Faire de la recherche en informatique

Aurore Guillevic

Université de Lorraine, CNRS, Inria, LORIA, Nancy, France

Lycée des Récollets, Longwy, 1er et 2 décembre 2022

https://members.loria.fr/AGuillevic/files/talks/22-longwy.pdf

Parcours

`https://members.loria.fr/AGuillevic/`

- 2002–2005 : lycée à Hennebont, Morbihan
- 2005 : bac S spécialité maths

Parcours

`https://members.loria.fr/AGuillevic/`

- 2002–2005 : lycée à Hennebont, Morbihan
- 2005 : bac S spécialité maths

Études supérieures, université gratuite et bourses du CROUS :

- 2005–2008 : licence de mathématiques et informatique
  Université de Bretagne Sud, Lorient, Vannes
- 2008–2010 : master de mathématiques appliquées à la cryptographie
  Sciences des codes secrets

Parcours

Ensuite on est rémunéré, d'abord comme stagiaire puis en CDD :

- 2010 : 6 mois de stage chez Thales Communications, Colombes, Hauts-de-Seine
- 2010–2013 : thèse de doctorat en alternance (CIFRE), Thales et École Normale Supérieure, Paris
- 2014, 2015, 2016 Post-doctorat, dont 8 mois au Canada

Pacific Institute *for the* Mathematical Sciences

UNIVERSITY OF CALGARY

Aujourd'hui

1. Un peu de chimie
2. Un peu d'algorithmique

Vous pouvez poser vos questions au fur et à mesure.

Un algorithme pour détecter la pollution de l'air

Collaboration en 2018–2021 avec des chimistes et physiciens de Zurich en Suisse
Détecter de nouvelles pollutions de l'air

L'air

l'Atmosphère (air sec, sans la vapeur d'eau ) est composé de

- 78% 
- 21% 
- 0,9% 
- 0,041% 
- autres gaz à l'état de traces

L'air

l'Atmosphère (air sec, sans la vapeur d'eau ) est composé de

- 78%  diazote N_2
- 21%  dioxygène O_2
- 0,9%  argon Ar
- 0,041%  dioxyde de carbone CO_2
- autres gaz à l'état de traces

Des atomes, des molécules

Hydrogène, Carbone, Azote, Oxygène, Fluor, Phosphore, Soufre, Chlore

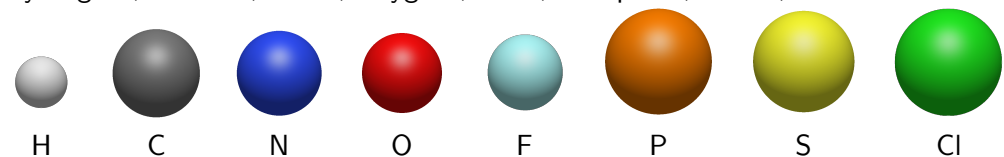
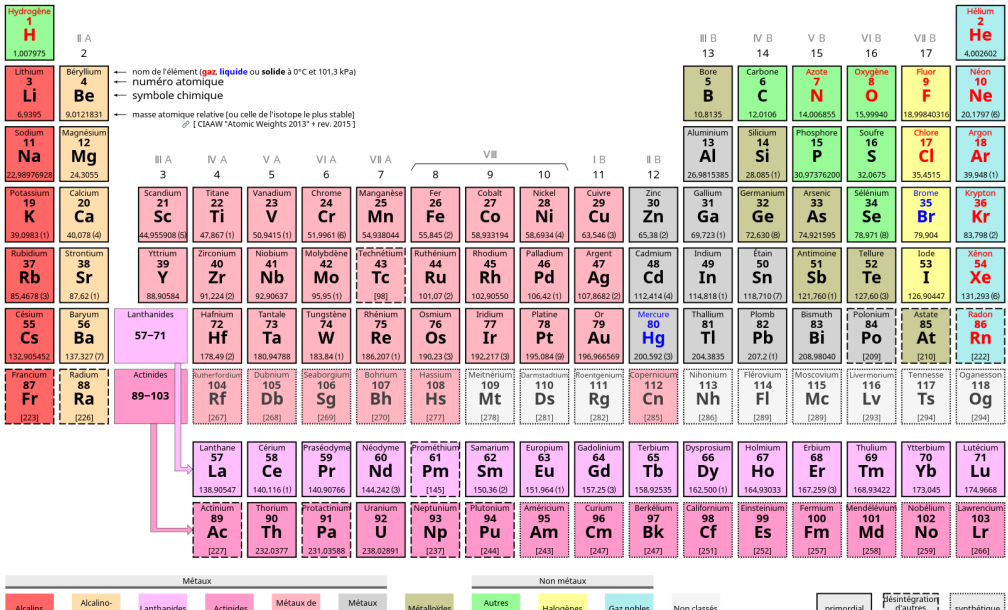


