=1S/( | 299-42-3 | 15407 | | C01575 | | CID:9294 | KWGRBVC | 8935 | |
| 237.092 | c1([C@@] | InChI=1S/( | 33643-46-8 | | | | | | YQEZLKZA | 158414 | |

| M | N | O | P | Q | R | S | T | U | V | W | X |
|---|---|---|---|---|---|---|---|---|---|---|---|
| CH$EXACT | CH$SMILE | CH$IUPAC | CH$LINK.C | CH$LINK.C | CH$LINK.H | CH$LINK.K | CH$LINK.L | CH$LINK.F | CH$LINK.I | CH$LINK.CHEMSPIDER | |
| 230.1031 | c1c(C(F)(F | InChI=1S/( | 15532-75-9 | | | | | CID:4296 | KKIMDKM | 4145 | |
| 277.183 | C1(\c2c(C( | InChI=1S/( | 50-48-6 | 2666 | | D07448 | | CID:2160 | KRMDCWI | 2075 | |
| 135.1048 | c1(ccccc1) | InChI=1S/( | 300-62-9 | | | D07445 | | CID:3007 | KWTSXDU | 13852819 | |
| 289.1314 | CN1C2CCC | InChI=1S/( | 519-09-5 | | | | | CID:2340 | GVGYEFKI | 2250 | |
| 303.1471 | CN1C2CCC | InChI=1S/( | 50-36-2 | 27958 | | C01416 | | CID:44622 | ZPUCINDJ | 10194104 | |
| 271.1936 | c12[C@]3 | InChI=1S/( | 125-71-3 | | | D03742 | | CID:69161 | MKXZASY | 13109865 | |
| 277.183 | C/C=C\1/C | InChI=1S/( | 30223-73-5 | | | | | CID:53780 | AJRJPORI | 4526936 | |
| 165.1154 | c1([C@H]( | InChI=1S/( | 299-42-3 | 15407 | | C01575 | | CID:9294 | KWGRBVC | 8935 | |
| 237.092 | c1([C@@] | InChI=1S/( | 33643-46-8 | | | | | | YQEZLKZA | 158414 | |

# MassBank Record Workflow

## Infolist Editing (6) – Check missing PubChem entry?

| M | N | O | P | Q | R | S | T | U | V | W | X |
|---|---|---|---|---|---|---|---|---|---|---|---|
| CH$EXACT | CH$SMILE | CH$IUPAC | CH$LINK.C | CH$LINK.C | CH$LINK.H | CH$LINK.K | CH$LINK.L | CH$LINK.P | CH$LINK.I | CH$LINK.CHEMSPIDER | |
| 230.1031 | c1c(C(F)(F | InChI=1S/( | 15532-75-9 | | | | | CID:4296 | KKIMDKM | 4145 | |
| 277.183 | C1(\c2c(C( | InChI=1S/( | 50-48-6 | 2666 | | D07448 | | CID:2160 | KRMDCWI | 2075 | |
| 135.1048 | c1(ccccc1)( | InChI=1S/( | 300-62-9 | | | D07445 | | CID:3007 | KWTSXDU | 13852819 | |
| 289.1314 | CN1C2CCC( | InChI=1S/( | 519-09-5 | | | | | CID:2340 | GVGYEFKI | 2250 | |
| 303.1471 | CN1C2CCC( | InChI=1S/( | 50-36-2 | 27958 | | C01416 | | CID:44622 | ZPUCINDJ | 10194104 | |
| 271.1936 | c12[C@]3( | InChI=1S/( | 125-71-3 | | | D03742 | | CID:69161 | MKXZASYA | 13109865 | |
| 277.183 | C/C=C\1/C( | InChI=1S/( | 30223-73-5 | | | | | CID:53780 | AJRJPORI( | 4526936 | |
| 165.1154 | c1([C@H](( | InChI=1S/( | 299-42-3 | 15407 | | C01575 | | CID:9294 | KWGRBVC | 8935 | |
| 237.092 | c1([C@@]( | InChI=1S/( | 33643-46-8 | | | | | | YQEZLKZA | 158414 | |



# MassBank Record Workflow

Infolist editing finished! Save under a new name and restart workflow

- Moving Infolists into a separate Infolist folder is recommended!

```
100
101 mb <- resetInfolists(mb)
102 mb <- loadInfolists(mb, "C:/DATA/RMassBank/InfoLists_demo")
103 # NOTE: if you didn't want to correct the whole list, use this one:
104 #mb <- loadInfolists(mb, system.file("infolists", package="RMassBankData"))
105 # and start the workflow again - "no new data added" is a good sign!
106 mb <- mb
107
```

```
> mb <- resetInfolists(mb)
> mb <- loadInfolists(mb, "C:/DATA/RMassBank/InfoLists_demo")
> # NOTE: if you didn't want to correct the whole list, use this one:
> #mb <- loadInfolists(mb, system.file("infolists", package="RMassBankData"))
> # and start the workflow again - "no new data added" is a good sign!
> mb <- mbWorkflow(mb)
mbWorkflow: Step 1
mbWorkflow: Step 2
No new data added.
mbWorkflow: Step 3
mbWorkflow: Step 4
Compiling: 1_3_Chlorophenyl_piperazin_2818_pos.mzML
Compiling: 1_3_Trifluoromethylphenyl_piperazin_2819_pos.mzML
Compiling: 1_Benzylpiperazin_2820_pos.mzML
Compiling: Amitriptylin_2821_pos.mzML
Compiling: Amphetamin_2822_pos.mzML
Compiling: Benzoylecgonin_2823_pos.mzML
```

