Development of a LC-HRMS workflow for the target, suspect and non-target screening of contaminants of emerging concern in environmental water samples

Nikolaos S. Thomaidis

Laboratory of Analytical Chemistry
Department of Chemistry
University of Athens

Emerging pollutants (EPs)

Pharmaceuticals
Illicit drugs
Personal care products
Endocrine disruptive compounds (EDCs)
Flame retardants
Food additives
Disinfection by-products
Pesticides
+
Metabolites &
Transformation Products **(TPs)**

aquatic environment

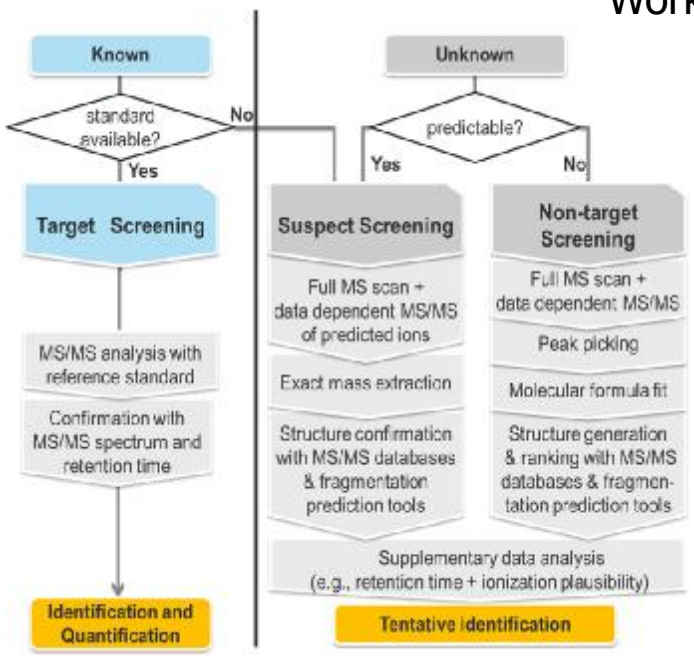
Identification Approaches

Target screening
- Known EPs
- Analytical standards available

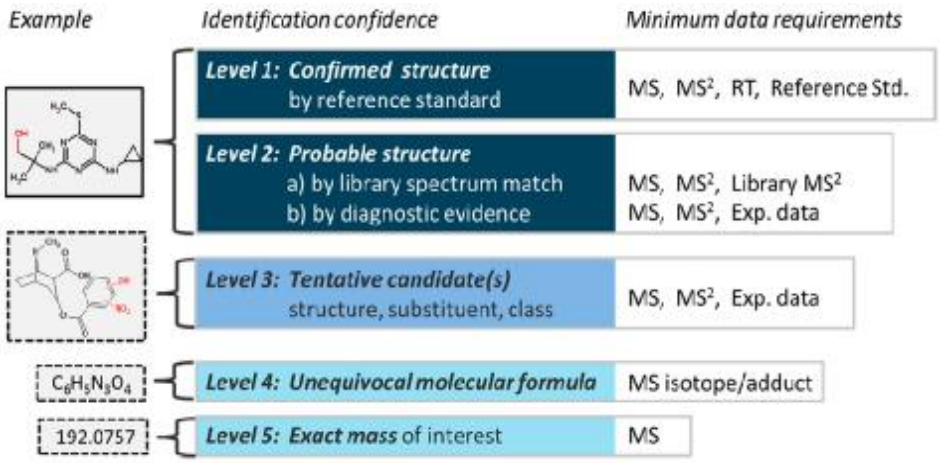
Suspect screening
- List of possible EPs and their TPs (literature & prediction models)

Non-target screening
- Unknown compounds
- post-acquisition data tools

Workflows



IDENTIFICATION LEVELS



Schymanski et al. Environmental Science and Technology (2014) 48(4):2097

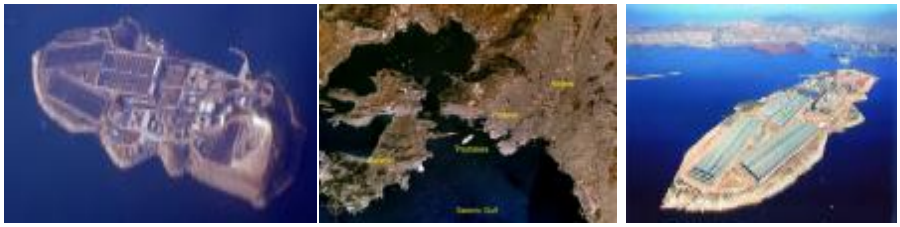
Sampling

Location: WWTP of Athens, Greece

Period: March 2014

Samples:

- *24-h composite* flow-proportional samples of *influent* wastewaters & *effluent* wastewaters over a week (7 consecutive days)
- *2-h composite* flow-proportional samples of *influent* wastewaters (Thursday & Saturday, 12 samples per day, from 02:00 to 00:00)



Sample Preparation - Analysis

- 200 mL filtered wastewater (pH adjusted to 6.5)
- Isotopically labelled internal standards were added (100 ng/L)
- Mixed SPE with 4 sorbents:
  (Strata X copolymer, Strata-X-AW, Strata-X-CW, IsoluteENV+)
- Extraction: Neutral, Basic & Acidic Compounds
- Evaporation/reconstitution to a final volume of 200 µL

HPLC-HRMS
-QTOF-MS/MS

MS & MS/MS data in a single run

Target & Suspect Analysis:
bbCID

Non-target screening:
AutoMS/MS

I. Target Screening

in-house database:

