A data analysis pipeline integrating ion mobility and high-resolution mass spectrometry for non-target screening in environmental studies

Julien Sade, Sabrina Guérin, Stéphane Mottelet, Vincent Rocher, Régis Moilleron and Julien Le Roux

Laboratoire Eau Environnement et Systèmes Urbains (LEESU)
Univ Paris Est Créteil, Créteil, France

julien.sade@u-pec.fr

SETAC Europe 2024, 5-9 May, Seville, Spain

What are the various HRMS strategies for characterising contaminants?

Technique	Targeted	Suspect	Non-target
Question	Are compounds x, y, & z present in this sample?	Which compounds of a defined list are present in this sample?	Which compounds are present in this sample?
Type	Known-knowns	Known-unknowns	Known-unknowns & unknown-unknowns

Confidence levels in contaminants annotation

Technique	Targeted	Suspect	Non-target
Question	Are compounds x, y, & z present in this sample?	Which compounds of a defined list are present in this sample?	Which compounds are present in this sample?
Type	Known-knowns	Known-unknowns	Known-unknowns & unknown-unknowns

Level	1		
data requirements	MS, MS/MS, RT, Reference standard		

Schymanski and al., (2014), Environmental Science & Technology, Identifying Small Molecules via High Resolution Mass Spectrometry: Communicating Confidence.

Confidence levels in contaminants annotation

Technique	Targeted	Suspect	Non-target
Question	Are compounds x, y, & z present in this sample?	Which compounds of a defined list are present in this sample?	Which compounds are present in this sample?
Type	Known-knowns	Known-unknowns	Known-unknowns & unknown-unknowns
Level	1	2 & 3	4 & 5
data requirements	MS, MS/MS, RT, Reference standard	MS, MS/MS, Library MS/MS	MS, isotope/adduct

Schymanski and al., (2014), Environmental Science & Technology, Identifying Small Molecules via High Resolution Mass Spectrometry: Communicating Confidence.

Confidence levels in contaminants annotation

Technique	Targeted	Suspect	Non-target
Question	Are compounds x, y, & z present in this sample?	Which compounds of a defined list are present in this sample?	Which compounds are present in this sample?
Type	Known-knowns	Known-unknowns	Known-unknowns & unknown-unknowns
Level	1	2 & 3	4 & 5
data requirements	MS, MS/MS, RT, Reference standard	MS, MS/MS, Library MS/MS	MS, isotope/adduct

Potential false positives

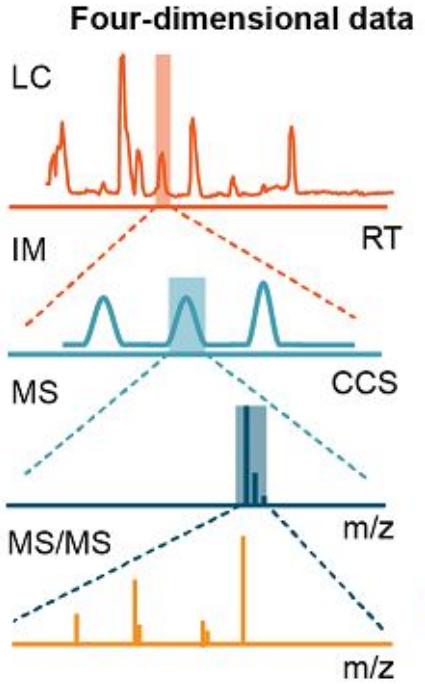
Schymanski and al., (2014), Environmental Science & Technology, Identifying Small Molecules via High Resolution Mass Spectrometry: Communicating Confidence.