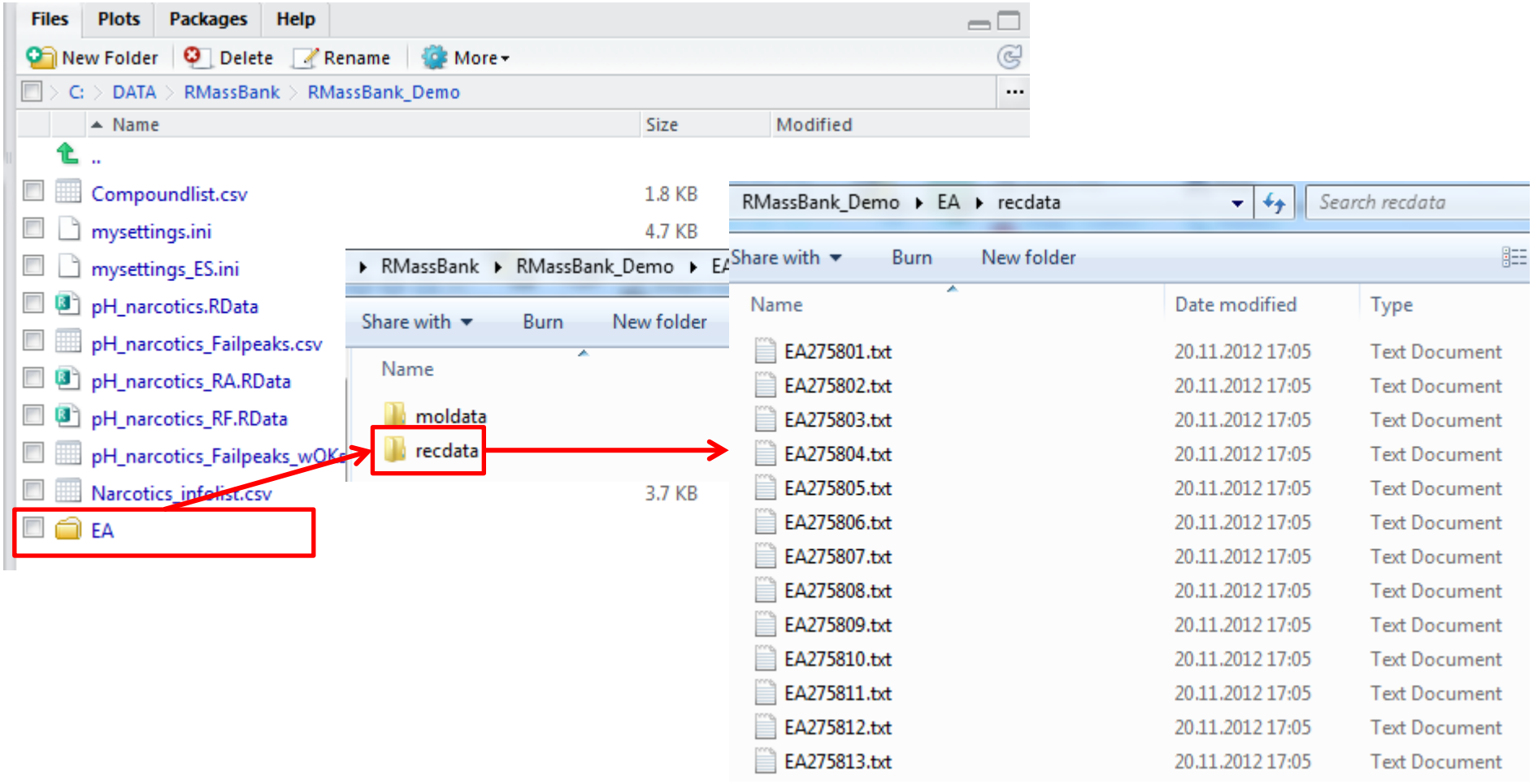
```
3
3
3
mbWorkflow: Step 5
mbWorkflow: Step 6
mbWorkflow: Step 7
mbWorkflow: Step 8
>
```

And that's it – DONE!

# MassBank Record Workflow

Take a look at your new records!