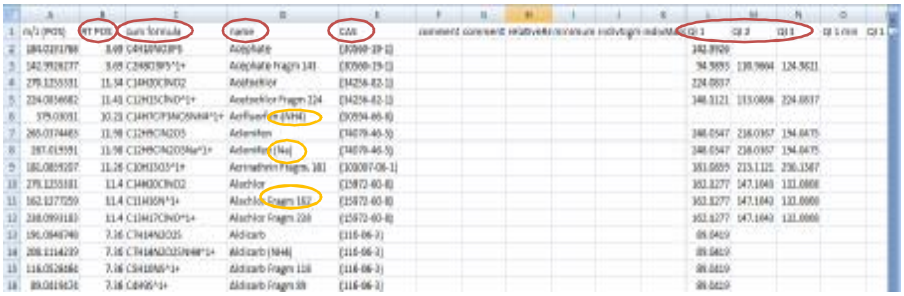
1500 compounds for positive ESI screening

500 compounds for negative ESI target screening

~200 common compounds

- *more* than 700 pesticides
- *more* than 800 EPs & TPs

…including information over:

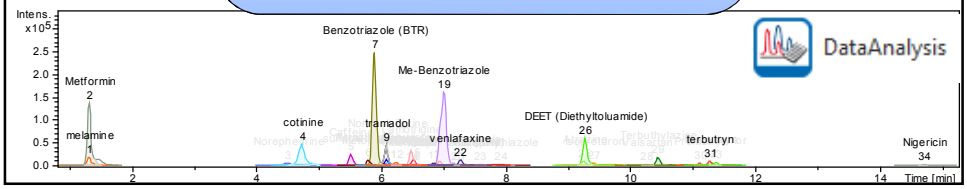


I. Target Screening

TargetAnalysis

Criteria

- deltaRT ≤ 0.05 min
- Accuracy: Error ≤ 5 ppm
- Isotopic fit: ≤ 20 mSigma
- MS/MS fragments, ion ratio
- Ion Intensity > 500 (+ESI) / 200 (-ESI)
- Area > 2000 (+ESI) / 800 (-ESI)



DataAnalysis

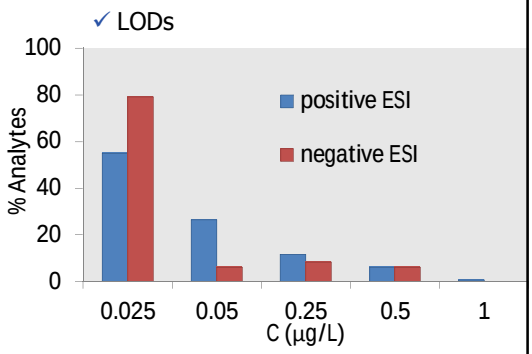
Validation

I. Target Screening

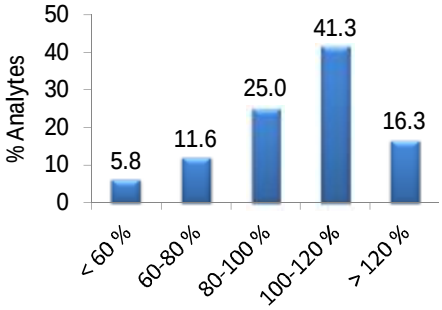
200 target compounds
- 170 positive ESI
- 50 negative ESI

Over the whole range of database

✓ LODs



✓ % Recoveries



✓ Linearity in stds, spiked samples & matrix-matched samples

R^2> 0.92- 0.9999

✓ Repeatability: %RSD <20%

83 % of analytes

Results

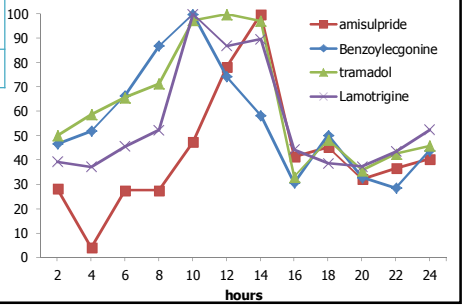
I. Target Screening

Sat. 15/03/14
24-h composite wastewater

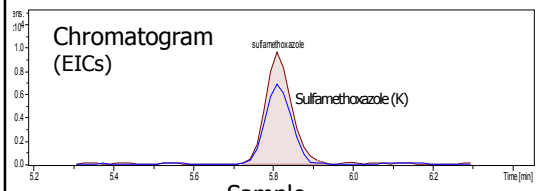
effluent		influent
123	**Compounds detected**	176
75	pharmaceuticals & drugs of abuse	103
23	pesticides	39
6	PFCs	6
4	sweeteners	4
10	Disinfection by-products & PCP	19
5	Aminoacids	5

4.0 **ng/L** (Primidone) –
26.1 µg/L (Caffeine)
0.5 **mg/L** (Metformin)

Sat. 15/03/14
2-h influent wastewater



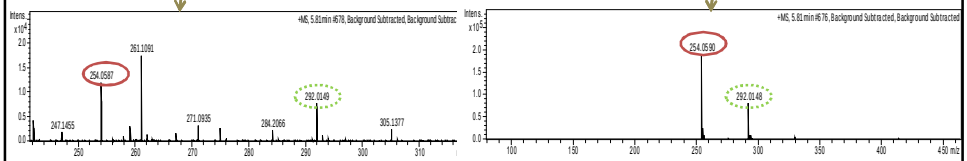
I. Target Screening



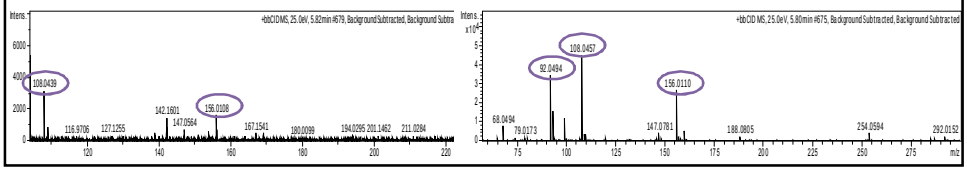
Sample

Reference Standard

MS



MS/MS spectra



II. Suspect Screening

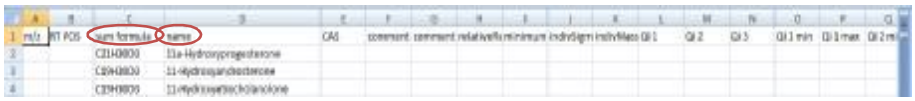
1. in-house database

- *more* than 10000 EPs and TPs → from *prediction models* (UM-PPS, Metabolite Predict)

from *literature*

from *regulation bodies* (REACH)

