After this there is not turning back

THIS IS YOUR LAST CHANCE....AFTER THIS, THERE IS NO TURNING BACK. YOU TAKE THE **BLUE** PILL - THE STORY IS ENDS,YOU WAKE UP IN YOUR BED,
AND BELIEVE WHATEVER YOU WANT TO BELIEVE.

YOU TAKE THE **RED** PILL, ...

YOU STAY IN WONDERLAND, AND I SHOW YOU, HOW DEEP THE RABBIT-HOLE GOES.


Welcome to the desert of the real....


The only good
is knowledge
The only evil
is ignorance

Plitvička jezera, Croatia

Kopački rit, Croatia

There is a difference between knowing the path and walking the path...

Kornati, Croatia

Statistical approaches for data-mining and non-target selection

Croatian Waters, Central water management laboratory
Draženka Stipaničev, Siniša Repec
(drazenka.stipanicev@voda.hr, sinisa.repec@voda.hr)

17.09.2014. Eawag: Dübendorf, Switzerland

-the challenging task for environmental researches is screening of surface waters because different organic substances present in surface waters are difficult to characterize by chemical analyses
-these complex mixtures occurs at a very low concentrations and requires both a specific analytical methods and instruments for identification

AGILENT 6550 i-Funnel UHPLC/QTOF-MS:
-40000 FWHM,mass accuracy <1 ppm
-satisfactory sensitivity in full-acquisition mode for the rapid screening and quantification of multi-class organic pollutants in water, with little sample manipulation open new multiple possibilities for challenging environmental analyses

- Impurity analysis
- Components introduced in water : where, how and from whom?
- What components change
- Evaluate impact of pollutants on the environment
- Which components are not known

QUESTIONS

Statistical Software for Comparing Data Sets

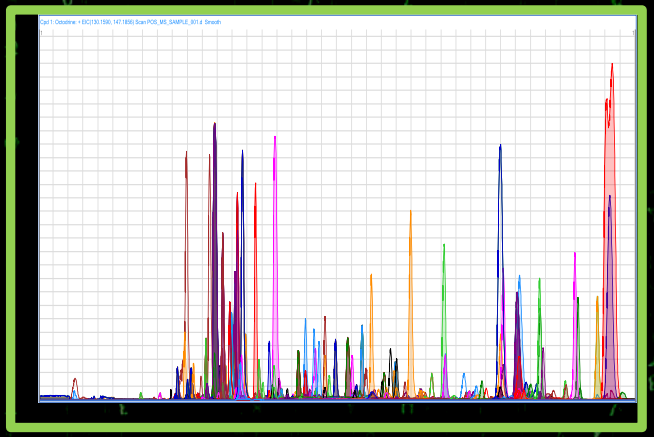
SAMPLE WITHOUT SAMPLE PREPARATION	Q-TOF MS AQUISITION full scan	MASS PROFINDER SOFTWARE	MASS PROFILER PROFESSIONAL with ID Browser	Q-TOF auto-MSMS fragmentation

Only Filtering ESI +

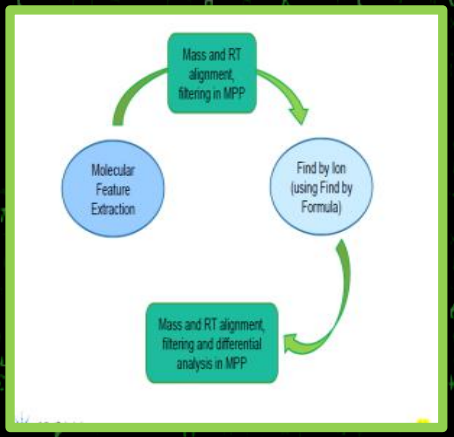
Advanced Batch Feature Extraction

Data Filtering Statistical Analyses Data Visualization

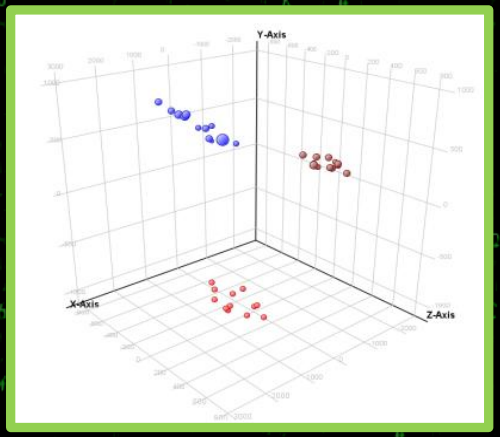
CE



Overlay EIC



Workflow



PCA

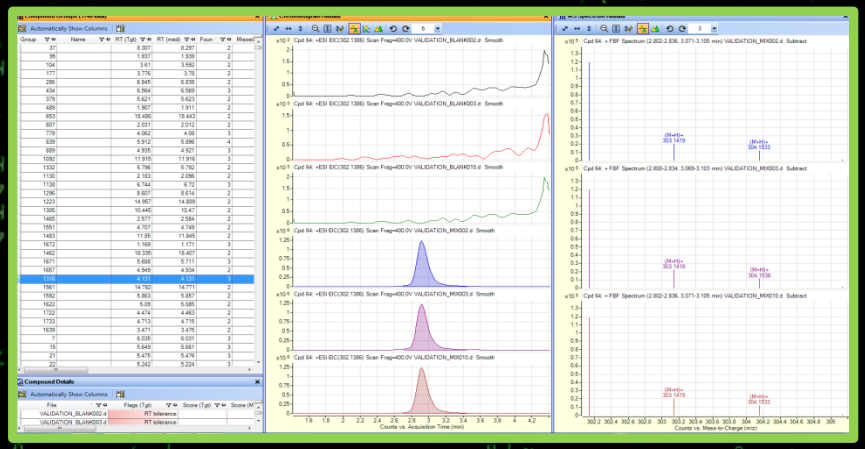
MassHunter Profinder Feature Finding

Molecular Feature Extraction

MFE → rMFE → Find by Ion

- From the raw data finds co-eluting ions that are related: (isotopes, adducts - such as Na+ / K+...), and dimers
- Filters noise
- Creates a compound chromatogram for the group of ions
- Sums all ion signals into one value: one Feature = one compound

- Batch processing of large, complex accurate mass LC/MS data
- Find by Ion reduces false negatives
- Allows manual editing of compounds
- Reduces the number of false positives and false negatives



rMFE → Filter 1 → rFind by Ion → Filter 2 → **Final Compound List**

improves the quality of target list for Find by Ion

improves the quality of final compound group list, amount of manual cleanup is reduced

Mass Profiler Professional

Statistical Software for Comparing Data Sets

multi-variate dataset is reduced (by filtering and statistical analyses) to a small set of significant and relevant compounds for further evaluation

Data Filtering

Filter by mass, retention time, frequency, abundance, mass error and alignment, using Venn Diagrams

Statistical Analyses

t-tests/ANOVA, Fold-change, Clustering, Find Similar Entities, Principal Component Analysis

Data Visualization

Scatter plots, Profile plots, Matrix plots, Box- Whisker plots, Histograms, Heat Maps, Venn Diagrams, Data spreadsheets