Confidence levels in contaminants annotation: impact of CCS data

Vion® IMS QTof

Waters

THE SCIENCE OF WHAT'S POSSIBLE.™

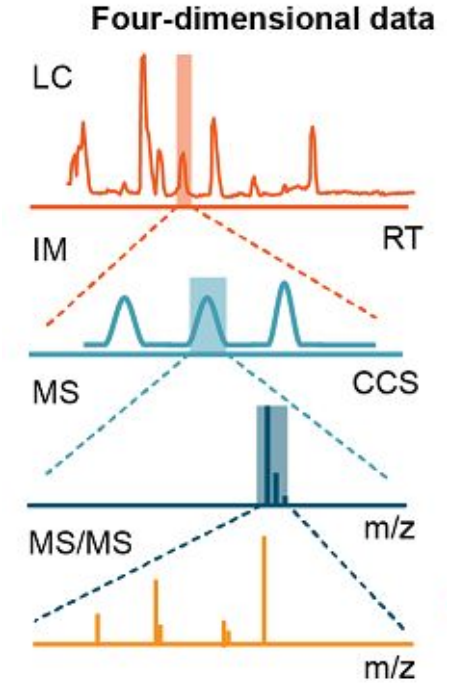


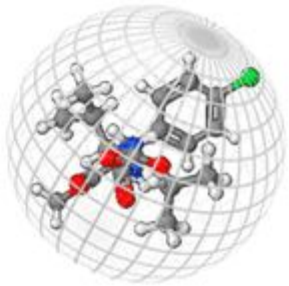
Zhou and al., (2019), Bioinformatics, LipidIMMS Analyzer: integrating multi-dimensional information to support lipid identification in ion mobility-mass spectrometry based lipidomics,

Confidence levels in contaminants annotation: impact of CCS data

Vion® IMS QTof

Waters
THE SCIENCE OF WHAT'S POSSIBLE.™





Why IMS?

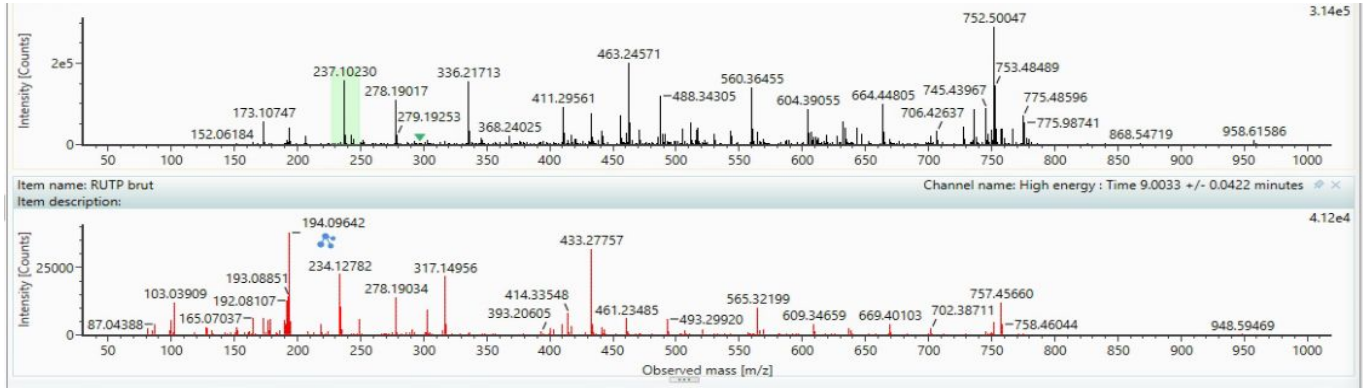
- CCS : an additional identification parameter
- Spectral cleaning
- Reduction of false positive rate

Celma and al., (2020), Environ Sci Technol. Improving Target and Suspect Screening High-Resolution Mass Spectrometry Workflows in Environmental Analysis by Ion Mobility Separation.