…including information over:



2. Retention time prediction tool
KNN-GA-SVM

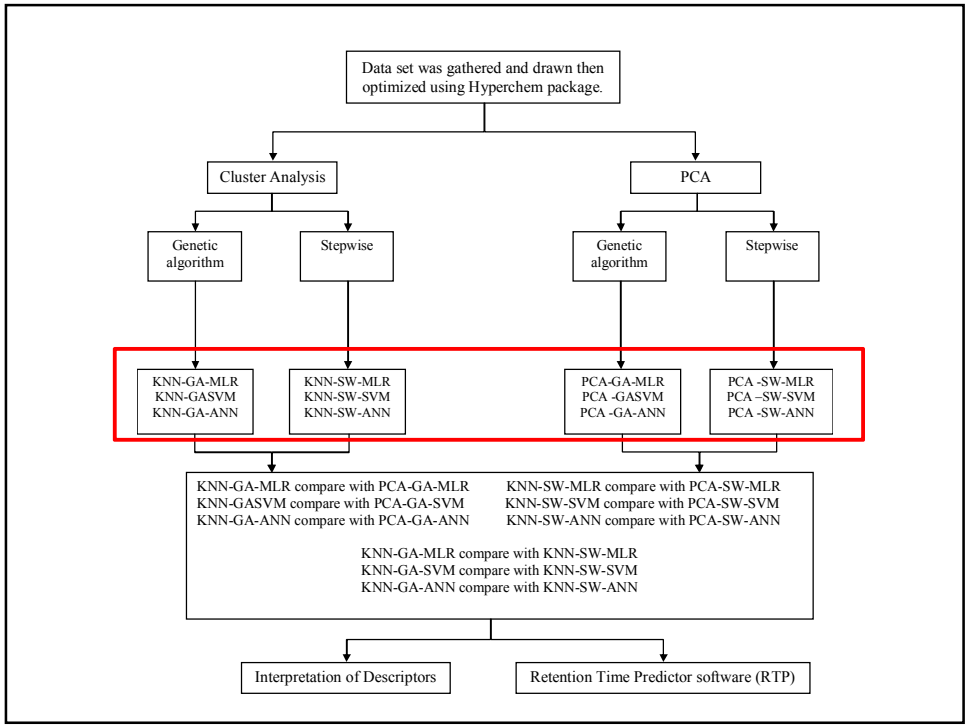
3. High Resolution Mass Spectral Libraries

for MS/MS data (MassBank, MetFrag)

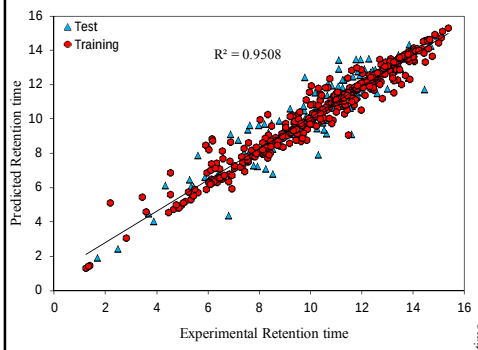
Retention Time Prediction Models

QSAR/QSPR procedure:

1. Optimization by HyperChem / MOPAC
2. Molecular descriptors by Dragon (zero, constant and near-constant, and collinear descriptors were removed)
3. Division of dataset to training and test datasets by clustering (KNN) or PCA
4. Selection of the relevant descriptors by Stepwise or Genetic algorithm
5. Build of models by MLR, ANNs, and SVM and their comparison

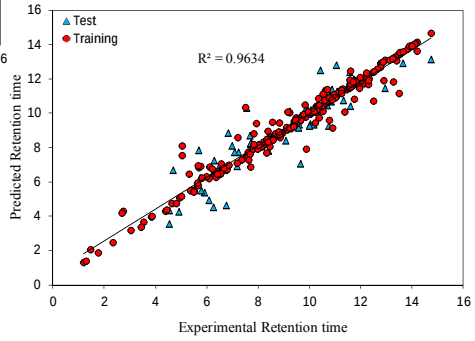


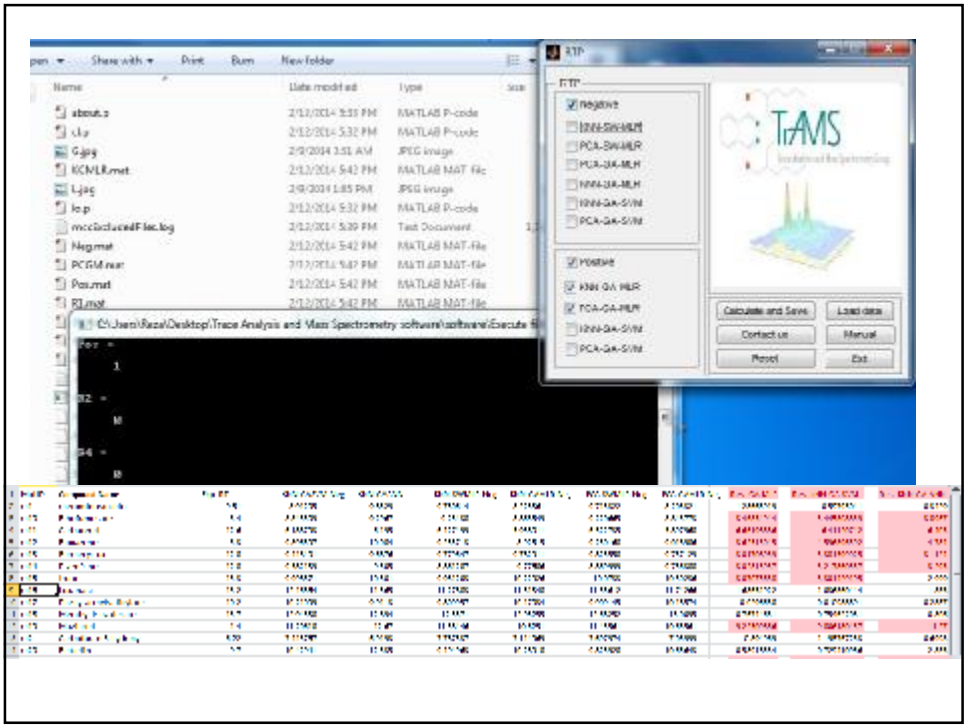
The best prediction accuracy was achieved by **KNN-GA-SVM model** for both positive and negative ESI compounds.