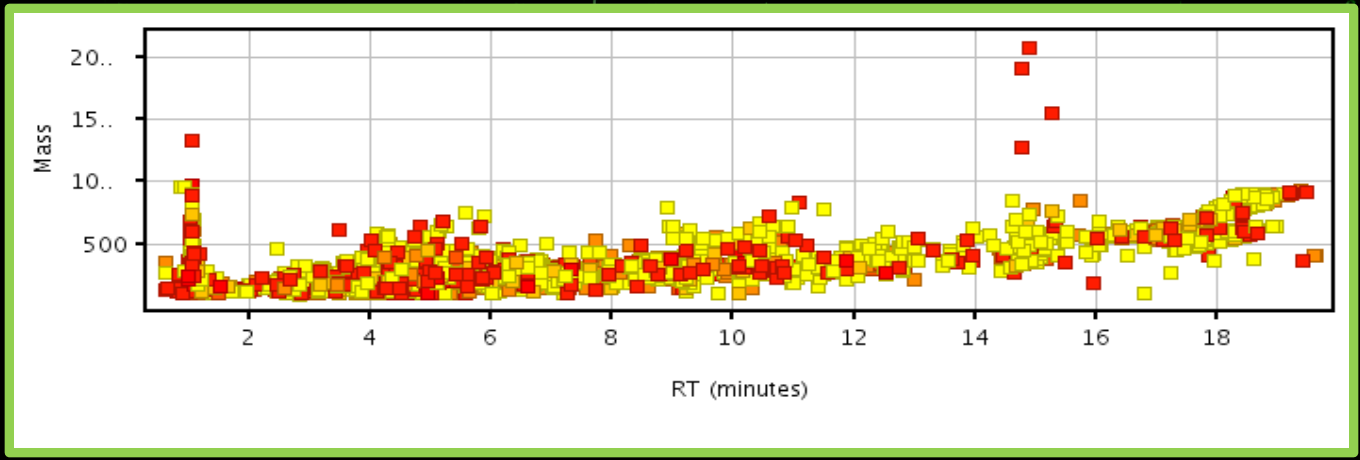
Experiment grouping | Interpretations | QC filters | Statistical analysis | Fold change | Results interpretation

Covariance analysis (PCA, Clustering)

Proof of MPP principle in spike-in experiment

VALIDATION

Experiment creation – data input and compound alignment



Each square represents a compound.

Plot presents information about the chromatography.

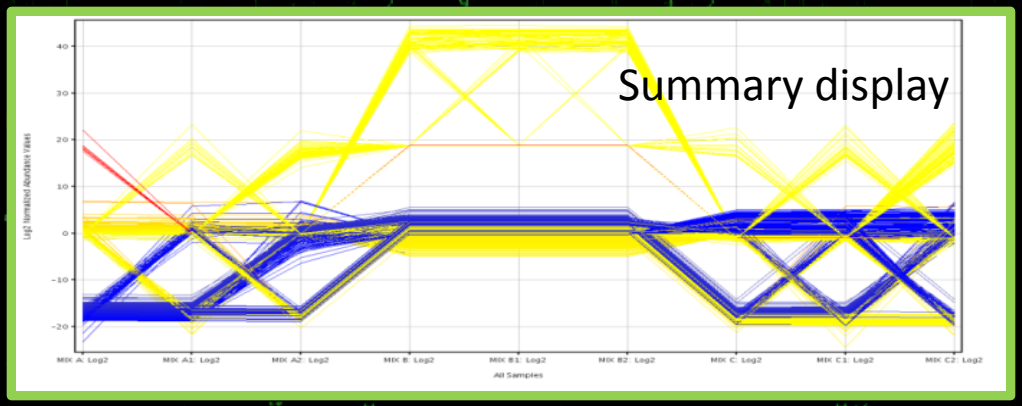
Mix A surface water (river Slunjcica)

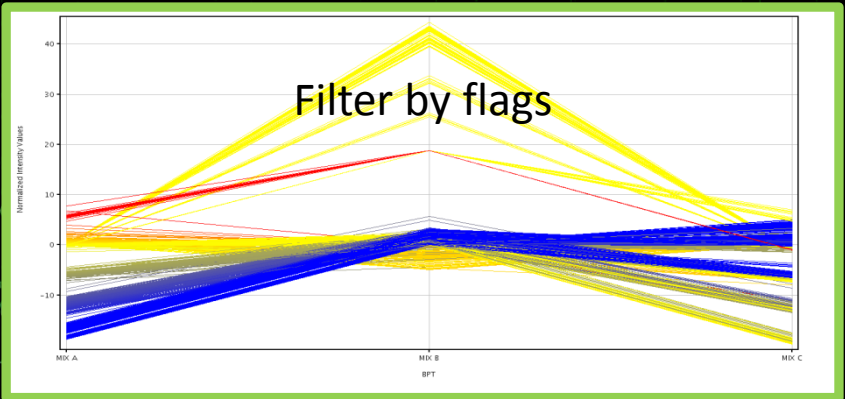
Mix B surface water + 30 spiked pest (50 ng/l)

Mix C surface water + 80 spiked For-Tox (50 ng/l)

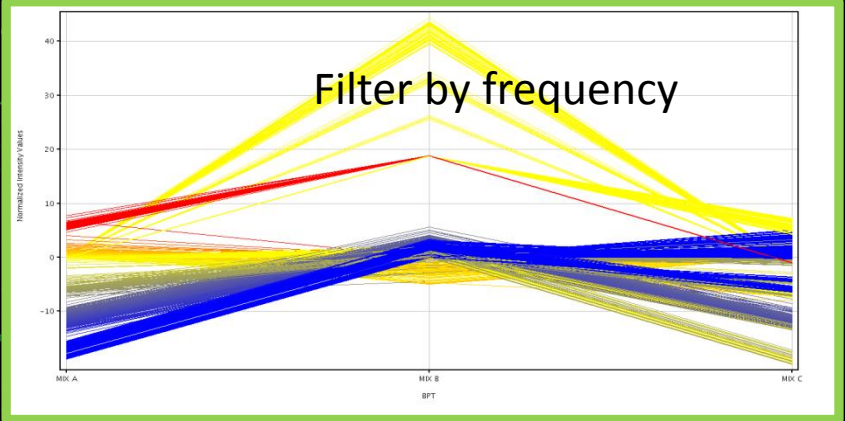
Experiment creation

- **aboundance normalization** (percentile shift 75 - 0 values are treated equally no positive fold changes)
- **baselining options** (to median of all samples treat all compounds equally regardless of their intensity) and 0 abundance point is in the middle of the plot – the median Log2 value for an entity in all samples is substracted from the Log2 abundance value of that entity in each sample
- after data import analysis steps starts
- **summary displays** all aligned compounds (presented in colored lines) and colour indicates the relative abundance of each compound. Indicated abundance is for the first sample compared to all other samples, but the higher aboundance does not necessarily mean that these compound are high in aboundance relative to overall aboundance.