Confidence levels in contaminants annotation: impact of CCS data

Low energy

High energy

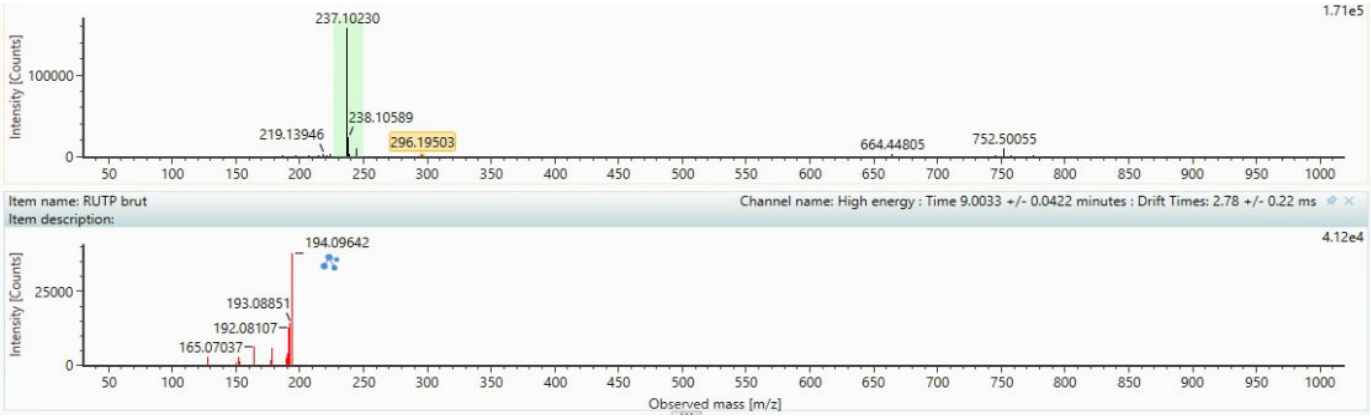


NO IMS

IMS filter

Low energy

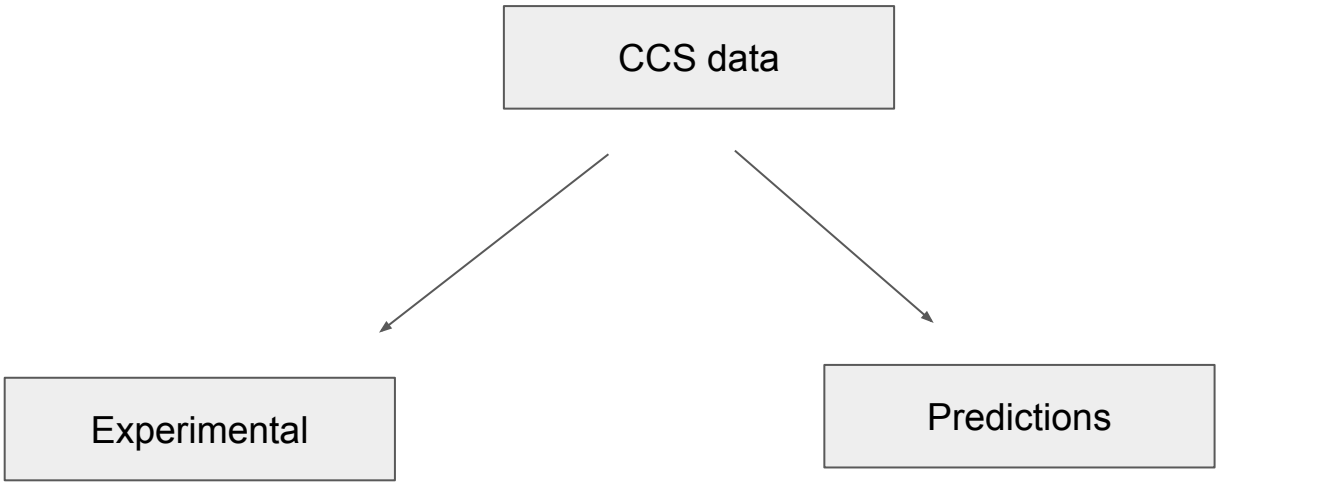
High energy



IMS

Better spectral matching with MS libraries!

CCS data for NTS



- Standards required
- Expensive
- Accurate

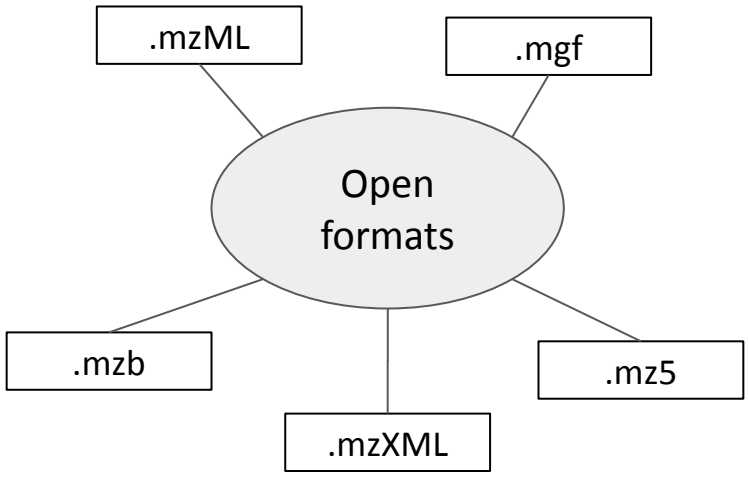
- Standards not required
- Cheaper
- Accurate ?