KNN-GA-SVM plot – (+) ESI Compounds

KNN-GA-SVM – (-) ESI compounds





Optimization & Validation

II. Suspect Screening

TargetAnalysis

Criteria

- ✓ Peak Area/Intensity Ratio > 4
- ✓ False Negative Results <10%

…in order exclude too many irrelevant peaks !

Thresholds

+ ESI: Peak Area: 20,000, Intensity: 5,000

- ESI: Peak Area: 2,000, Intensity: 500

Application to "artificial" suspect

at 5 concentration levels (1 – 0.025 µg/L)

% False Negative

C (µg/L)	+ESI	-ESI
1	9.8	0
0.5	11	6.2
0.25	19	27
0.05	61	40
0.025	88	50

II. Suspect Screening

Database of **10,000** compounds

Ion Intensity > 20,000 (+ESI) / 2,000 (-ESI)
Area > 5,000 (+ESI) / 500 (-ESI)

effluent

3590 hits (+ESI) / **1493** hits (-ESI)

influent

4438 hits (+ESI) / **3245** hits (-ESI)

Isotopic fit: ≤ 100 mSigma

2612 hits (+ESI) / **943** hits (-ESI)

3325 hits (+ESI) / **2727** hits (-ESI)

Accuracy: Error ≤ 5 ppm

2061 hits (+ESI) / **703** hits (-ESI)

2997 hits (+ESI) / **2249** hits (-ESI)

Area/Intensity : 7-25 *

* Validation mean ±2 SD

610 hits (+ESI) / **150** hits (-ESI)

663 hits (+ESI) / **564** hits (-ESI)

Retention time prediction model & (+)ESI or (-)ESI amenable

...in progress

MS/MS data in on-line spectral libraries

II. Suspect Screening

focus on..

METABOLITES & TPs

Database of **1,500** human metabolites of consumed pharmaceuticals

- 2 generations
  - Phase I
  - Phase II
  - Cytochrome P450

influent

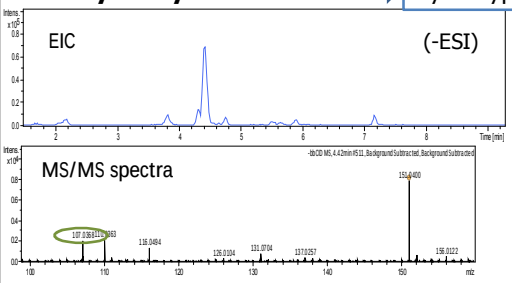
274 hits (+ESI)

254 hits (-ESI)

Retention time prediction
Reject ~30% suspect compounds

Examples

Acetylsalicylic acid Met12 → Hydroxyphenyl acetate



t_R = 4.43 min Predicted t_R = 4.39 min
Accuracy: 1.3 ppm
Isotopic fit: 0.6 mSigma

Suspect Screening – Human Metabolites

Theophylline Met8

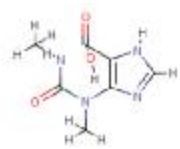
- ESI

t_R = 3.0 min

Predicted t_R = 3.9 min

Accuracy: 0.3 ppm

Isotopic fit: 10 mSigma



+ ESI

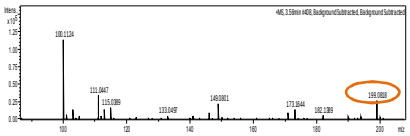
t_R = 3.6 min

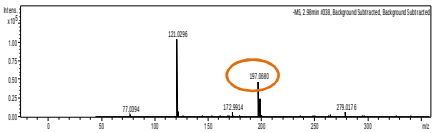
Predicted t_R = 3.5 min

Accuracy: 3.2 ppm

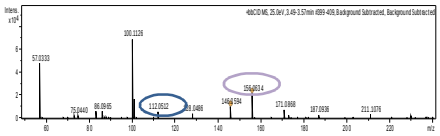
Isotopic fit: 10.7 mSigma

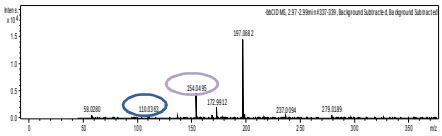
MS spectra





MS/MS spectra





III. Non-target Screening

WHY NON-TARGET?

TARGET SCREENING	✓ Known substance ✓ Reference standard available	✓ Unequivocal identification ✓ Possible quantification
SUSPECT SCREENING	✓ Suspect substance ✓ No reference standard available	✓ Qualitative detection possible

What proportion of substances present in the samples are actually detected with target and suspect screening?

III. Non-target Screening

- ✓ Usually, many of the most intense peaks do not correspond to substances included in the target and suspect screening lists.
- ✓ These substances are potentially relevant, due to their high concentration.

