

Faire de la recherche en informatique

Aurore Guillevic

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

Lycée Jean-Mermoz, Saint-Louis, 2 février 2023



<https://members.loria.fr/AGuillevic/files/talks/23-saint-louis.pdf>

Parcours

<https://members.loria.fr/AGuillevic/>

- 2002–2005 : lycée à Hennebont, Morbihan
- 2005 : bac S (maths, physique, SVT) spécialité maths

Parcours

<https://members.loria.fr/AGuillevic/>

- 2002–2005 : lycée à Hennebont, Morbihan
- 2005 : bac S (maths, physique, SVT) spécialité maths

Études supérieures, université gratuite et bourses du CROUS :



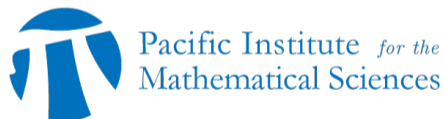
Universités de Haute Alsace, Strasbourg, Lorraine, Franche-Comté, UTBM Belfort
CROUS Alsace, Lorraine

- 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



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



H



C



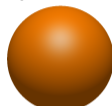
N



O



F



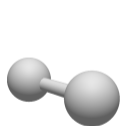
P



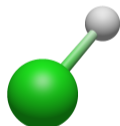
S



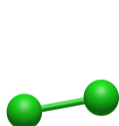
Cl



H₂



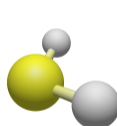
HCl



Cl₂



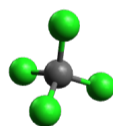
H₂O



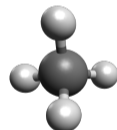
H₂S



NH₃



CCl₄

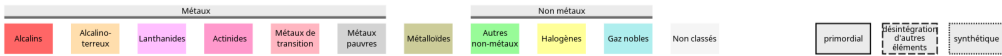


CH₄

Tableau périodique des éléments chimiques

Gruppe	1	II A	III A	IV A	V A	VI A	VII A	VIII	IX	X	XI	II B	III B	IV B	V B	VI B	VII B	18
1	Hydrogène 1 H 1,007975																	Hélium 2 He 4,002602
2	Lithium 3 Li 6,9395	Béryllium 4 Be 9,0121831											Bore 5 B 10,8135	Carbone 6 C 12,0106	Azote 7 N 14,006855	Oxygène 8 O 15,99940	Fluor 9 F 18,99840316	Néon 10 Ne 20,1797(6)
3	Sodium 11 Na 22,98976928	Magnésium 12 Mg 24,3055											Aluminium 13 Al 26,9815385	Silicium 14 Si 28,085(1)	Phosphore 15 P 30,97376200	Soufre 16 S 32,0675	Chlore 17 Cl 35,4515	Argon 18 Ar 39,948(1)
4	Potassium 19 K 39,0983(1)	Calcium 20 Ca 40,078(4)	Scandium 21 Sc 44,955908(5)	Titane 22 Ti 47,867(1)	Vanadium 23 V 50,9415(1)	Chrome 24 Cr 51,9961(6)	Manganèse 25 Mn 54,938044	Fer 26 Fe 55,845(2)	Cobalt 27 Co 58,933194	Nickel 28 Ni 58,6934(4)	Cuivre 29 Cu 63,546(3)	Zinc 30 Zn 65,38(2)	Gallium 31 Ga 69,723(1)	Germanium 32 Ge 72,630(8)	Arsenic 33 As 74,921595	Sélénium 34 Se 78,971(8)	Brome 35 Br 79,904	Krypton 36 Kr 83,798(2)
5	Rubidium 37 Rb 85,4678(3)	Strontium 38 Sr 87,62(1)	Yttrium 39 Y 88,90584	Zirconium 40 Zr 91,224(2)	Niobium 41 Nb 92,90637	Molybdène 42 Mo 95,95(1)	Technétium 43 Tc [98]	Ruthénium 44 Ru 101,07(2)	Rhodium 45 Rh 102,90550	Palladium 46 Pd 106,42(1)	Argent 47 Ag 107,8682(2)	Cadmium 48 Cd 112,414(4)	Indium 49 In 114,818(1)	Étain 50 Sn 118,710(7)	Antimoine 51 Sb 121,760(1)	Tellure 52 Te 127,60(3)	Iode 53 I 126,90447	Xénon 54 Xe 131,293(6)
6	Césium 55 Cs 132,905452	Baryum 56 Ba 137,327(7)	Lanthanides 57-71	Hafnium 72 Hf 178,49(2)	Tantale 73 Ta 180,94788	Tungstène 74 W 183,84(1)	Rhénium 75 Re 186,207(1)	Osmium 76 Os 190,23(3)	Iridium 77 Ir 192,217(3)	Platine 78 Pt 195,084(6)	Or 79 Au 196,966569	Mercury 80 Hg 200,592(3)	Thallium 81 Tl 204,3835	Plomb 82 Pb 207,2(1)	Bismuth 83 Bi 208,98040	Polonium 84 Po [209]	Astate 85 At [210]	Radon 86 Rn [222]
7	Francium 87 Fr [223]	Radium 88 Ra [226]	Actinides 89-103	Rutherfordium 104 Rf [267]	Dubnium 105 Db [268]	Seaborgium 106 Sg [269]	Bohrium 107 Bh [270]	Hassium 108 Hs [277]	Méitnérium 109 Mt [278]	Darmstadtium 110 Ds [281]	Roentgenium 111 Rg [282]	Copernicium 112 Cn [285]	Nihonium 113 Nh [286]	Flerovium 114 Fl [289]	Moscovium 115 Mc [289]	Livermorium 116 Lv [293]	Tennessee 117 Ts [294]	Oganesson 118 Og [294]
				Lanthane 57 La 138,90547	Cérium 58 Ce 140,116(1)	Praséodyme 59 Pr 140,90766	Néodyme 60 Nd 144,242(3)	Prométhium 61 Pm [145]	Samarium 62 Sm 150,36(2)	Europium 63 Eu 151,964(1)	Gadolinium 64 Gd 157,25(3)	Terbium 65 Tb 158,92535	Dysprosium 66 Dy 162,500(1)	Holmium 67 Ho 164,93033	Erbium 68 Er 167,259(3)	Thulium 69 Tm 168,93422	Ytterbium 70 Yb 173,045	Lucétium 71 Lu 174,9668
				Actinium 89 Ac [227]	Thorium 90 Th 232,0377	Protactinium 91 Pa 231,03588	Uranium 92 U 238,02891	Neptunium 93 Np [237]	Plutonium 94 Pu [244]	Amercium 95 Am [243]	Curium 96 Cm [247]	Berkélium 97 Bk [247]	Californium 98 Cf [251]	Einsteinium 99 Es [252]	Fermium 100 Fm [257]	Mendelevium 101 Md [258]	Nobelium 102 No [259]	Lawrencium 103 Lr [266]