AllCCS

CCSbase

Zhou and al., (2020), Nat Commun, Ion mobility collision cross-section atlas for known and unknown metabolite annotation in untargeted metabolomics.

Challenges with formats and softwares including ion mobility (IMS)



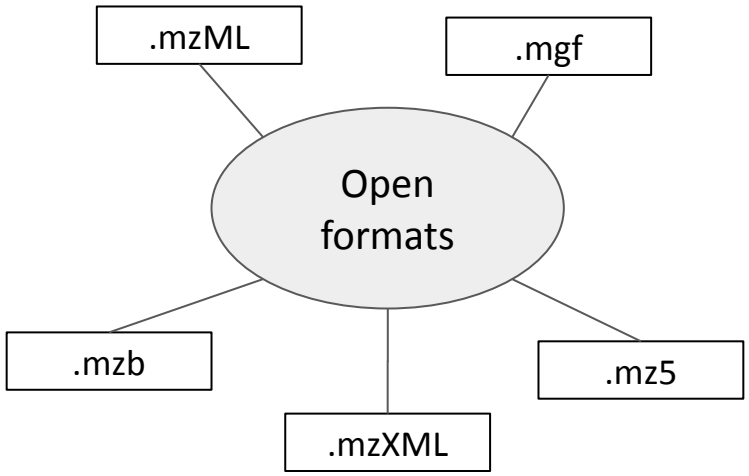
Issues with proprietary formats: lack of interoperability, closed ecosystems, unknown algorithms…

Data conversion takes a long time
~ 20 minutes for 1 replicate

File size ~ 2-10 GB per file

Most tools do not read or process open-format data including IMS (yet)
Mzmine, MS-dial, XCMS, patRoon…

Challenges with formats and softwares including ion mobility (IMS)



Issues with proprietary formats: lack of interoperability, closed ecosystems, unknown algorithms…

Data conversion takes a long time
~ 20 minutes for 1 replicate

File size ~ 2-10 GB per file

Most tools do not read or process open-format data including IMS (yet)
Mzmine, MS-dial, XCMS, patRoon…

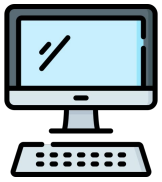
Our solution: arcMS

- R package
- Free, open-source
- Including IMS
- Under our control
- parquet file format

Parquet

arcMS workflow

A UNIFI to .parquet conversion package



asynchronous parallel requests

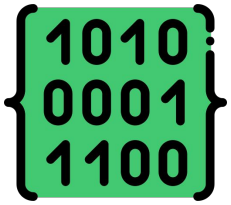
UNIFI SCIENTIFIC INFORMATION SYSTEM

Waters THE SCIENCE OF WHAT'S POSSIBLE.

Vion IMS QTof

collect MS data
locally or network

binary response



deserialization

	scanId	retention_time	intensity	drift_time	mz
0	1	0.007410	0.0	0.284	556.259277
1	1	0.007410	4.0	0.284	556.262268
2	1	0.007410	0.0	0.284	556.265259
3	1	0.007410	0.0	0.426	556.250244
4	1	0.007410	2.0	0.426	556.253235