✓ **Identification of these substances is environmentally relevant**

NON-TARGET SCREENING

- ✓ No former information on the analytes
- ✓ Molecular structures can be assigned on the basis of the exact mass, isotopic pattern and fragmentation information

✓ *Nevertheless, full identification of unknown compounds is often difficult & there is no guarantee of a successful outcome*

III. Non-target Screening

STANDARD SCREENING WORKFLOW

Full scan (MS) and Product ion spectra (MS/MS)
Accurate mass measurements

↓

Automatic peak detection using Algorithms
(High number of peaks)

↓

Determination of the Elemental compositions of the unknowns

↓

Determination and evaluation of candidates
(Tentative) Identification of TPs
- *Interpretation of the fragmentation pathway*
- *Chromatographic retention time plausibility*

↓

Confirmation: RT and MS/MS of chemical standards, when available

- ✓ Large effort on manual data evaluation
- ✓ Systematic strategies with automated approaches are required to prioritize relevant peaks on which the identification efforts should focus

PROPOSED APPROACH

III. Non-target Screening

✓ Analyses are carried out in the same way previously described for target and suspect screening, except that AutoMSMS is performed (MS/MS data of the 5 most intense peaks per scan event).

Non-target steps:

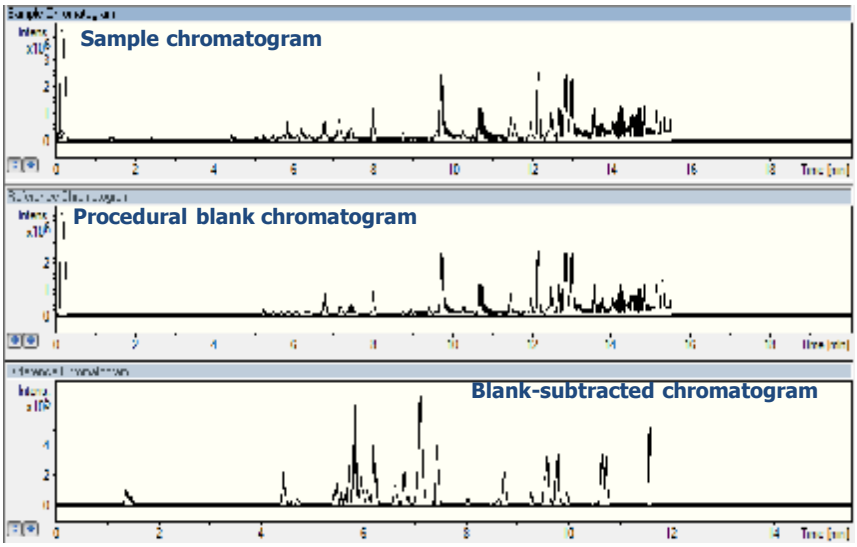
- ✓ Blank subtraction
- ✓ Peak peaking procedure
- ✓ Prioritization of peaks for further evaluation
- ✓ Determination of elemental composition
- ✓ Evaluation of possible candidates → Tentative identification

BLANK SUBTRACTION

III. Non-target Screening

✓Use of metabolomics tools

Metabolite Detect



III. Non-target Screening

PEAK PEAKING PROCEDURE

✓ Peak peaking: **Molecular features Algorithm**

- Using *Data analysis* and *Target analysis (Bruker)*
- Threshold: Signal/Noise >10

➡ **A high number of peaks (> 3500) was obtained**

#	RT [min]	Area	Int. Type	I	S/N	Max. m/z
1	1,1	12025,0	MolFeature	1108	12,1	151,0084
2	1,1	[illegible]	MolFeature	[illegible]	[illegible]	[illegible]
3	1,1	30852,8	MolFeature	2927	11,1	181,038
4	1,1	150714,1	MolFeature	14085	76,1	232,9988
5	1,1	[illegible]	MolFeature	[illegible]	[illegible]	[illegible]
6	1,1	74232,5	MolFeature	7411	44,9	166,9858
7	1,1	13137,0	MolFeature	1538	17	261,9721
8	1,1	[illegible]	MolFeature	[illegible]	[illegible]	[illegible]
9	1,1	9764,1	MolFeature	472	10,5	634,8444
10	1,1	[illegible]	MolFeature	[illegible]	[illegible]	[illegible]
11	1,1	9017,2	MolFeature	927	10,3	406,9612
12	1,1	8880,0	MolFeature	804	17,9	444,9128
13	1,1	[illegible]	MolFeature	[illegible]	[illegible]	[illegible]
14	1,1	73705,8	MolFeature	5180	57,6	378,9341
15	1,1	[illegible]	MolFeature	528	14	334,9604
16	1,1	[illegible]	MolFeature	[illegible]	[illegible]	[illegible]
17	1,1	51524,7	MolFeature	4804	32,5	181,0011
18	1,1	[illegible]	MolFeature	[illegible]	[illegible]	[illegible]
19	1,1	9466,9	MolFeature	1016	11,3	318,8721
20	1,1	142217,7	MolFeature	10180	62,1	133,034
21	1,1	[illegible]	MolFeature	[illegible]	[illegible]	[illegible]
22	1,1	109310,9	MolFeature	16292	38,8	216,9499
23	1,1	17145,6	MolFeature	1157	15,3	412,9716
24	1,1	[illegible]	MolFeature	[illegible]	[illegible]	[illegible]

III. Non-target Screening

PRIORITIZATION OF PEAKS FOR FURTHER EVALUATION

✓ Selection of the **most relevant** from the large peak list (Not included either in the target or the suspect screening)

Criteria:

- Intensity
- Presence of a distinctive isotopic pattern

Non-target identification was performed on selected masses from the top most intense peaks