Tableau périodique des éléments chimiques



La pollution de l'air

Combustion incomplète : Monoxyde de Carbone CO
NO_x (NO, NO_2) oxydes d'azote
particules fines pm 10, pm 2.5 taille inférieure à 10 microns, 2,5 microns
abrasion des freins, usure des pneus sur la route
suies de fours et cheminées
Pollutions très dangereuses : les dioxynes (avec du Benzène)
Usines et sites "Seveso" : catastrophe industrielle en Italie, 1976

- Lubrizol à Rouen en 2019
- Beyrouth en 2020 (explosion du port)
- Leverkusen en 2021, incendie d'entrepots chimiques

Détecter la pollution

Mesure de concentrations de molécules connues

Airparif *surveillance de certaines pollutions*
pollution due au trafic routier et à l'activité humaine : Ozone O_3, dioxyne d'azote NO_2, dioxyde de soufre SO_2, particules fines

Détections de nouvelles molécules connues

Les nouvelles molécules connues (sur catalogue) sont recherchées dans l'air pour détecter leurs émissions par l'industrie
surveillance des suspects réseau international NOAA `https://www.noaa.gov/`

Recherche en aveugle de pollutions inconnues

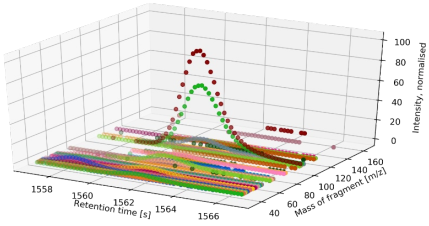
search for unknown unknowns
cas des catastrophes industrielles
Doit-on évacuer la population ?
Les pompiers peuvent-ils intervenir, avec quel equipement ?
Production agricole contaminée ?

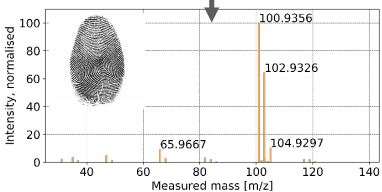
How to search for unknown unknowns?

Empa
Materials Science and Technology

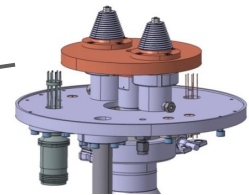
Target screening : Aprecon – GC – ToF-MS

Empa
Materials Science and Technology

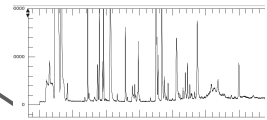




Electron impact (EI)
Time-of-Flight
Mass spectrometer
Tofwerk AG

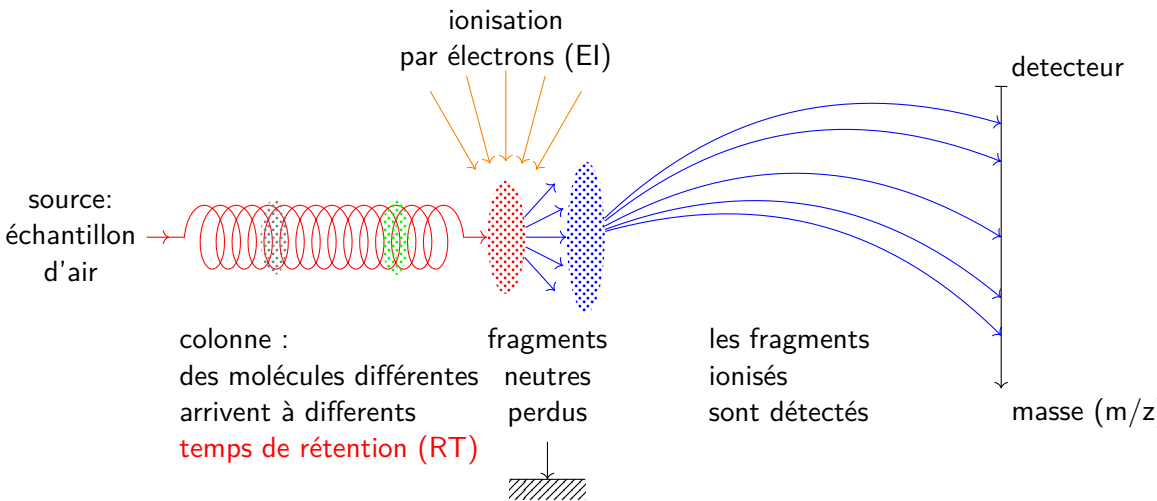


Pre-concentration
(APRECON)