- You can open these in any text editor



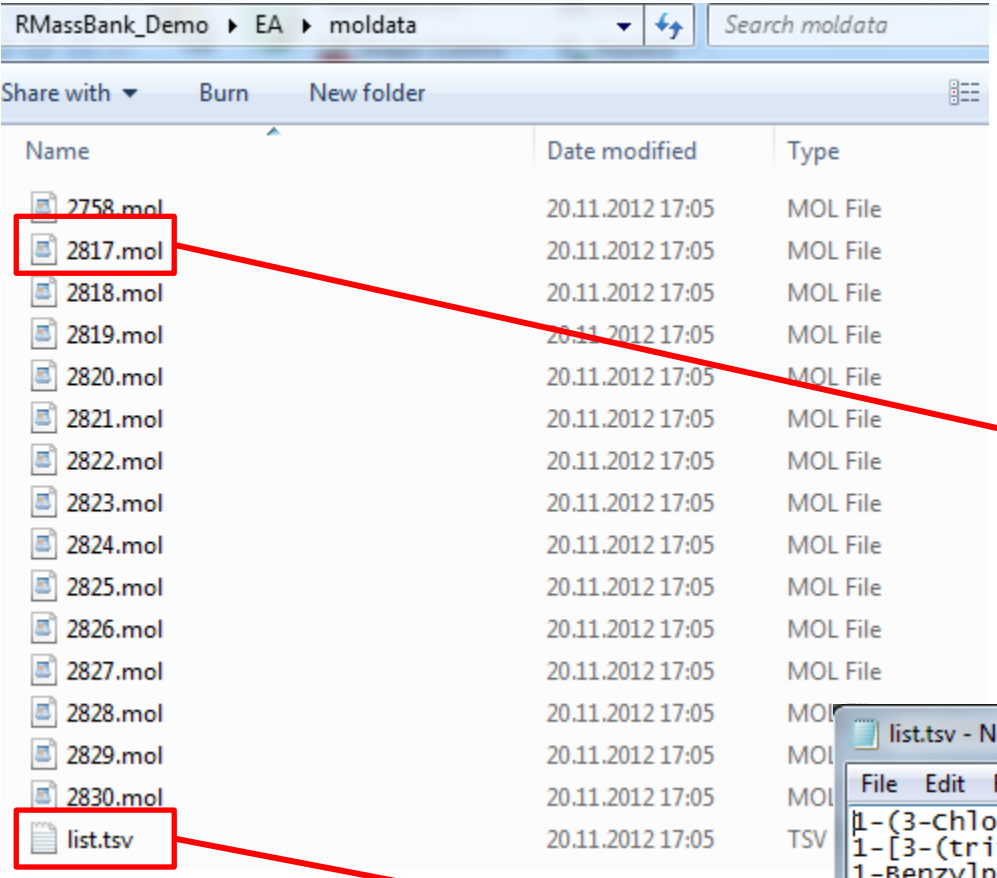
# MassBank Records “recdata”