columnar format

compression and saving

Parquet

compact

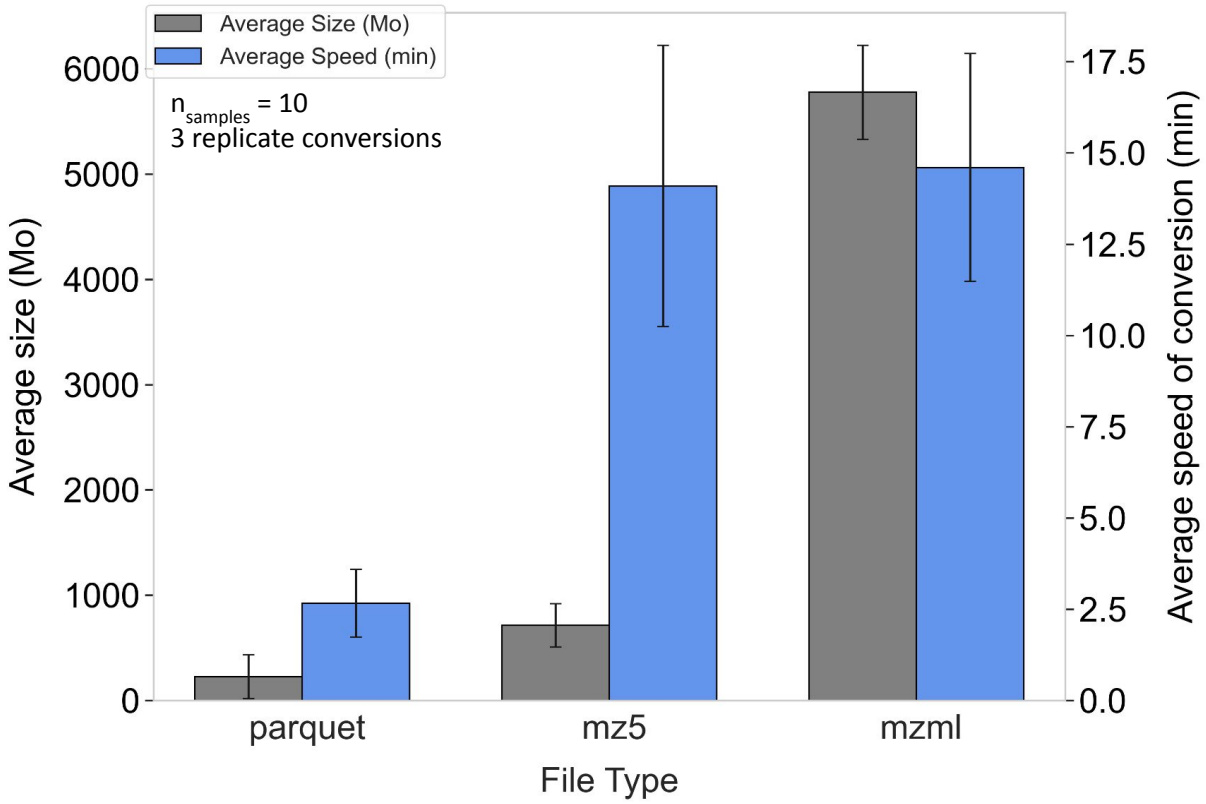
arcMS performance

Data collection and conversion speed

arcMS vs proteowizard

File size

arcMS vs mzML and mz5



- More compact
- Faster collection

arcMS philosophy

Opening and visualization of raw data

Columnar format

	scanId	retention_time	intensity	drift_time	mz
0	1	0.007410	0.0	0.284	556.259277
1	1	0.007410	4.0	0.284	556.262268
2	1	0.007410	0.0	0.284	556.265259
3	1	0.007410	0.0	0.426	556.250244
4	1	0.007410	2.0	0.426	556.253235

- Open source
- Modular, extensible, easy to manipulate, fast to read
- Compatible with several programming languages (R, python, Java, C++, C#,etc…)
- Compatible with the most commonly used libraries
  - For visualization : plotly, matplotlib, ggplot…
  - For data analysis : pandas, data.table, scikit-learn, numpy…