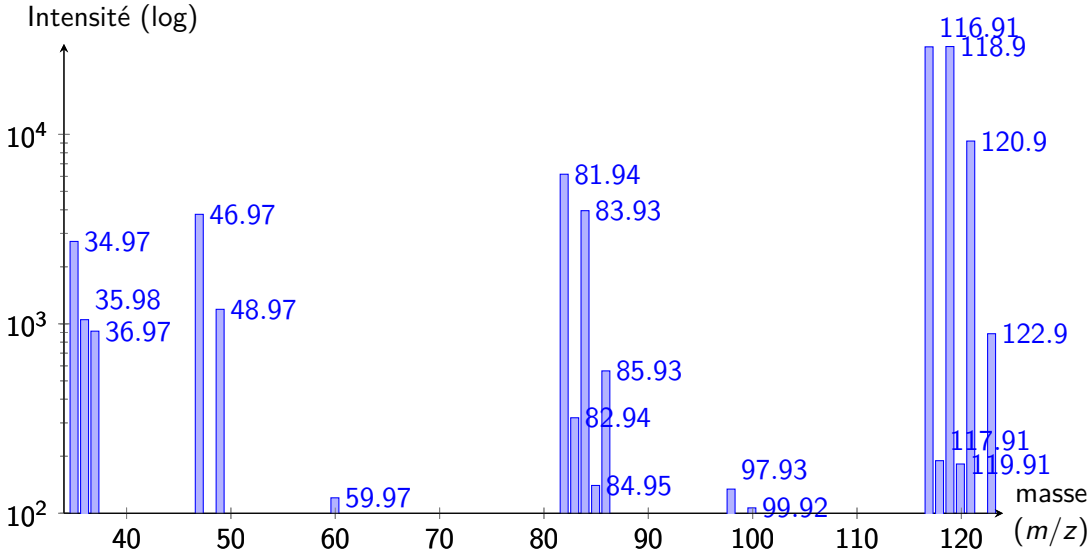


Gas Chromatography
GasPro column

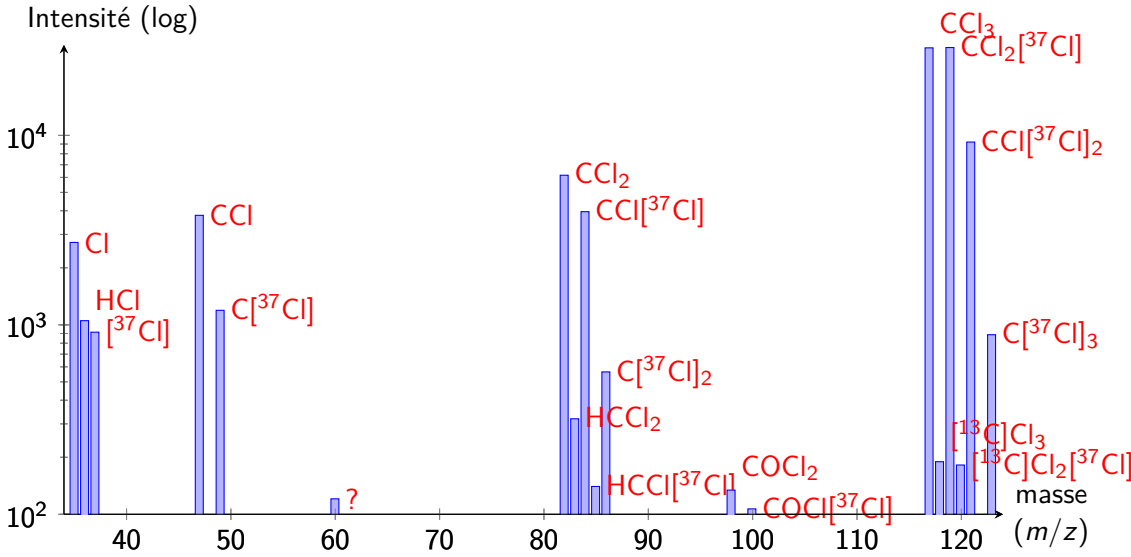
Spectromètre de masse à temps de vol et ionisation par électrons