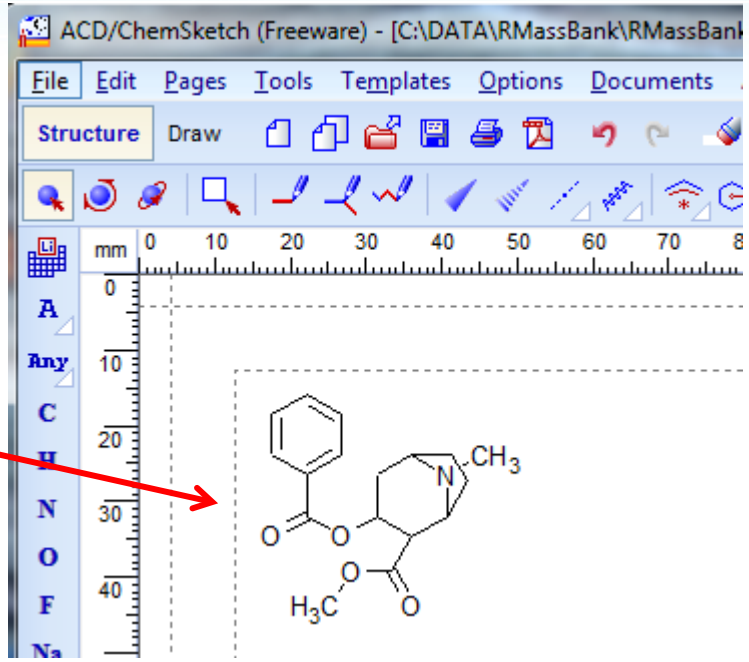
EA275805.txt - Notepad

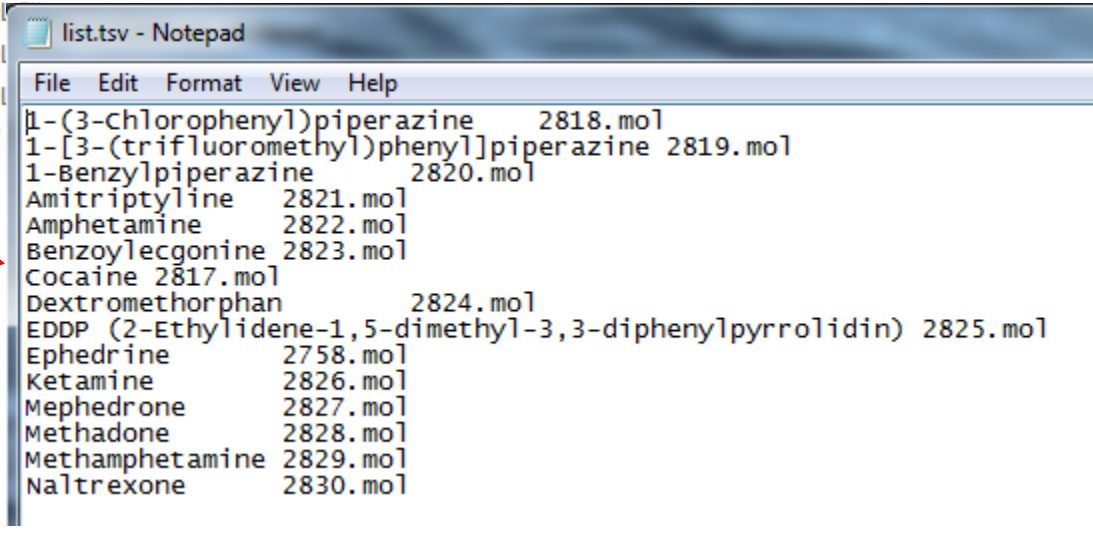
File Edit Format View Help

```
ACCESSION: EA275805
RECORD_TITLE: Ephedrine; LC-ESI-ITFT; MS2; 60%; R=7500; [M+H]+
DATE: 2012.11.20
AUTHORS: M. Stravs, E. Schymanski, H. Singer, Eawag
LICENSE: CC BY-SA
COPYRIGHT: Copyright (C) Eawag, 2012
COMMENT: CONFIDENCE standard compound
COMMENT: INTERNAL_ID 2758
CH$NAME: Ephedrine
CH$NAME: (1R,2S)-2-(methylamino)-1-phenyl-1-propanol
CH$COMPOUND_CLASS: N/A; Environmental Standard
CH$FORMULA: C10H15NO
CH$EXACT_MASS: 165.1154
CH$SMILES: c1([C@H]([C@@H](N
CH$IUPAC: InChI=1S/C10H15NO/
-/m0/s1
CH$LINK: CAS 299-42-3
CH$LINK: CHEBI 15407
CH$LINK: KEGG C01575
CH$LINK: PUBCHEM CID:9294
CH$LINK: INCHIKEY KWGRBVOPPL
CH$LINK: CHEMSPIDER 8935
AC$INSTRUMENT: LTQ Orbitrap
AC$INSTRUMENT_TYPE: LC-ESI-I
AC$MASS_SPECTROMETRY: MS_TYP
AC$MASS_SPECTROMETRY: IONIZA
AC$MASS_SPECTROMETRY: ION_MO
AC$MASS_SPECTROMETRY: FRAGME
AC$MASS_SPECTROMETRY: COLLIS
AC$MASS_SPECTROMETRY: RESOLU
AC$CHROMATOGRAPHY: COLUMN_NA
AC$CHROMATOGRAPHY: FLOW_GRAD
5/95 at 25 min, 90/10 at 25.
AC$CHROMATOGRAPHY: FLOW_RATE
AC$CHROMATOGRAPHY: RETENTION
AC$CHROMATOGRAPHY: SOLVENT A
AC$CHROMATOGRAPHY: SOLVENT B
MS$FOCUSED_ION: BASE_PEAK 16
MS$FOCUSED_ION: PRECURSOR_M/Z 166.1226
MS$FOCUSED_ION: PRECURSOR_TYPE [M+H]+
MS$DATA_PROCESSING: RECALIBRATE loess on assigned fragments and MS1
MS$DATA_PROCESSING: WHOLE RMassBank
PK$ANNOTATION: m/z num {formula mass error(ppm)}
  56.0495 1 C3H6N+ 56.0495 -0.46
  70.0652 1 C4H8N+ 70.0651 0.49
  91.0542 1 C7H7+ 91.0542 -0.51
  93.07 1 C7H9+ 93.0699 1.22
  115.0542 1 C9H7+ 115.0542 -0.58
  117.0699 1 C9H9+ 117.0699 0.11
  118.0647 1 C8H8N+ 118.0651 -3.86
  132.081 1 C9H10N+ 132.0808 1.55
  133.0886 1 C9H11N+ 133.0886 0.14
  135.0804 1 C9H11O+ 135.0804 -0.68
  148.1121 1 C10H14N+ 148.1121 -0.04
PK$NUM_PEAK: 11
PK$PEAK: m/z int. rel.int.
  56.0495 151421.4 24
  70.0652 152358.5 24
  91.0542 150051.6 24
  93.07 24399.9 3
  115.0542 229068.4 37
  117.0699 1548461.5 251
  118.0647 12712.5 2
  132.081 35161.8 5
  133.0886 1423545.6 231
  135.0804 94627.2 15
  148.1121 6145796.5 999
//
```

# MassBank Record Structures “moldata”



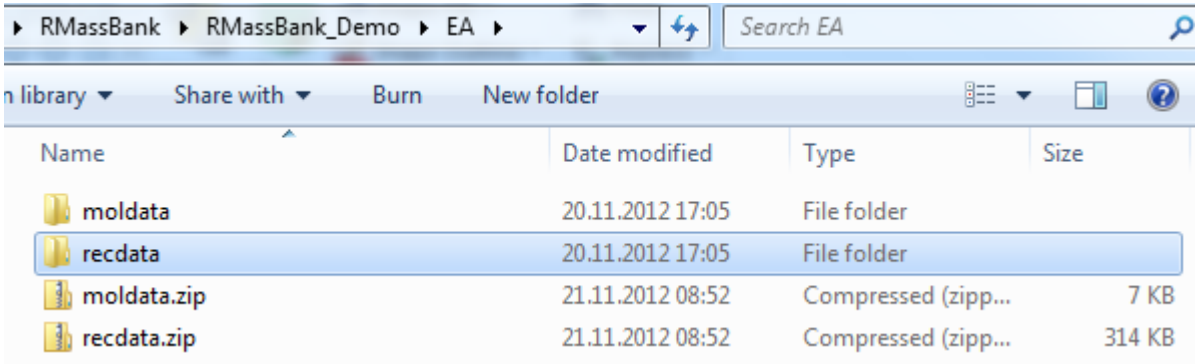




This is the list that connects structure and record

# Final Tasks for MassBank records:

Create “recdata.zip” and “moldata.zip” (e.g. with 7Zip)