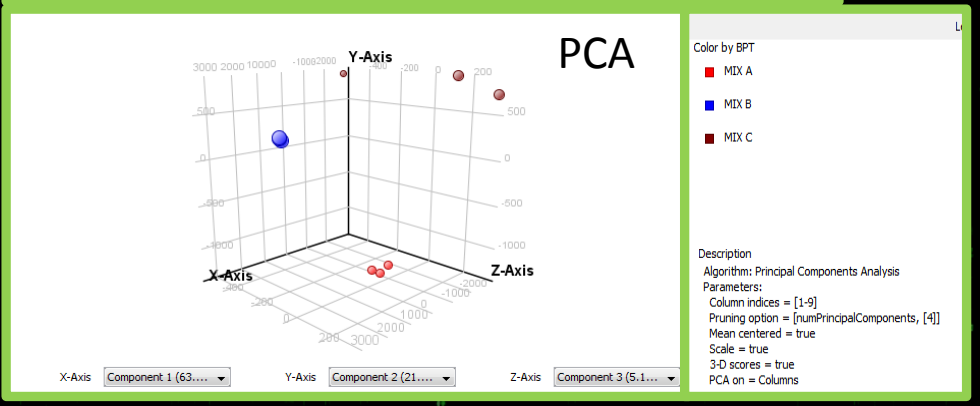




Mean of eliminating less reliable compounds (compounds not found in at least 1 of 9 total sampes). Filtering out entities based on flags A,P,M.



Based on frequency of accurance across samples. Limit analysis to entities present in minimal number of samples.



QC on samples ,we can assess data quality as indicated by covariance clustering using a principal component analysis

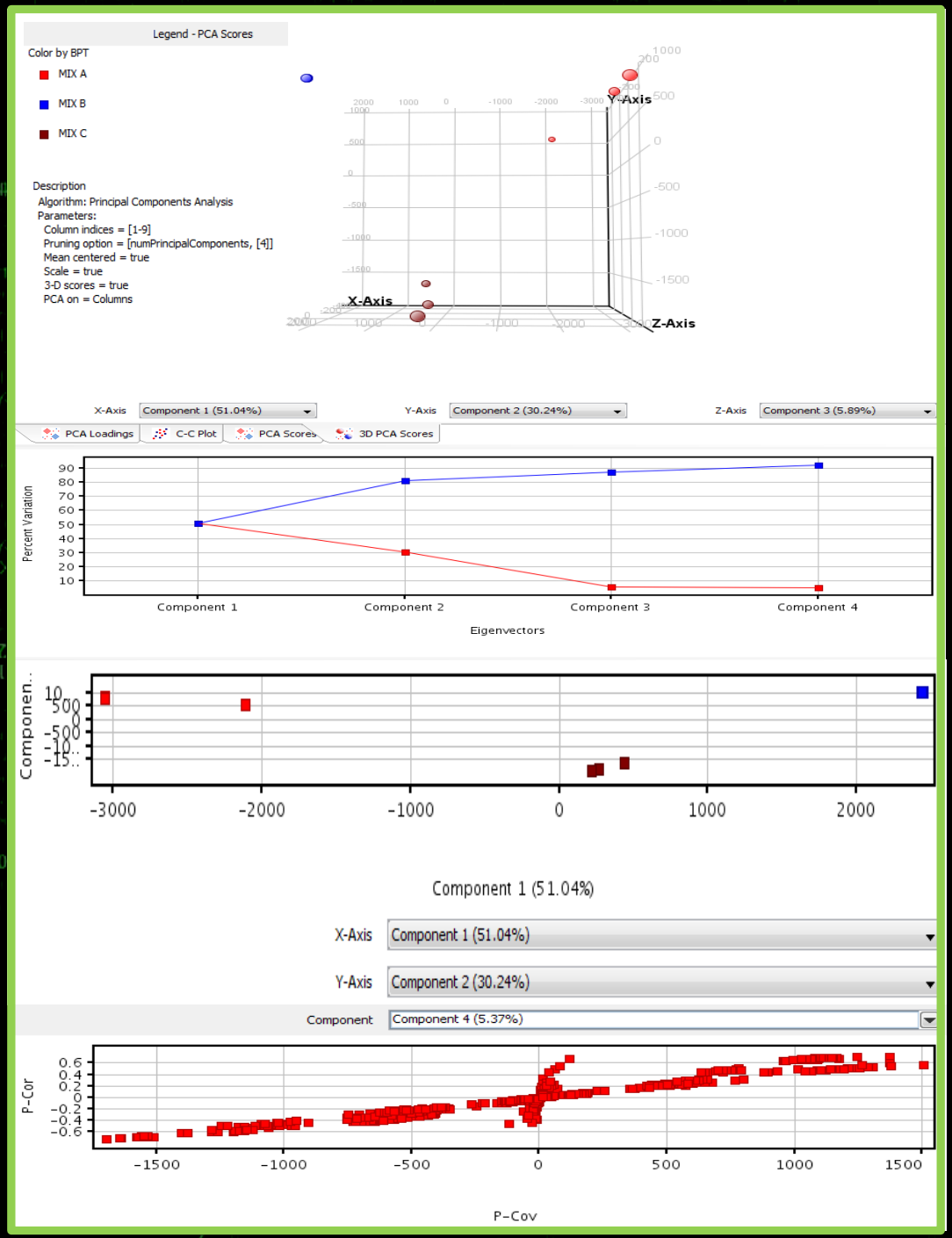
Principal Component Analysis

-reduces multi dimensional data to a few dimensions

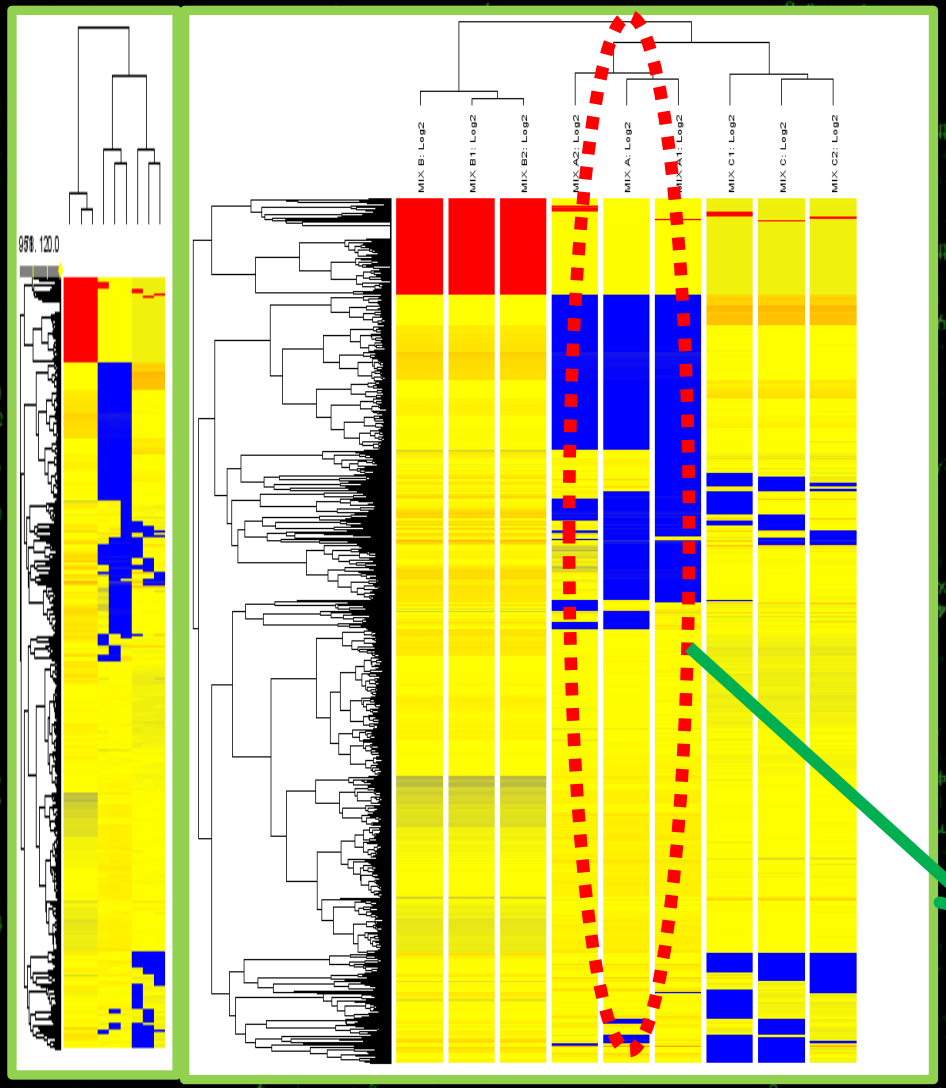
-reveal simplified structures

-simple method can be considered a positive feature because the answer is unique and independant of the user