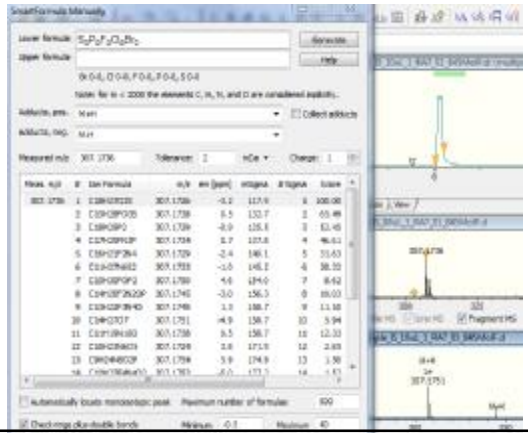
III. Non-target Screening

DETERMINATION OF ELEMENTAL COMPOSITION

1st step: **Generation of possible molecular formula(s)**

Criteria:

- Mass accuracy → threshold: 5 ppm
- Agreement of the theoretical and measured isotopic pattern



DETERMINATION OF ELEMENTAL COMPOSITION: SEVEN GOLDEN RULES (SGR)

✓**Plausibility of the generated molecules** → Use of the **Seven Golden Rules** software

"Seven golden rules for heuristic filtering of molecular formulas obtained by accurate mass spectrometry"

i. Element number restrictions
ii. Lewis and Senior chemical rules check
iii. Isotopic pattern filter
iv. Hydrogen/carbon ratio check
v. Element ratio of nitrogen, oxygen, phosphorus and sulphur vs carbon check
vi. Element ratio probability check
vii. Check of the presence of trimethylsilylated compounds

30 million compounds database → Great reduction of the possibilities

✓ The correct molecular formula is assigned with a probability of 98%, if the formula exists in a compound database

Kind and Fiehn. BMC Bioinformatics 8:105 (2007)

EVALUATION OF POSSIBLE CANDIDATES

✓ Number of candidates to one molecular formula: **1** - **>2000**
(*Chemspider, Pubmed databases*)

Approaches for tentative identification:

✓**Databases** (e.g. MassBank) →*Still very limited number of compounds (not very useful for non-target screening)*
✓Deep **study** of the **MS/MS spectra (AutoMSMS analysis)**
✓ ***In-silico* fragmentation software**
- Smart formula 3D *(Bruker)*
- Metfrag

✓ **Chromatographic retention time** plausibility → Application of models
✓ **Number of data sources and references** in different data bases (e.g. Chemspider)

✓ To confirm the identity of a substance, purchase of reference standard is required (if available)

EXAMPLE 1: TREATING METFORMIN AS UNKNOWN

PEAK PEAKING PRIORIZATION

✓ Peak peaking: **Molecular features Algorithm**
- Threshold: Signal/Noise >10

➡ **A large amount of peaks (> 3500) obtained**

#	RT [min]	Area	I	S/N	Chromatogram	Max. m/z
31	1.38	23464522.0	2383580	26.3	Metformin, 130.1087±0.002, C 4H 11N 5, 1.4min	130.1091
53	1.91	39184028.0	2186079	508.8		145.0977

Metformin

#	RT [min]	Area	I	S/N	Chromatogram	Max. m/z
31	1.38	23464522.0	2383580	26.3	Metformin, 130.1087±0.002, C 4H 11N 5, 1.4min	130.1091

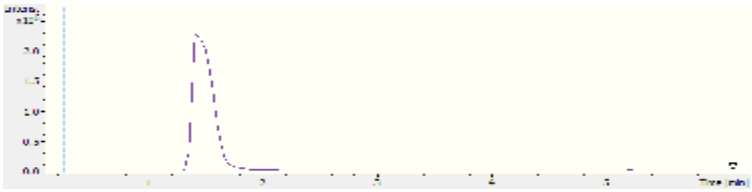
Compound detected using the TARGET ANALYSIS approach

324	6.44	18125204.0	1262588	379.6		608.3899
322	6.13	17853796.0	1262524	262.7		520.3356

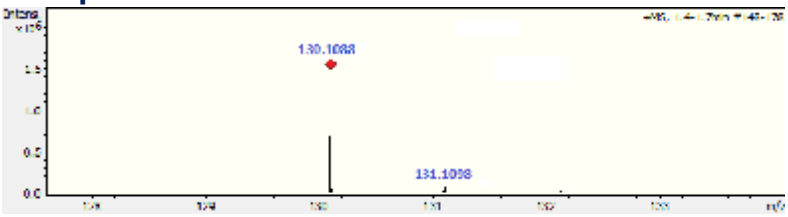
Non-target identification was performed on 16 selected masses from the top most intense peaks

EXAMPLE 1: TREATING METFORMIN AS UNKNOWN

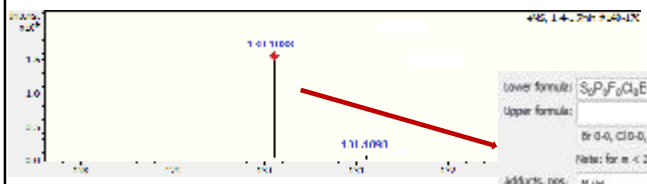
- **Experimental accurate mass: 130.1088**
- **Retention time: 1.4 min**

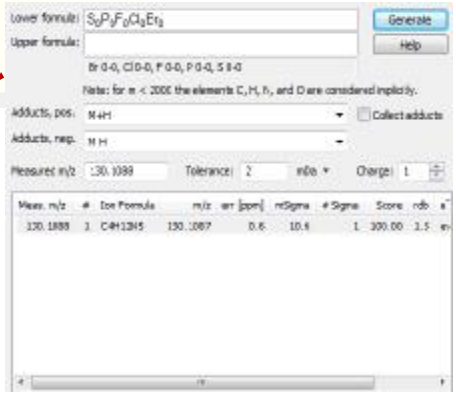


MS spectra



EXAMPLE 1: ELEMENTAL COMPOSITION DETERMINATION





Number of possible formulas → 1
(Threshold of 5 ppm and 50 mSigma)