- This wasn’t automated as this requires additional packages for Windows

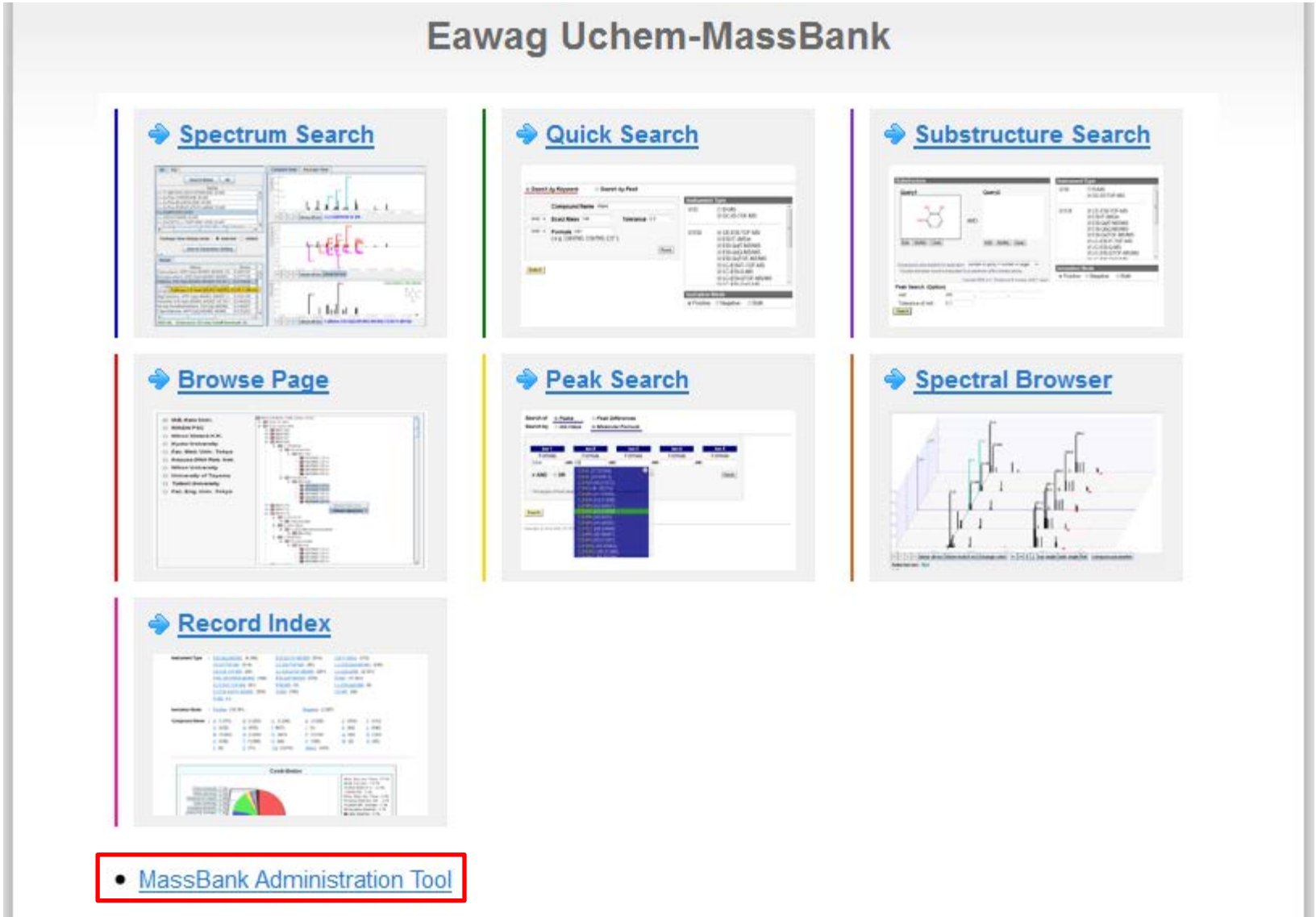



These zip files are now ready for upload to MassBank

- If these are new records for NORMAN MassBank
  - Email these to massbank@normandata.eu
- If these are private records, you can upload to your own MassBank
  - I’ll give a few screenshots in the next few slides