-visual way to to explore variance and identify patterns in data



Unsupervised hierarchical clustering (measurement of variability)



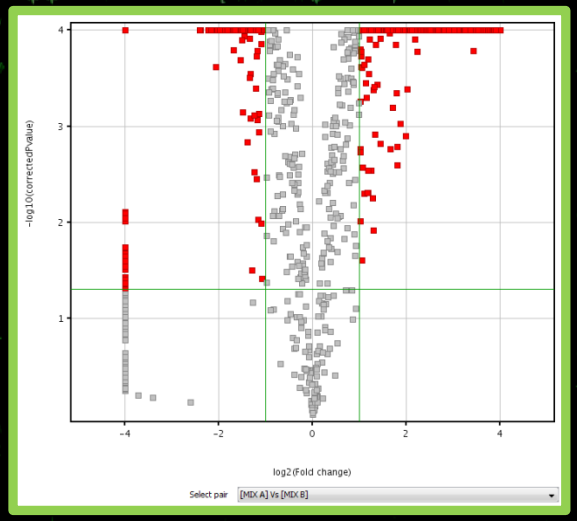
- can help identify major sources of variation that have influenced sample covariance
- analysis tool that groups samples or sample groups according to their similarities
- results are displayed in a dendrogram
- compounds that contribute to clustering can be isolated and saved on entity list
- samples connected to the same node are more alike than samples connected to other nodes

Significance analysis – VOLCANO PLOT

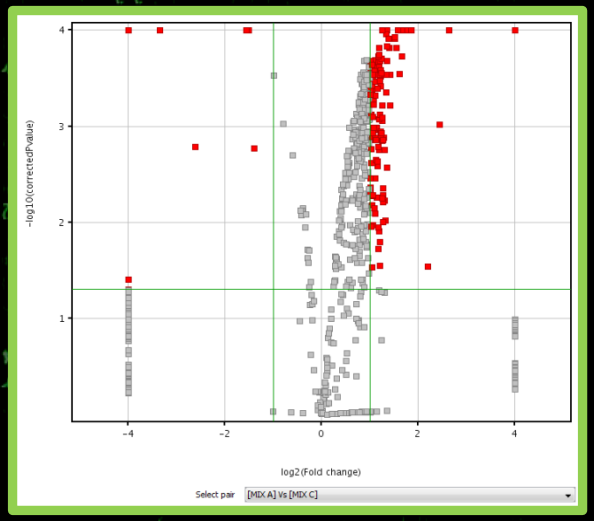
VOLCANO
- because in very complex data sets with many measurable differences between groups the up- and down-regulated entities appear on both sides on center and form an image like erupting volcano
- simultaneously applies t-test and a fold change filter

Compounds above the green horizontal line (p value cutoff) and outside one of the vertical green lines (fold change cutoff) are colored red indicating they pass both tests

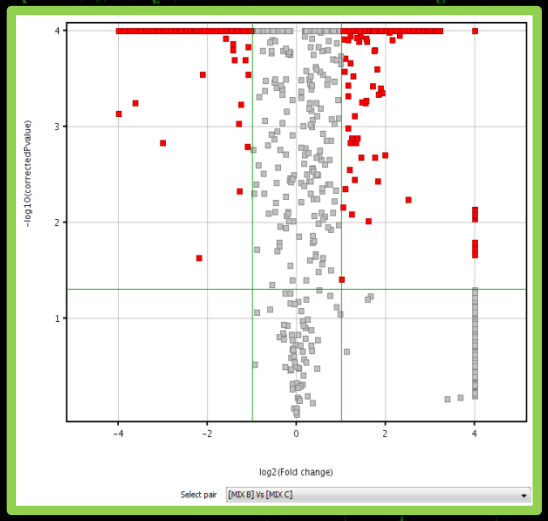
mix A vs mix B



mix A vs mix C

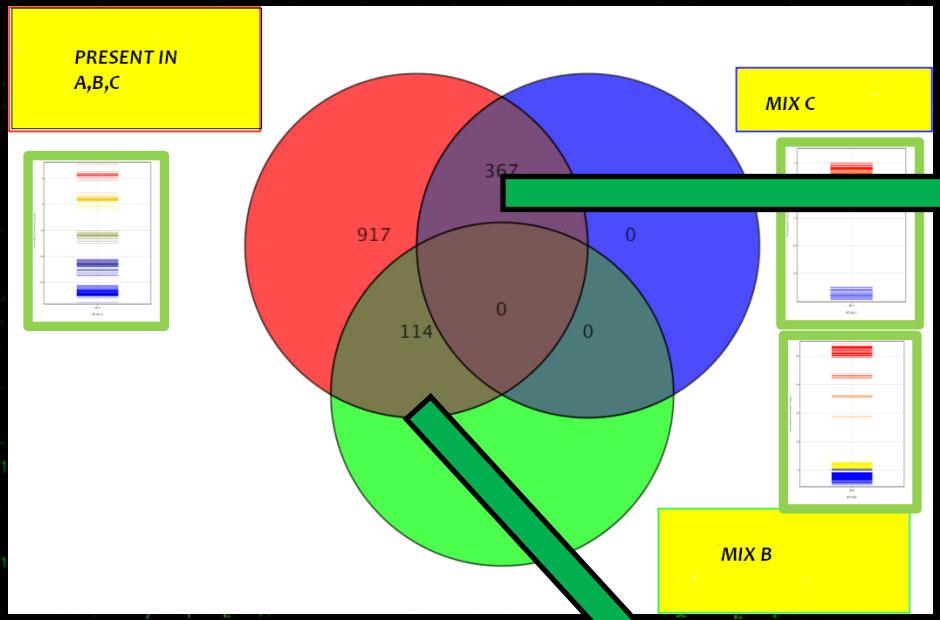


mix B vs mix C



Venn diagram

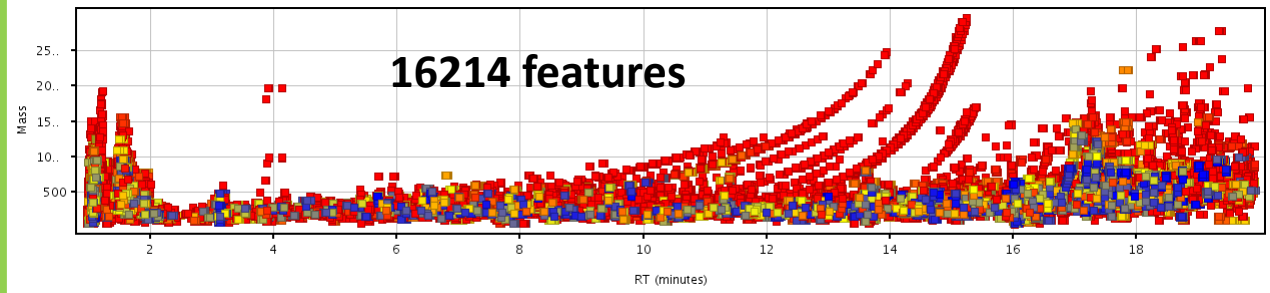
- for creation list of entities that are unique to a specific condition or common to multiple conditions