Données d'entrée : un spectre de masse

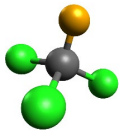


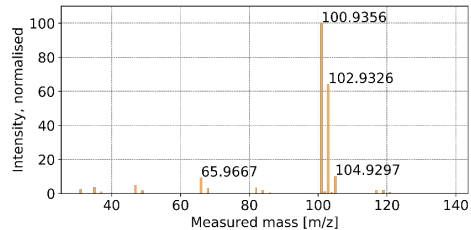
But : annoter la figure



Target vs non-target screening

Target screening:





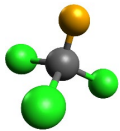
CFC-11
$CFCl_3$

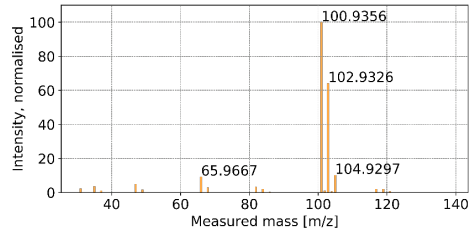
Instrumental fingerprint

Target vs non-target screening

Empa
Materials Science and Technology

Target screening:





CFC-11
$CFCl_3$

Instrumental fingerprint

Non-target screening

Workflow: Knapsack algorithm

Weight: 100.936 ± 0.0002 g/mol
(U = 2 ppm)

Which **atoms** can be packed together to match the measured masses, ± u?

9 atoms: H, C, N, O, S, F, Cl, Br, I

Un exemple : trouver la molécule inconnue

Données d'entrée :

- masse détectée : ≈ 16 m/z
- liste des masses IUPAC des atomes

Quelles possibilités ?

H	1.0078250319
B	11.00930536
C	12.
N	14.0030740074
O	15.9949146223
F	18.99840316
P	30.973762
S	31.97207073
Cl	34.96885271
Br	78.9183376
I	126.9044719

Informations supplémentaires

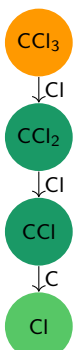
- Les atomes ont une valence : une propension à former des molécules ou pas
- Les éléments ont des isotopes : connaissez-vous le Carbone 14 ?
- Les spectres de masse sont issus de la fragmentation de molécules identiques : théorie des graphes

Hiérarchiser les informations

Algorithmes de graphes pour CCl_4

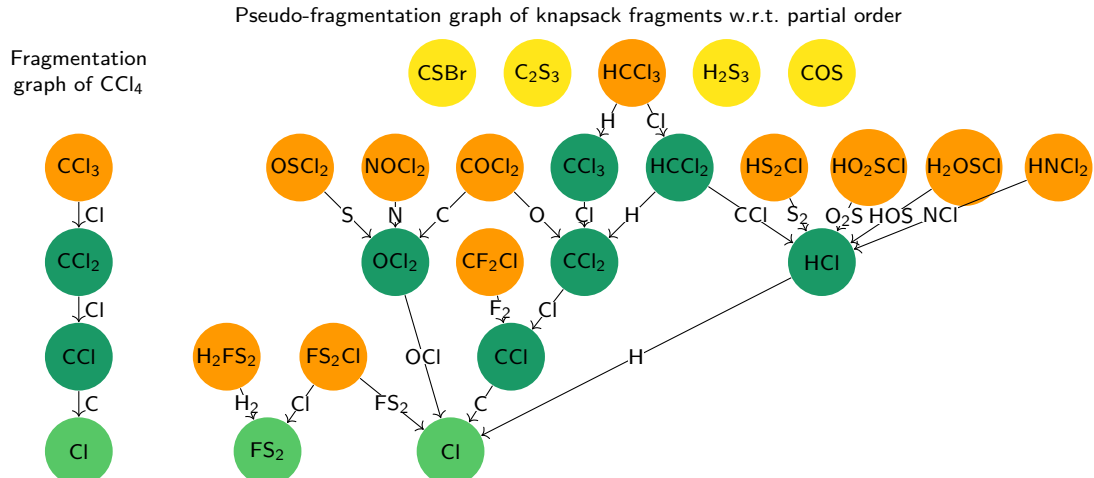
tout seul, Maximal, Nœud, feuille.