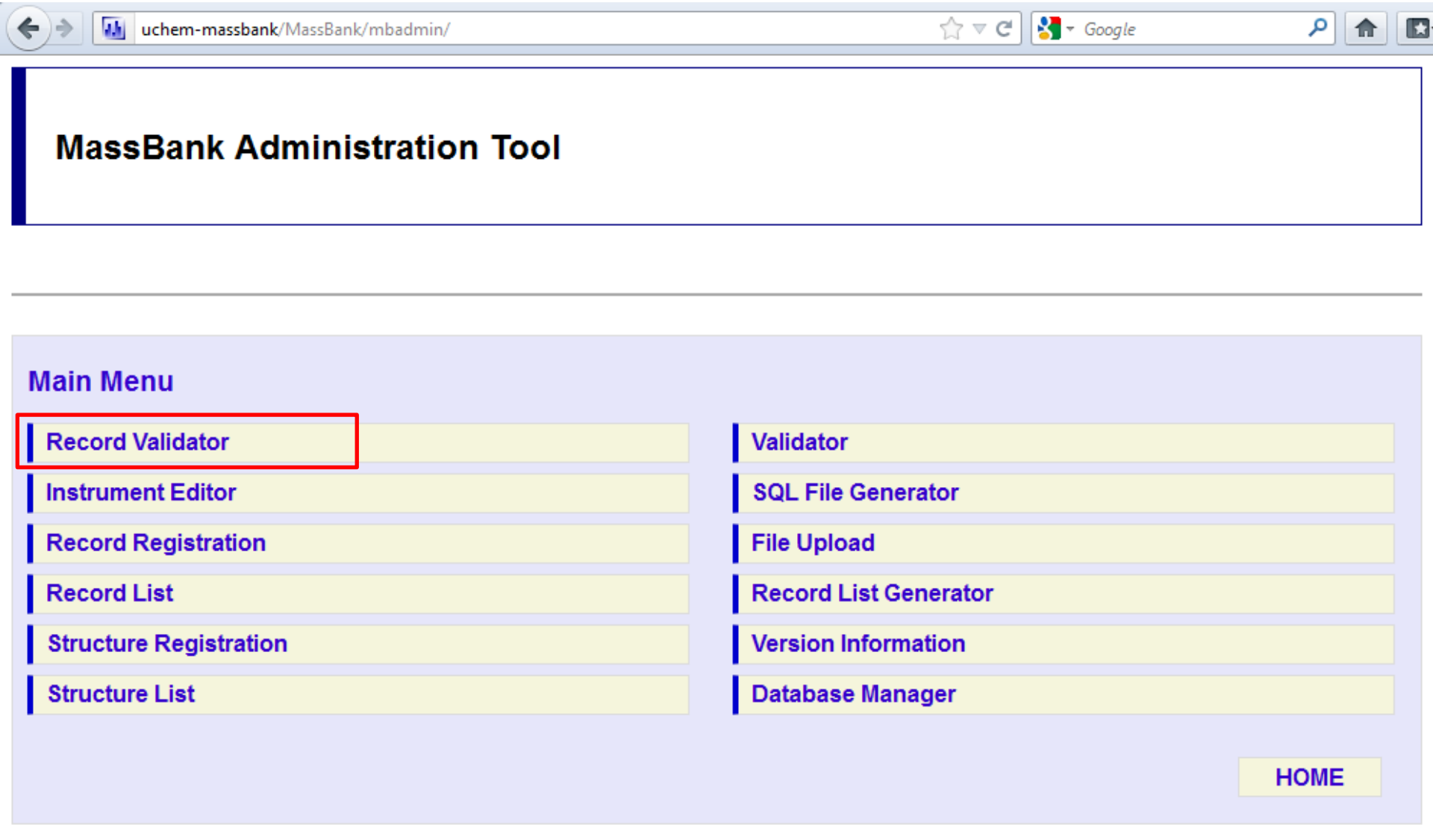
# Uploading MassBank Records

Example with “uchem-massbank” screenshots



# Uploading MassBank Records

Example with “uchem-massbank” screenshots



# Uploading MassBank Records

## Record validation


| Name | Status | Details |
|---|---|---|
| EA275801.txt | warn | value of required item [CH$COMPOUND_CLASS: ] is not compound class format. |
| EA275802.txt | warn | value of required item [CH$COMPOUND_CLASS: ] is not compound class format. |
| EA275803.txt | warn | value of required item [CH$COMPOUND_CLASS: ] is not compound class format. |
| EA275804.txt | warn | value of required item [CH$COMPOUND_CLASS: ] is not compound class format. |
| EA275805.txt | warn | value of required item [CH$COMPOUND_CLASS: ] is not compound class format. |
| EA275806.txt | warn | value of required item [CH$COMPOUND_CLASS: ] is not compound class format. |
| EA275807.txt | warn | value of required item [CH$COMPOUND_CLASS: ] is not compound class format. |
| EA275808.txt | warn | value of required item [CH$COMPOUND_CLASS: ] is not compound class format. |
| EA275809.txt | warn | value of required item [CH$COMPOUND_CLASS: ] is not compound class format. |
| EA275810.txt | warn | value of required item [CH$COMPOUND_CLASS: ] is not compound class format. |
| EA275811.txt | warn | value of required item [CH$COMPOUND_CLASS: ] is not compound class format. |

Technically, this should define if the compound is a natural product or not; there is no way to automate this (unambiguously) – hence this warning