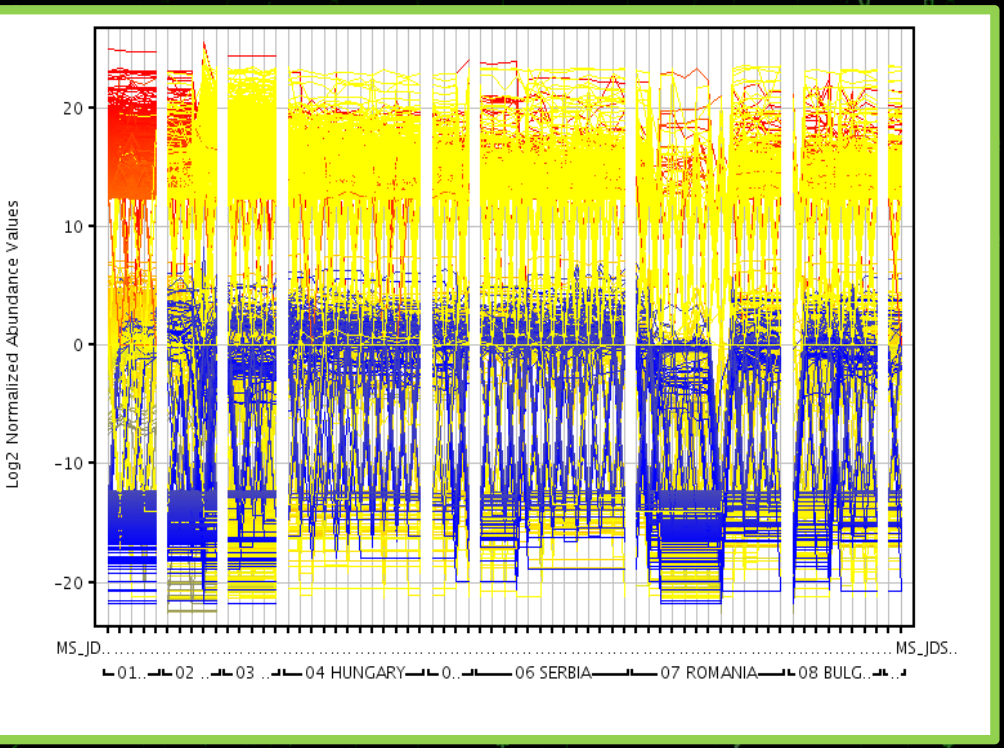


- autoMSMS confirmed 80 For- Tox compounds

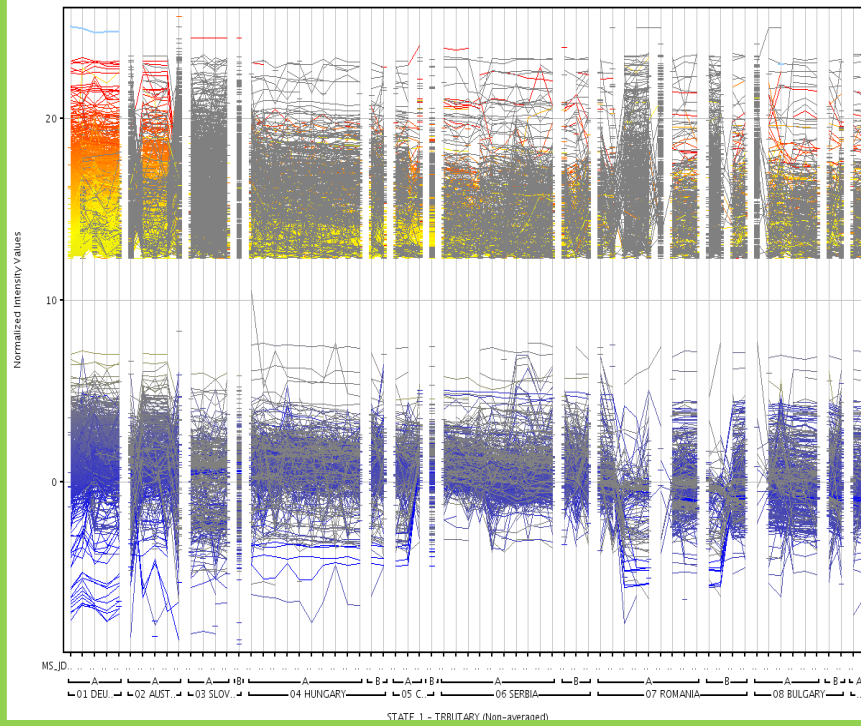
- autoMSMS confirmed 30 pest compoundsS

Joint Danube Survey 3



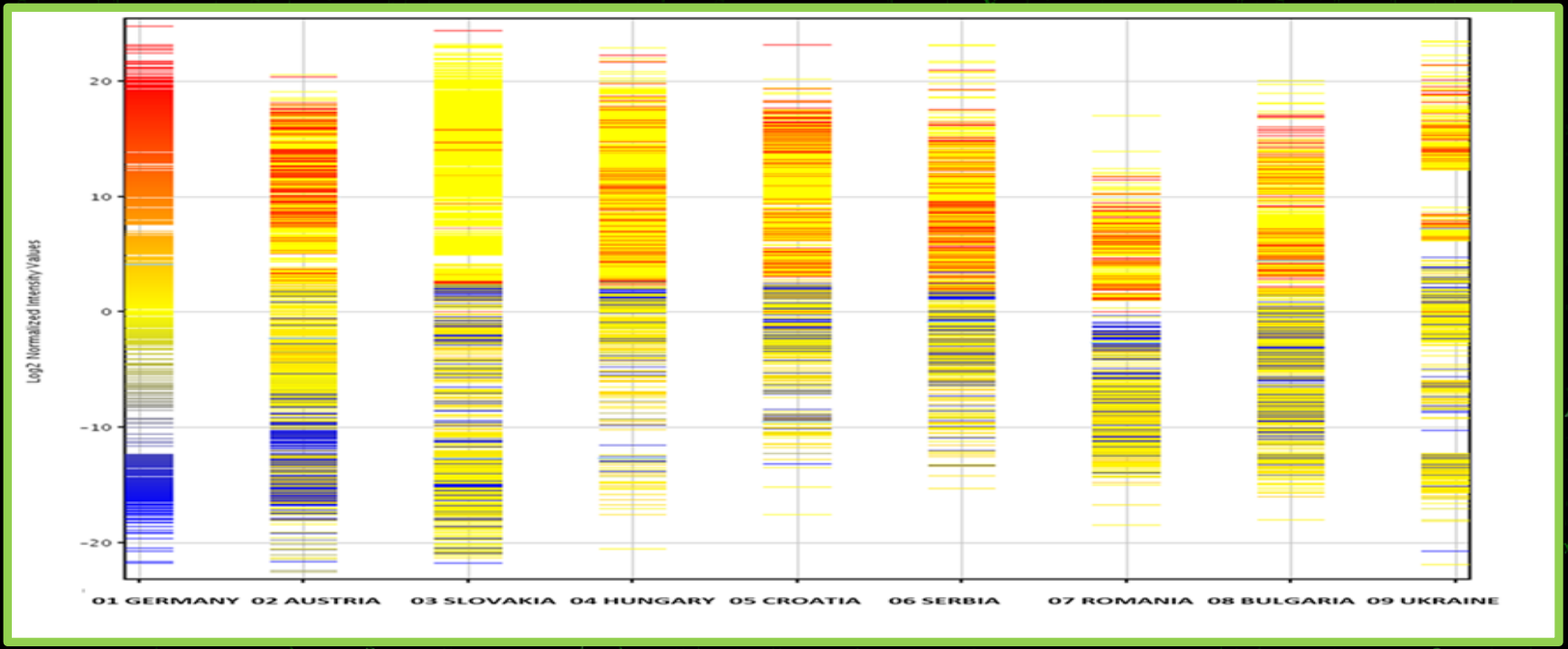


Summary display parameter state



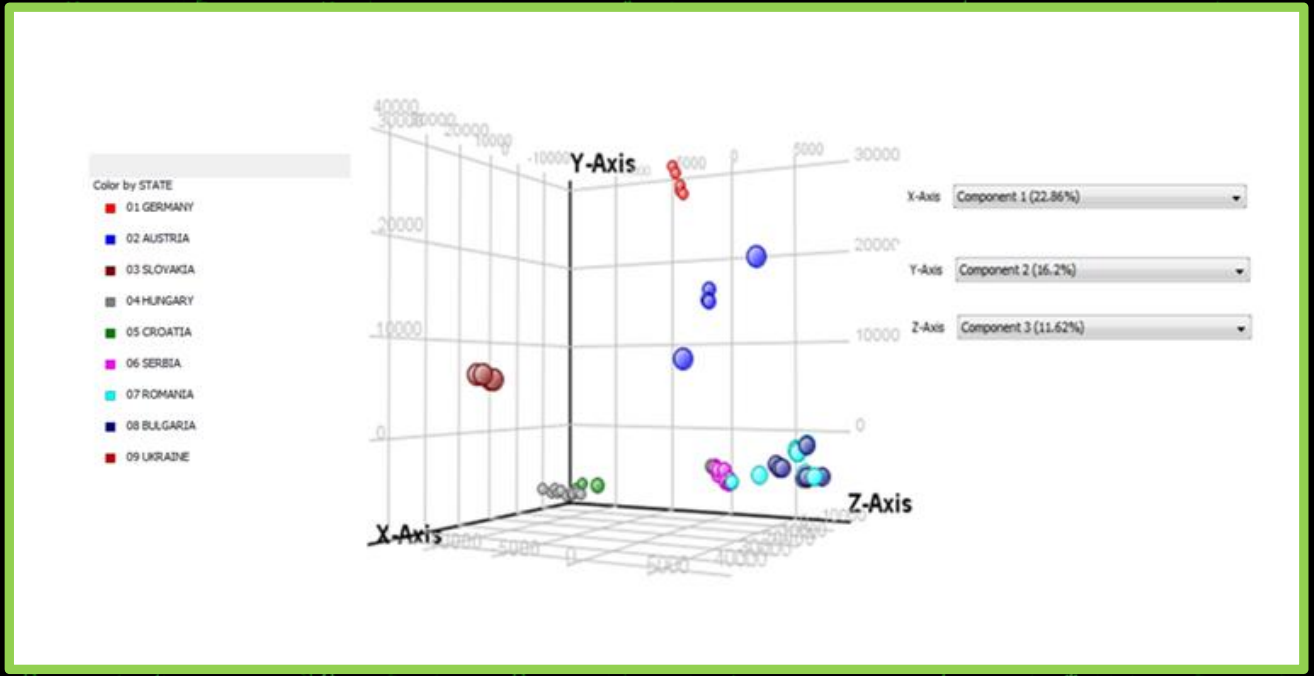
Summary display parameter state and tributary

JDS3 - STATES



Initial quality control on acquired 16214 features in MPP with filtering by frequency, sample variability, flags, abundance, significance testing and fold change resulted in 7767 features that were detected in 68 JDS3 samples. All targeted compounds were excluded from analysis.

PCA was performed for detection of similarity between states discriminated by the major trends