Fragmentation graph of CCl_4



Algorithmes de graphes pour CCl_4

tout seul, Maximal, Nœud, feuille.



Un peu de programmation Python

`https://gitlab.inria.fr/guillevi/alpinac/`
`https://members.loria.fr/AGuillevic/files/talks/knapsack.py`

Aspects du métier en recherche

Faire de la recherche :

- Trouver une question
- Réfléchir
- Tester
- Calculer
- Rédiger
- Publier

Guillevic *et al. J Cheminform* (2021) 13:78
https://doi.org/10.1186/s13321-021-00544-w

Journal of Cheminformatics

RESEARCH ARTICLE Open Access

Automated fragment formula annotation for electron ionisation, high resolution mass spectrometry: application to atmospheric measurements of halocarbons

Myriam Guillevic[1*], Aurore Guillevic[2], Martin K. Vollmer[1], Paul Schlauri[1], Matthias Hill[1], Lukas Emmenegger[1] and Stefan Reimann[1]

Abstract

Background: Non-target screening consists in searching a sample for all present substances, suspected or unknown, with very little prior knowledge about the sample. This approach has been introduced more than a decade ago in the field of water analysis, together with dedicated compound identification tools, but is still very scarce for indoor and atmospheric trace gas measurements, despite the clear need for a better understanding of the atmospheric trace gas composition. For a systematic detection of emerging trace gases in the atmosphere, a new and powerful analytical method is gas chromatography (GC) of preconcentrated samples, followed by electron ionisation, high resolution mass spectrometry (EI-HRMS). In this work, we present data analysis tools to enable automated fragment formula annotation for unknown compounds measured by GC-EI-HRMS.

Results: Based on co-eluting mass/charge fragments, we developed an innovative data analysis method to reliably reconstruct the chemical formulae of the fragments, using efficient combinatorics and graph theory. The method does not require the presence of the molecular ion, which is absent in ~40% of EI spectra. Our method has been trained and validated on >50 halocarbons and hydrocarbons, with 3–20 atoms and molar masses of 30–330 g mol^{-1}, measured with a mass resolution of approx. 3500. For >90% of the compounds, more than 90% of the annotated fragment formulae are correct. Cases of wrong identification can be attributed to the scarcity of detected fragments per compound or the lack of isotopic constraint (no minor isotopocule detected).

Conclusions: Our method enables to reconstruct most probable chemical formulae independently from spectral databases. Therefore, it demonstrates the suitability of EI-HRMS data for non-target analysis and paves the way for the identification of substances for which no EI mass spectrum is registered in databases. We illustrate the performances of our method for atmospheric trace gases and suggest that it may be well suited for many other types of samples. The L-GPL licenced Python code is released under the name ALPINAC for ALgorithmic Process for Identification of Non-targeted Atmospheric Compounds.

Aspects du métier en recherche

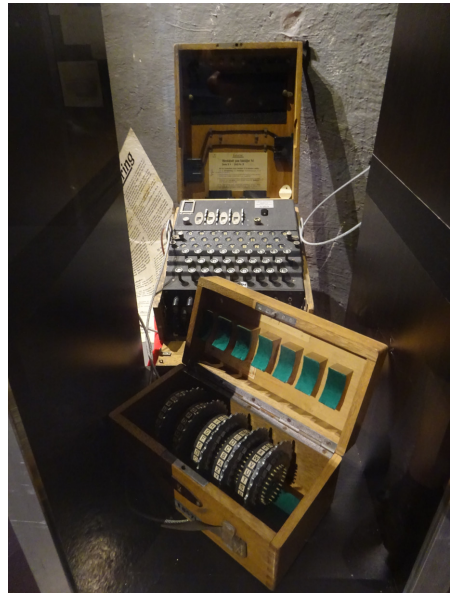
- travailler à plusieurs, souvent à distance
- co-auteurs
- articles en anglais
- conférences en Europe et dans le monde (US, Asie)
- Enseignement
- encadrement de stagiaires et doctorants
- médiation (comme aujourd'hui)
- administratif (comités de spécialistes, jury)
- chercher des financements
- déplacements, séjours à l'étranger

New Delhi, Inde, décembre 2017

Questions ?

Bonus : les débuts de l'informatique

Bletchley Park et le Colossus pour casser le chiffrement de la machine Enigma



Une Enigma 4 rotors au musée de la résistance de Bergen, Norvège