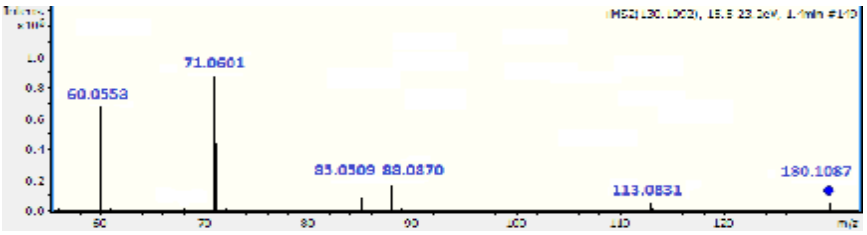
Seven Golden Rules

1 Plausible Molecular formula

$C_4H_{12}N_5$

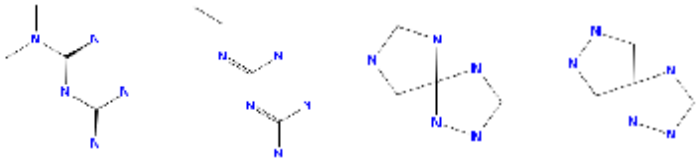
EXAMPLE 1: TENTATIVE IDENTIFICATION

MS/MS Spectra

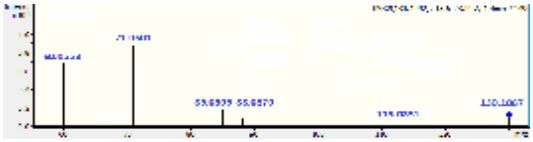


✓$C_{19}H_{28}N_4$

- ✓ Hits Chemspider: 12
- ✓ **Compounds with score > 0.8 → 4**

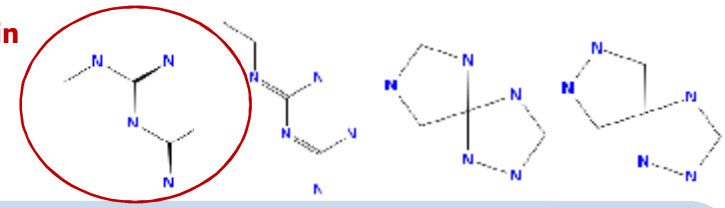


EXAMPLE 1: TENTATIVE IDENTIFICATION



✓$C_{19}H_{28}N_4$

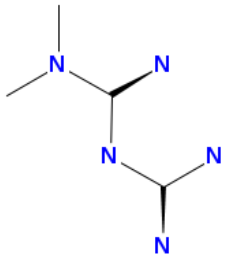
Metformin



✓ Metfrag Score:	1	0.86	0.81	0.81
✓ Explained Fragments:	5	3	2	2
✓ RT Pred. Model:	✓	✓	x	x
✓ Chemspider data sources:	59	5	4	4
✓ Chemspider References:	293	6	4	4

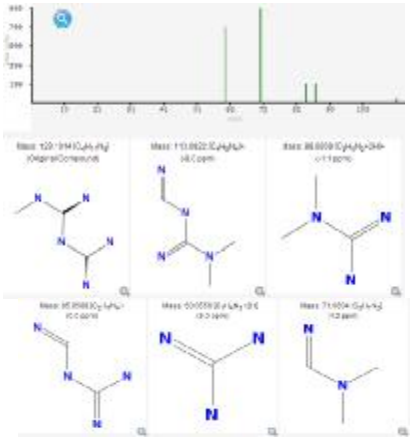
EXAMPLE 1: TENTATIVE IDENTIFICATION

Metformin



APPLICATION OF RETENTION TIME PREDICTION MODEL:
Experimental RT = 1.38 min
Predicted RT = 2.5 min

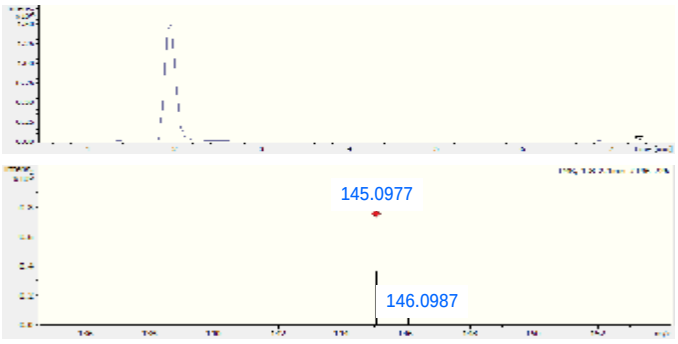
Metfrag peak explanation



- The developed workflow was applied successfully to identify unambiguously this compound as Metformin

EXAMPLE 2: APPLICATION OF THE WORKFLOW TO A REAL UNKNOWN

- **Experimental accurate mass: 145.0977**
- **Retention time: 1.9 min**



Number of possible formulas (Threshold of 5 ppm, 50 mSigma) and Seven Golden Rules