Similarities in pollution pattern exist among Serbia, Romania, Bulgaria and Ukraine and between Croatia and Hungary whereas rather unique character of pollution can be seen in the upstream countries (Germany, Austria, Slovakia).

OCCURRENCE OF ALL FEATURES (PCDL match compounds, unknowns, total unknowns)

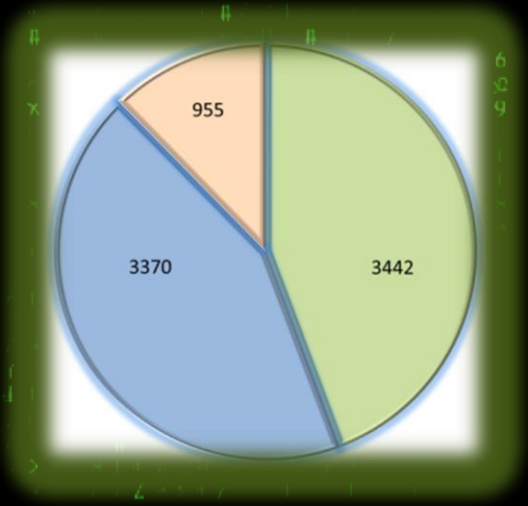
MPP

TOTAL UNKNOWN
defined only with mass and retention time

Level 5

Level 4

UNKNOWN
Calculated formula, accurate mass, Rt, isotopic pattern



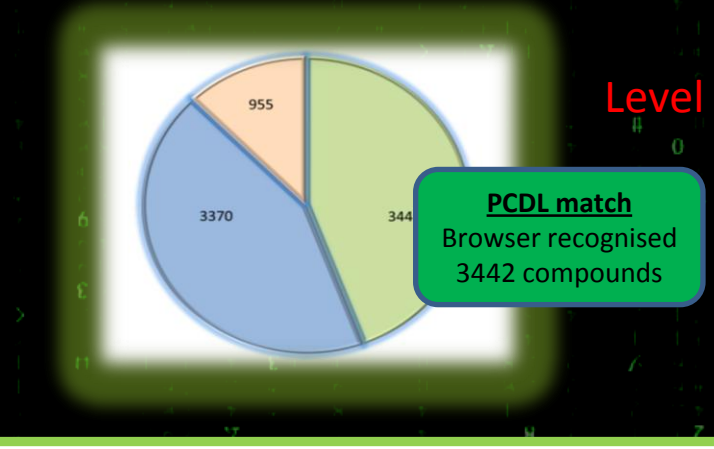
Level 3

PCDL MATCH
IDBrowser recognised (PCDL library) match compounds assigned with a defined name, accurate mass, molecular formula, Rt, CAS and isotopic pattern.