arcMS philosophy

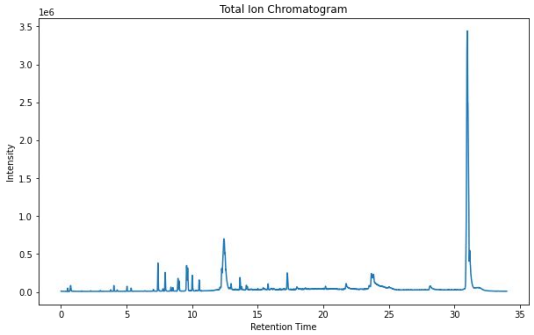
Opening and visualization of raw data

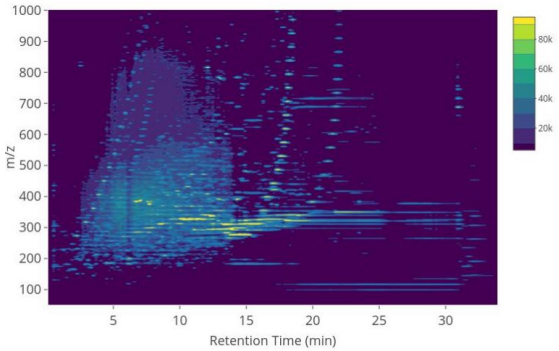
Columnar format

	scanId	retention_time	intensity	drift_time	mz
0	1	0.007410	0.0	0.284	556.259277
1	1	0.007410	4.0	0.284	556.262268
2	1	0.007410	0.0	0.284	556.265259
3	1	0.007410	0.0	0.426	556.250244
4	1	0.007410	2.0	0.426	556.253235

Parquet

matplotlib ggplot2 plotly bokeh





- Open source
- Modular, extensible, easy to manipulate, fast to read
- Compatible with several programming languages (R, python, Java, C++, C#,etc…)
- Compatible with the most commonly used libraries
  - For visualization : plotly, matplotlib, ggplot…
  - For data analysis : pandas, data.table, scikit-learn, numpy…

\+ DuckDB for on-disk operations! (low RAM usage)

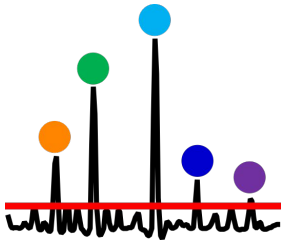
arcMS integration with DEIMoS: a package to process LC-IMS-MS data

DEIMoS

arcMS data →

Feature detection

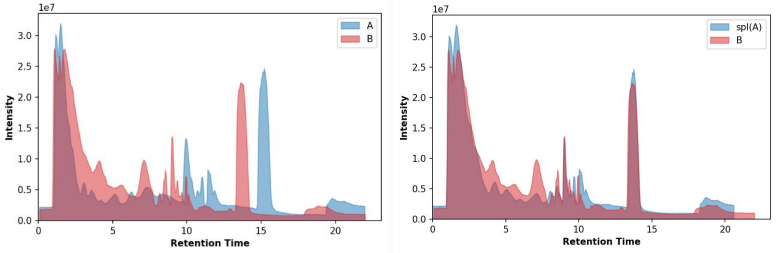
- Organize data into N-dim array
- Detection is based on 3D dimensions, including IMS



↓

Alignment

- Reference-based alignment (regression models)
- Cross-sample alignment (hierarchical clustering)

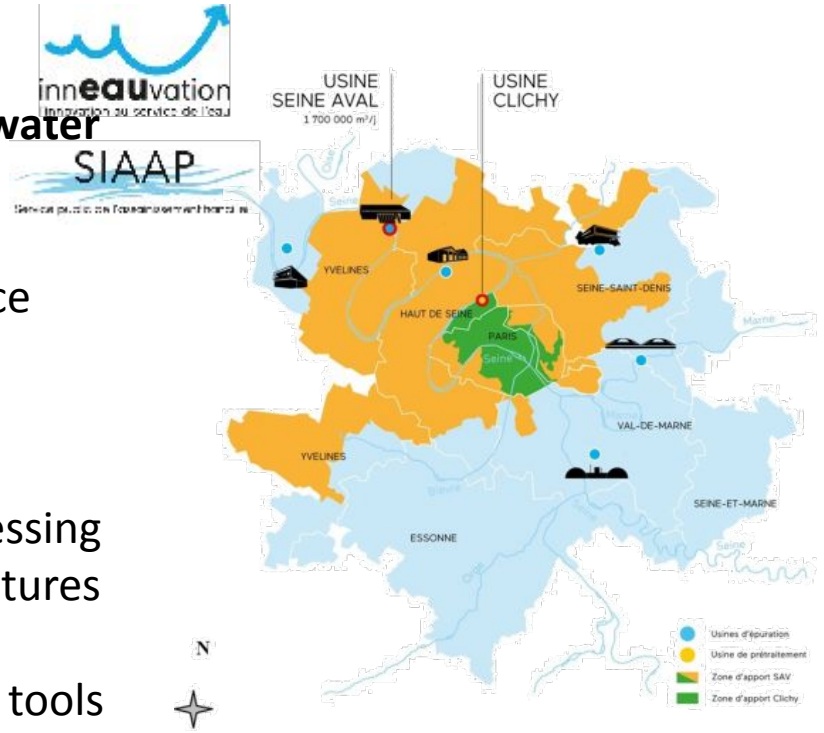


before alignment | after alignment

Colby and al., (2022), Analytical Chemistry, DEIMoS: an open-source tool for processing high-dimensional mass spectrometry data.