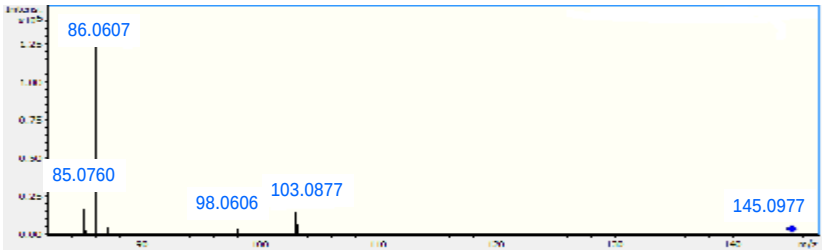
→

1 Plausible Molecular formula

$C_6H_{12}N_2O_2$

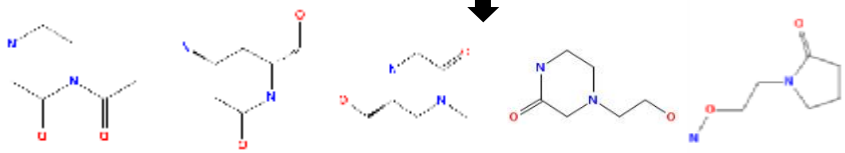
EXAMPLE 2: TENTATIVE IDENTIFICATION

MS/MS Spectra



✓ $C_6H_{12}N_2O_2$

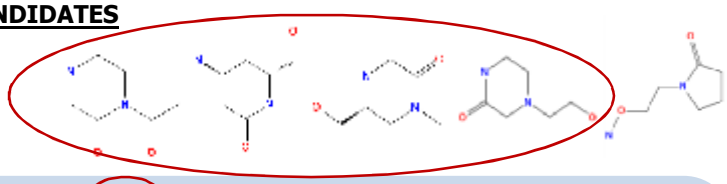
- ✓ Hits Chemspider: 336
- ✓ Compounds with score > 0.9 → 28
- ✓ Only few with more than 3 fragment matches



EXAMPLE 2: TENTATIVE IDENTIFICATION

✓ $C_6H_{12}N_2O_2$

TENTATIVE CANDIDATES



✓ Metfrag Score:	1	0.97	0.95	0.95	0.95
✓ Explained Fragments:	4	4	4	4	3
✓ RT Pred. Model:	✓	✓	✓	✓	x
✓ Chemspider data sources:	1	4	5	17	2
✓ Chemspider References:	1	4	5	18	2

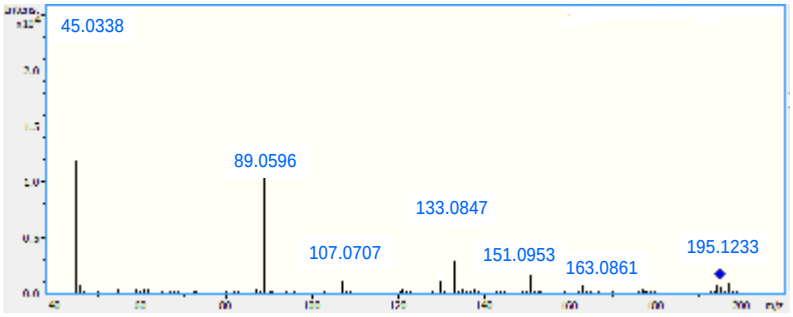
EXAMPLE 3: TENTATIVE IDENTIFICATION

- **Experimental accurate mass: 195.1233**
- **Retention time: 4.2 min**

Number of possible formulas
(Threshold of 5 ppm, 50 mSigma)
and Seven Golden Rules

➡

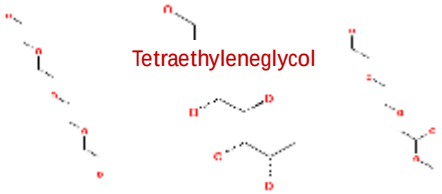
1 Plausible Molecular formula
$C_8H_{18}O_5$



EXAMPLE 3: TENTATIVE IDENTIFICATION

- ✓ Hits Chemspider: 13
- ✓ 3 compounds with Metfrag score > 0.95 and the others below 0.5

✓ $C_8H_{18}O_5$



✓ Metfrag Score:	1	1	0.95
✓ Explained Fragments:	5	5	5
✓ RT Pred. Model:	✓	✓	✓
✓ Chemspider data sources:	1	67	2
✓ Chemspider References:	1	379	2

SUMMARY OF THE LEVELS OF IDENTIFICATION

Retention time (min)	Mass of ion [m/z] (peak of component)	Ion type	Intensity	Molecular formula	Proposed identification name	Level of confirmation of identification
1.28	164.1282	$[M+H]^+$	1508655	C7H17NO3		Unequivocal molecular formula
1.91	145.0977	$[M+H]^+$	2186079	C6H12N2O2	e.g. 4-(2-Hydroxyethyl)-2-piperazinone	Tentative candidates
2.27	96.0452	$[M+H]^+$	1145713	C5H5NO	2-Formyl-1H-pyrrole	Probable structure
4.19						
4.68						formula
4.98						
5.09						formula
5.16						formula
5.2						
5.24						formula
5.73						formula
6.13						
6.44						
9.1					hydroxyethyl)octanamide	
9.4	191.1647	$[M+H]^+$	1410087	C10H22O3		Unequivocal molecular formula
12.69	316.1955	$[M+H]^+$	1137576	C16H29NO3S	e.g. 1-{(2-Methoxyethyl)[(5-methyl-2-thienyl)methyl] amino}-3-[(2-methyl-2-propanyl)oxy]-2-propanol	Tentative candidates

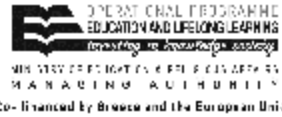
- ✓ **16 evaluated top intense peaks in +ESI mode**
  - ✓ 5 Tentatively candidates
  - ✓ 7 Unequivocal molecular formula
  - ✓ 4 Exact mass of interest

Conclusions

- Target and suspect HRMS screening workflows were developed and validated
- Target screening can identify app. 10% of the obtained peaks from a LC-QTOFMS analysis
- Suspect screening can explained app. 20% of the obtained peaks
- Non-target workflows are needed for the tentative identification of the highly abundant peaks

Acknowledgments

Pablo Gago Ferrero
Anna Bletsou
Katerina Psoma
Reza Aalizadeh



This research has been co-financed by the European Union and Greek national funds through the Operational Program "Education and Lifelong Learning" of the National Strategic Reference Framework (NSRF)

Thank you for your attention!

National and Kapodistrian
UNIVERSITY OF ATHENS
Faculty of Chemistry

http://trams.chem.uoa.gr/