ID browser identification - compound identification wizard, database search performed by using molecular formula, mas or mass and time, mass match tolerance is specified, for formula generation allowed elements and min and max numbe are specified, formula generation can be performed on all compounds or only unidentified

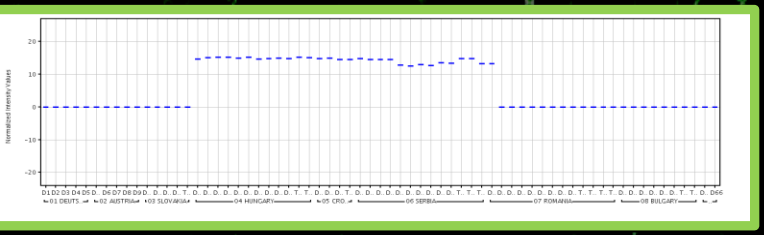
Agilent PCDL:

- **MassHunter *METLIN metabolite* PCDL** ver 5 (database 64092 compounds, MS/MS library > 8040 compounds at three collision energies: 10, 20 and 40eV)
- **MassHunter *Forensic Toxicology* PCDL** ver. 4.1 (database 7509 compounds, MS/MS library > 2500 compounds at three collision energies: 10, 20 and 40eV)
- **MassHunter *Pesticide* PCDL** ver. 4.1 (database 1664 compounds, MS/MS library > 600 compounds at three collision energies: 10, 20 and 40eV)