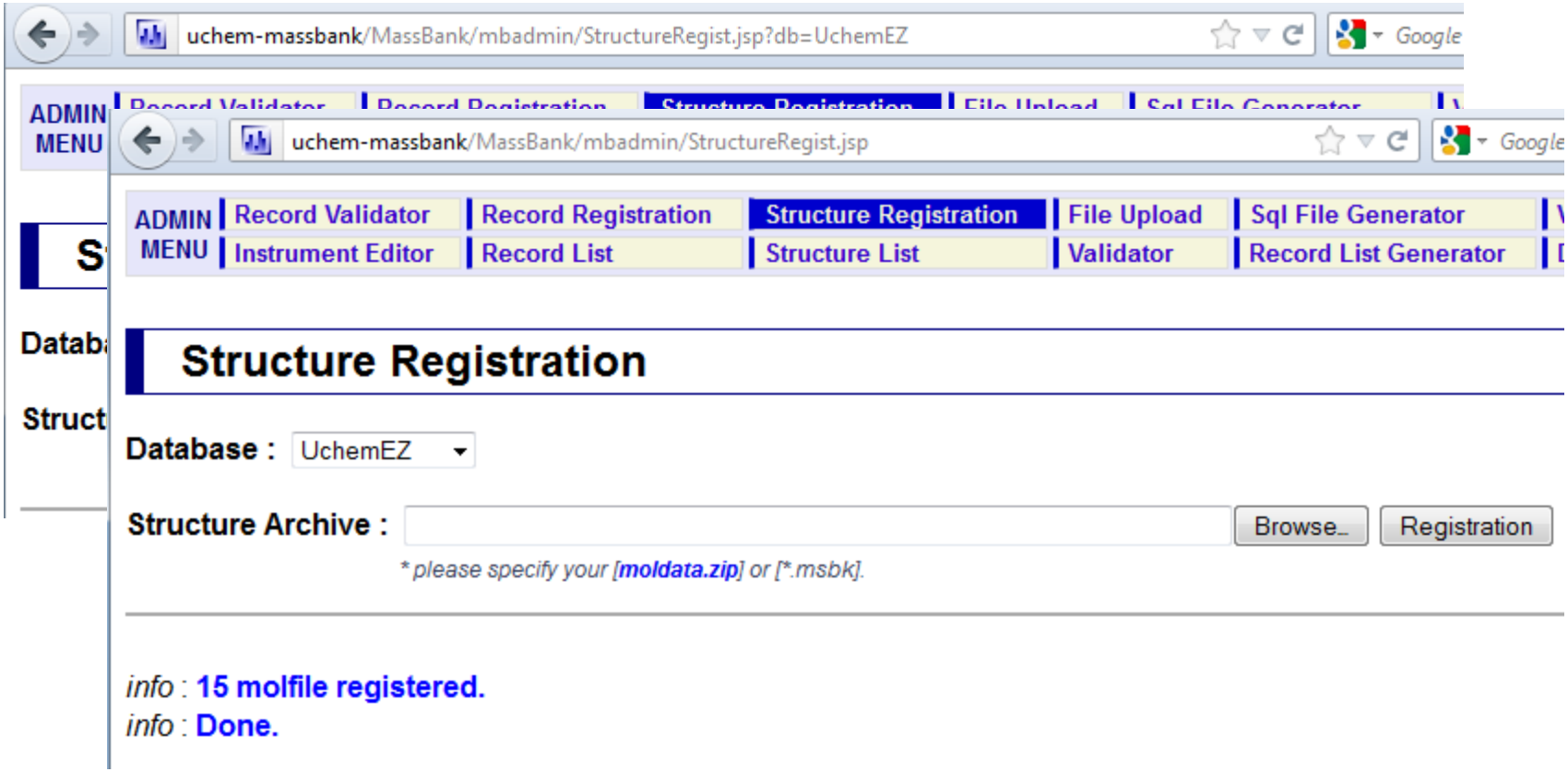
We print “N/A: Environmental Standard

# Uploading MassBank Records

## Record Registration

# Uploading MassBank Records

## Structure Registration



# Records Registered

Browse MassBank to see what they look like!

## Record Index

mass calculator

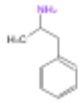
Home | Spectrum | Quick | Peak | Substructure | Browser | Browse | Index | MassBank ID: Go

**Contributor** :

CASMI (42) | Chubu Univ. (2,628) | EA Uchem Orbi (5,132)

EA Uchem Orbi Test (641) | EQ Uchem Q Ex (1,262) | Eawag Uchem Adducts (153)

Fukuyama Univ. (340) | IMM, CAMS & PUMC, China (192) | IPB Halle (528)

| | | | | |
|---|---|---|---|---|
| | **Amphetamine** 14 spectra | **C9H13N**  | **135.10480** | |
| | LC-ESI-ITFT; MS2; 15%; R=15000; [M+H]+ | | | EA282208 |
| | LC-ESI-ITFT; MS2; 15%; R=7500; [M+H]+ | | | EA282202 |
| | LC-ESI-ITFT; MS2; 30%; R=15000; [M+H]+ | | | EA282209 |
| | LC-ESI-ITFT; MS2; 30%; R=7500; [M+H]+ | | | EA282203 |
| | LC-ESI-ITFT; MS2; 35%; R=15000; [M+H]+ | | | EA282214 |
| | LC-ESI-ITFT; MS2; 35%; R=7500; [M+H]+ | | | EA282201 |
| | LC-ESI-ITFT; MS2; 45%; R=15000; [M+H]+ | | | EA282210 |
| | LC-ESI-ITFT; MS2; 45%; R=7500; [M+H]+ | | | EA282204 |
| | LC-ESI-ITFT; MS2; 60%; R=15000; [M+H]+ | | | EA282211 |
| | LC-ESI-ITFT; MS2; 60%; R=7500; [M+H]+ | | | EA282205 |
| | LC-ESI-ITFT; MS2; 75%; R=15000; [M+H]+ | | | EA282212 |
| | LC-ESI-ITFT; MS2; 75%; R=7500; [M+H]+ | | | EA282206 |
| | LC-ESI-ITFT; MS2; 90%; R=15000; [M+H]+ | | | EA282213 |
| | LC-ESI-ITFT; MS2; 90%; R=7500; [M+H]+ | | | EA282207 |