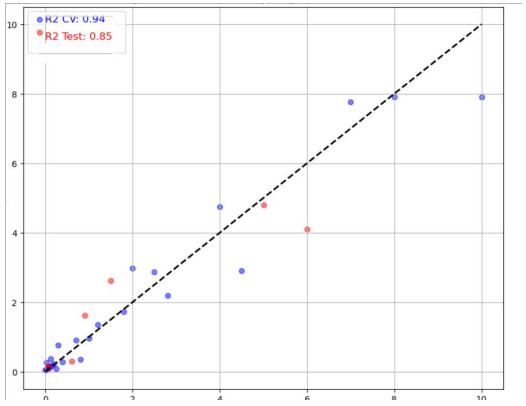
Applications of the data pipeline

- **Observatory of Paris wastewater (Greater Paris sanitation authority)**:
  - Goal: 10-year analyses
  - achieving a 15x reduction in storage space
  - Cost savings



- **Data processing and statistics:**
  - Development of an application for processing spectral data including IMS features (peak picking, alignment…)
  - Visualization tools (chromatograms, contour plot…)
  - Statistical analysis (descriptive statistics, unsupervised models (PCA), and supervised models)

- **Contamination model:**
  - Development of predictive models for contamination of surface water samples



Annotation with CCS

urban wastewater discharges

retention_time	mz	intensity	drift_time	CCS
18.215646	705.409058	717.0	0.071	63.983055
17.262668	485.280487	9355.0	0.071	66.071616
17.580382	485.277679	1835.0	0.071	66.071648
18.198939	705.409058	946.0	0.213	66.197677
17.287745	509.290314	1073.0	0.142	66.925016
...	...	...	...	...

- mz tolerance 5 ppm
- CCS tolerance 6%
- All IMS

Annotation with experimental CCS data (11K)

retention_time	mz	intensity	drift_time	CCS	Matching_Molecule
16.452719	399.251129	25720.756255	6.603	206.414514	tris(2-butoxyethyl) phosphate
4.790156	371.227448	25374.575866	5.609	181.595989	Nonoxynol-9_met115
10.366448	429.239502	24420.327584	6.603	205.519580	IRBESARTAN
4.840312	416.211487	18761.324646	6.035	191.086060	Famprofazone
4.028913	195.086945	16011.000000	3.550	139.806489	Caffeine
...	...	...	...	...	...

AllCCS

CCSbase

500 matches

Level 4

mz + CSS

Annotation with CCS

urban wastewater discharges

retention_time	mz	intensity	drift_time	CCS
18.215646	705.409058	717.0	0.071	63.983055
17.262668	485.280487	9355.0	0.071	66.071616
17.580382	485.277679	1835.0	0.071	66.071648
18.198939	705.409058	946.0	0.213	66.197677
17.287745	509.290314	1073.0	0.142	66.925016
...	...	...	...	...

- mz tolerance 5 ppm
- CCS tolerance 6%
- All IMS

Annotation with experimental CCS data (11K)

retention_time	mz	intensity	drift_time	CCS	Matching_Molecule
16.452719	399.251129	25720.756255	6.603	206.414514	tris(2-butoxyethyl) phosphate
4.790156	371.227448	25374.575866	5.609	181.595989	Nonoxynol-9_met115
10.366448	429.239502	24420.327584	6.603	205.519580	IRBESARTAN
4.840312	416.211487	18761.324646	6.035	191.086060	Famprofazone
4.028913	195.086945	16011.000000	3.550	139.806489	Caffeine
...	...	...	...	...	...