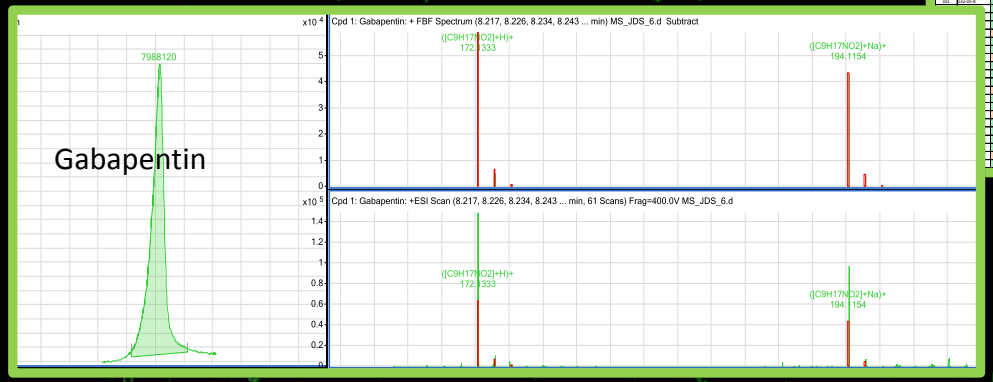


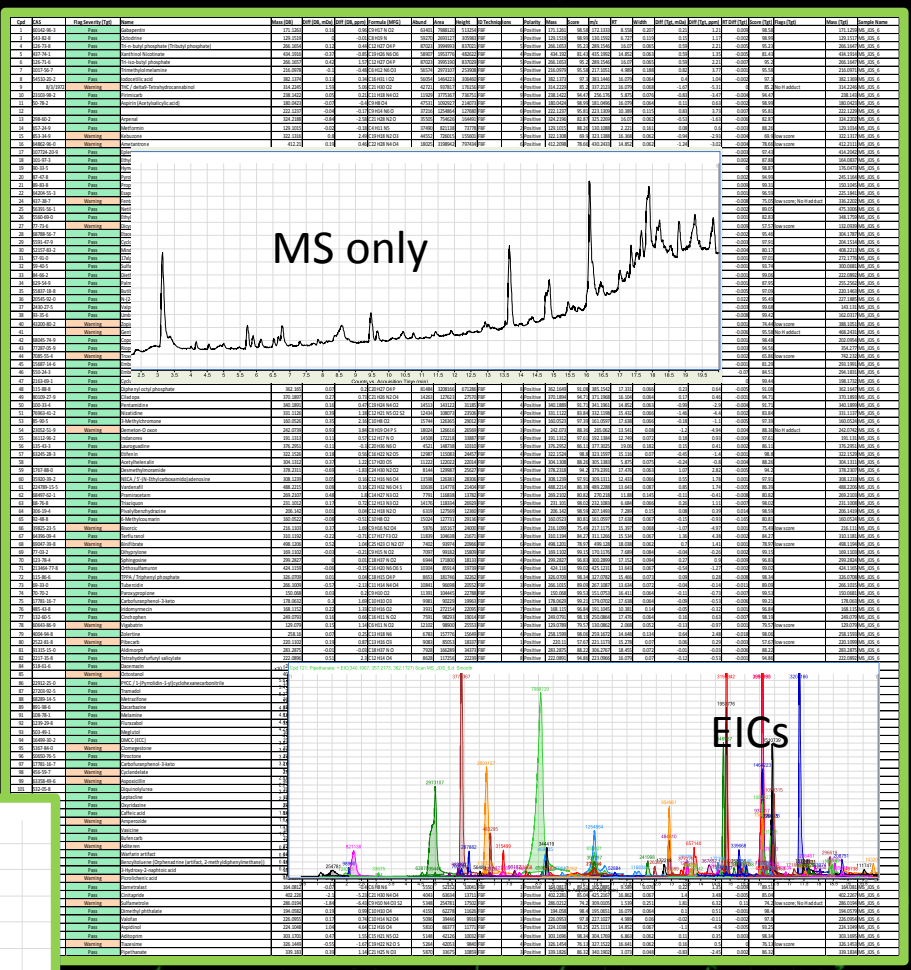
Level 3

PCDL match
Browser recognised
3442 compounds

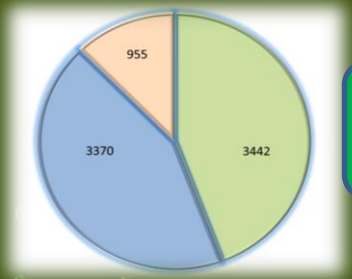


Example: Gabapentin





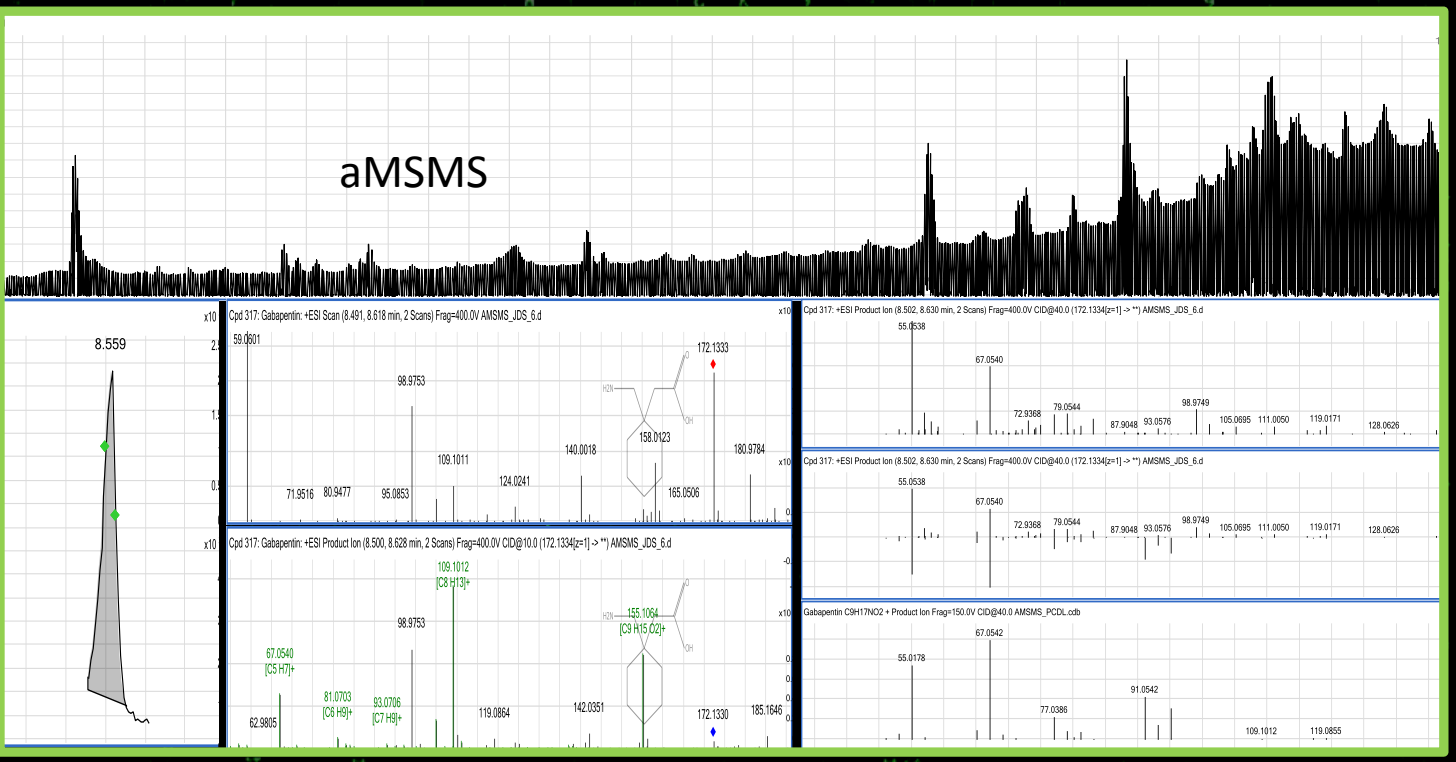
CONFIRMED COMPOUNDS



PCDL match
Browser recognised

AutoMSMS method was applied (MFE, accurate mass library search 10, 20 and 40 CE)

Example: Gabapentin



EIC FRAGMENTATION PCDL comparison

IDENTIFICATION CONFIDENCE

Level 2

Probable structure

MS, MS_2, library MS_2

CONCLUSIONS

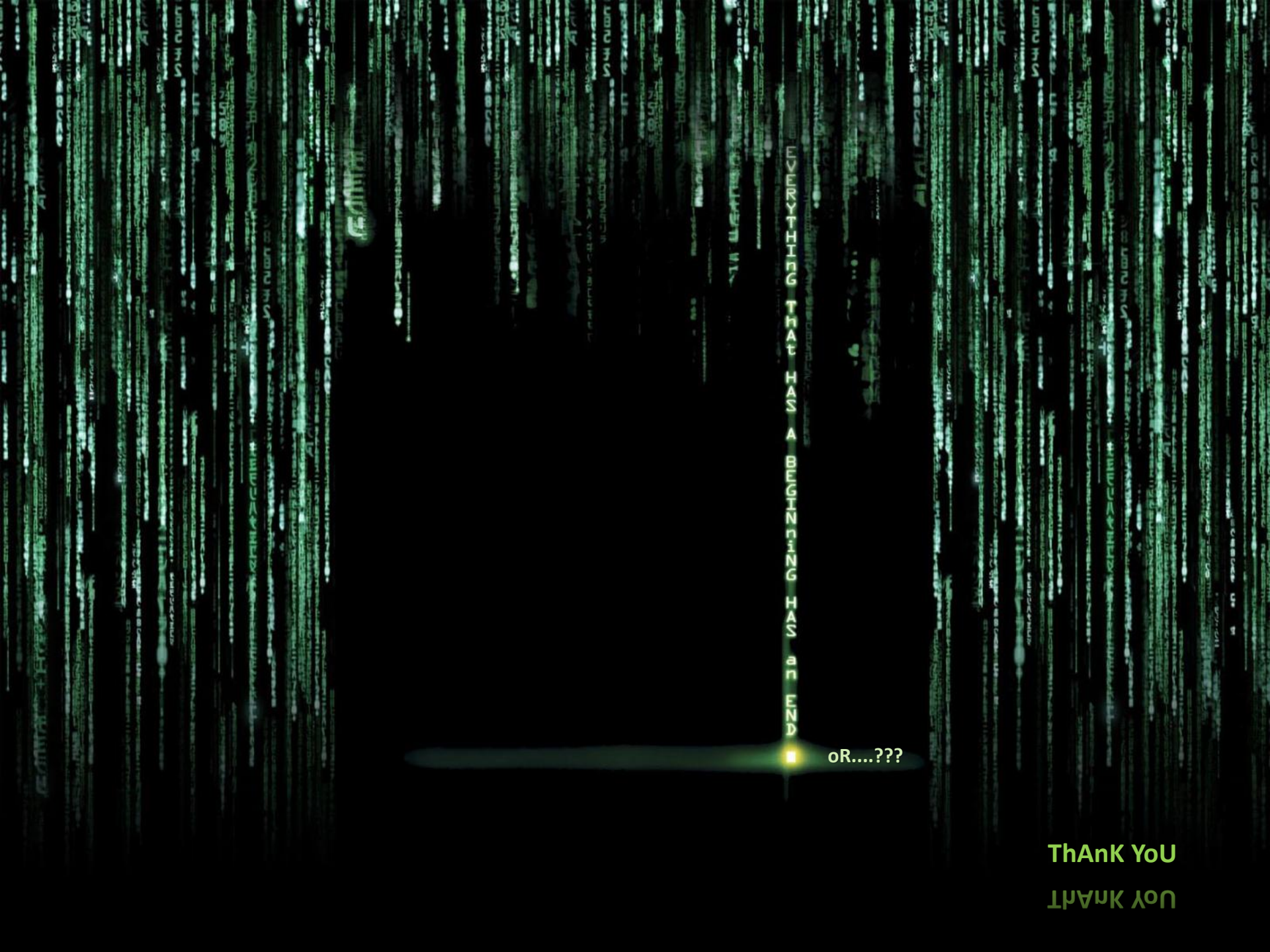
MS systems generate vast amounts of data and therefore there is a need for strategy to reduce the amount of detected (thousands of) substances in a single sample to ‘workable’ numbers (top 10 – 100 substances).

Combination of high resolution technique with different algorithms and the availability of comprehensive mass spectral libraries with accurate mass fragmentation information was shown to be important at the detected compounds identification.

International databases equipped with various structure elucidation tools, such as NORMAN MassBank (Schulze et al., 2012, NORMAN Association, 2014), would be of great benefit for identification of present and future emerging substances.

The statistical software (user friendly) and LC-QTOF-MS allowed clear differentiation in pollution patterns for the river stretches and countries within the basin.

Unfortunately, no one can be told what the (Matrix) MPP is, you have to see it for yourself.


EVERYTHING THAT HAS A BEGINNING HAS an END
oR....???
ThAnK YoU