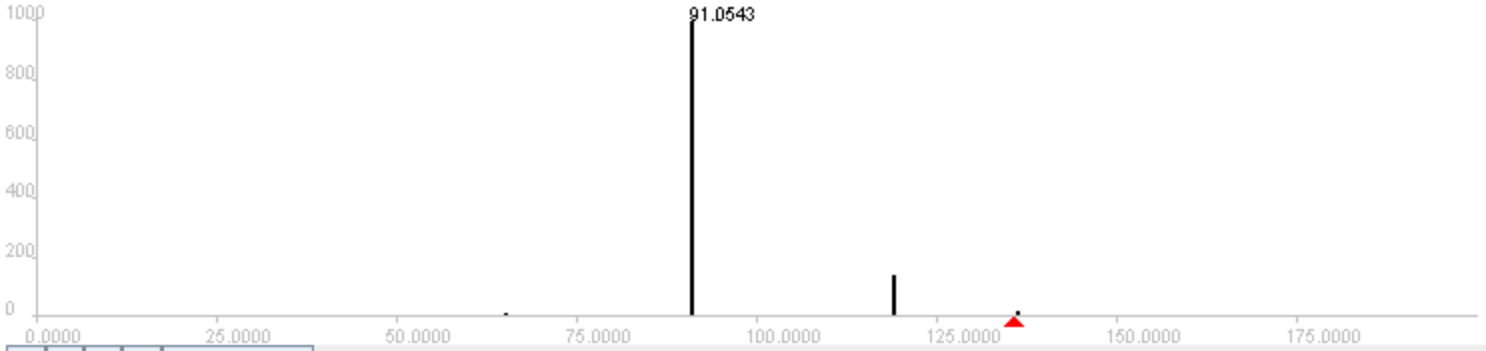
uchem-massbank/MassBank/jsp/Dispatcher.jsp?type=disp&id=EA282210&site=1

# MassBank Record: EA282210

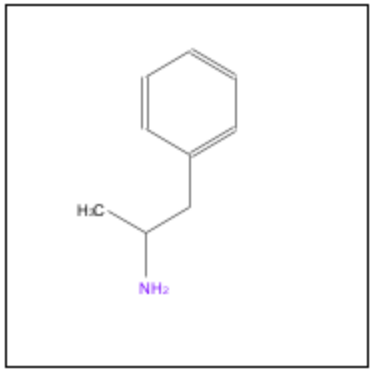
Home | Spectrum | Quick | Peak | Substructure | Browser | Browse | Index | MassBank ID: Go

## Amphetamine; LC-ESI-ITFT; MS2; 45%; R=15000; [M+H]+

Mass Spectrum




Chemical Structure




```
ACCESSION: EA282210
RECORD_TITLE: Amphetamine; LC-ESI-ITFT; MS2; 45%; R=15000; [M+H]+
DATE: 2012.11.20
AUTHORS: M. Stravs, E. Schymanski, H. Singer, Eawag
LICENSE: CC BY-SA
COPYRIGHT: Copyright (C) Eawag, 2012
COMMENT: CONFIDENCE standard compound
COMMENT: INTERNAL_ID 2822

CH$NAME: Amphetamine
CH$NAME: Amfetamine
```

# Take-Home Messages: RMassBank Demo

That was a 60+-slide introduction to RMassBank

- Install (and update) RMassBank and necessary programs
- Run RMassBank for trial data
  - Get a basic understanding for the workflow
- Understand the manual checking required
  - “Fail peaks” – checking these in raw data
  - Automatic annotation with CTS – searching/editing this data
- Generate records with RMassBank
- A quick impression how to upload these records

It will be impossible to remember everything!

- vignette(“RMassBank”) – this contains an extensive explanation
- Stravs et al. 2012, *J. Mass Spectrom.*, DOI: 10.1002/jms.3131

# Acknowledgements


- Marie-Curie Post Doctoral Fellowship (E. Schymanski),
- Eawag Discretionary Funds (M. Stravs)
- Coauthors: ***Michael Stravs***, Heinz Singer, Juliane Hollender
- R & MassBank help: Steffen Neumann, Michael Gerlich, Erik Müller, (IPB)
- Discussions: Matthias Ruff, Martin Loos (Eawag);

  Tobias Schulze, Martin Krauss, Werner Brack (UFZ)
- MassBank & Naming Rights: Prof. Takaaki Nishioka
- NORMAN Association


- And thank you all for listening!

RMassBank

eawag aquatic research

# Any Questions?

**massbank@normandata.eu (Tobias – record upload)**

**massbank@eawag.ch (RMassBank: E. Schymanski, M. Stravs)**

**http://metabolomics-forum.com/viewforum.php?f=29**

**emma.schymanski@eawag.ch**

**DOI: 10.1002/jms.3131**

**All Details Contained Within!**