MS/MS confirmation

47 matches

Level 3

AllCCS

CCSbase

500 matches

Level 4

mz + CSS

Kyoto Encyclopedia of Genes and Genomes

MetFrag

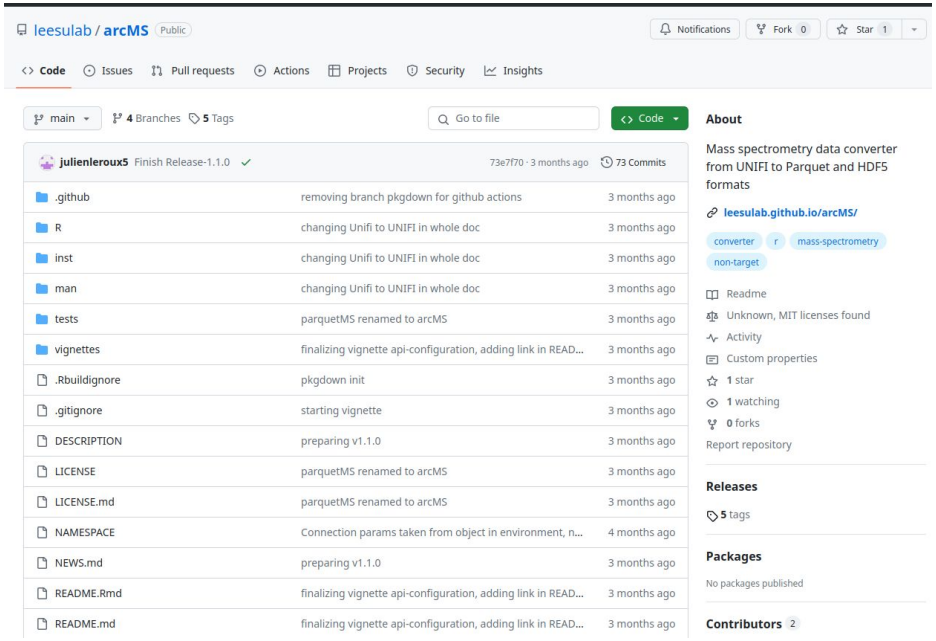
Conclusions

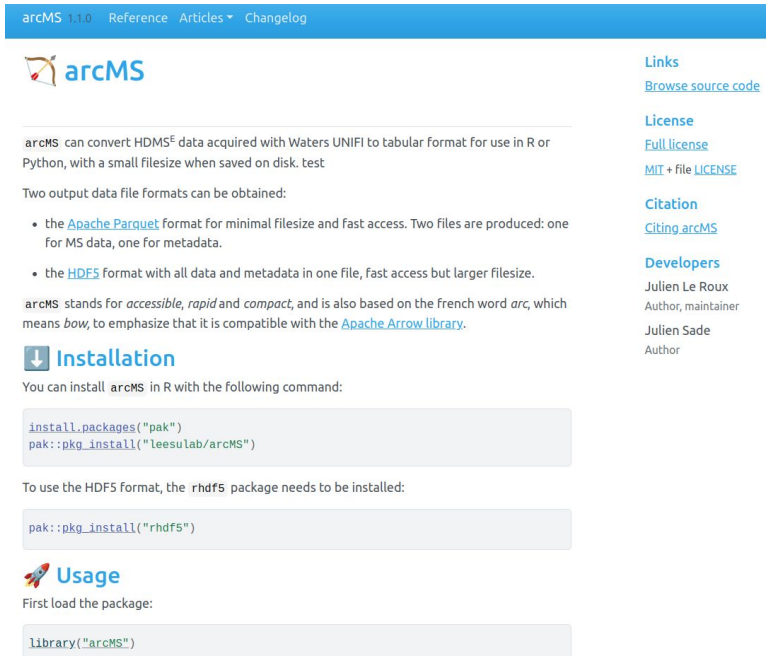
- **Fast automated pipeline for HRMS data including IMS**
- **Could be adapted to other file formats**
- **Versatility of Parquet format (space storage, reading and handling)**

UNIFI to .parquet conversion package:

https://github.com/leesulab/arcMS

https://leesulab.github.io/arcMS/





Acknowledgements

SIAAP
Service public de l'assainissement francilien

SETAC®

leesu
laboratoire eau environnement systemes urbains

PRAMMICS

utc
Université de Technologie
Compiègne