← nom de l'élément (**gaz**, **liquide** ou **solide** à 0°C et 101,3 kPa)
 ← numéro atomique
 ← symbole chimique
 ← masse atomique relative [ou celle de l'isotope le plus stable]
 © [CIAAW "Atomic Weights 2013" + rev. 2015]



Par Scaler, Michka B – Travail personnel, CC BY-SA 3.0, <https://commons.wikimedia.org/w/index.php?curid=8985780>

La pollution de l'air

Combustion incomplète : Monoxyde de Carbone CO

NO_x (NO, NO₂) 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'entrepôts 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 O₃, dioxyde d'azote NO₂, dioxyde de soufre SO₂, 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 ?

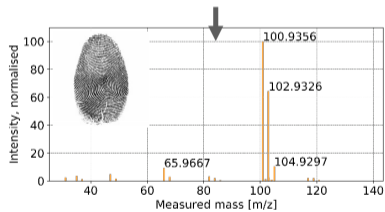
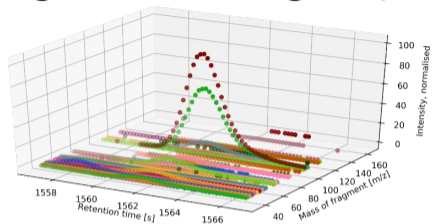
Les pompiers peuvent-ils intervenir, avec quel équipement ?

Production agricole contaminée ?

How to search for unknown unknowns?



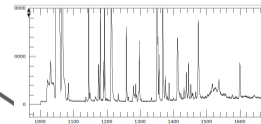
Target screening : Aprecon – GC – ToF-MS



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

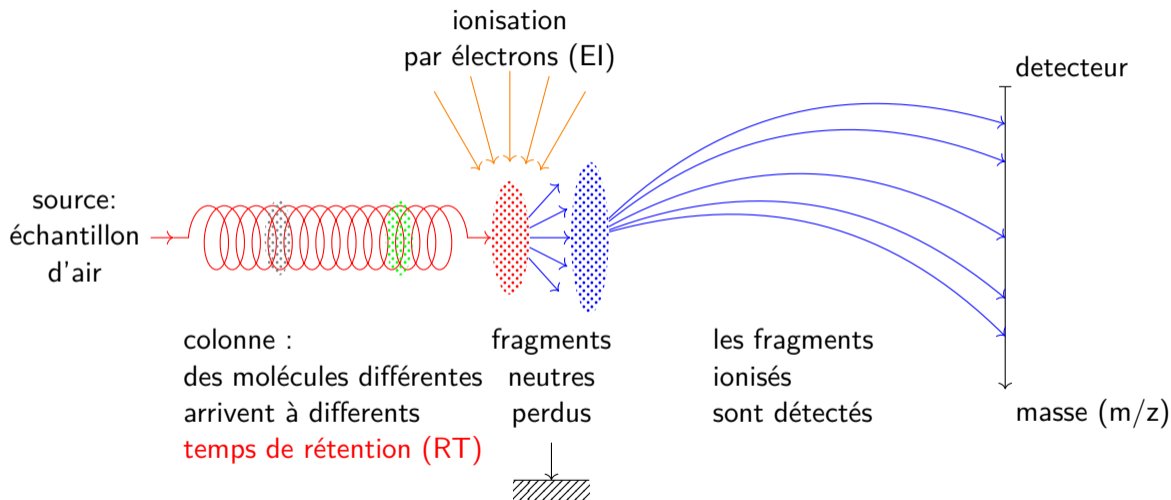


Pre-concentration
(APRECON)

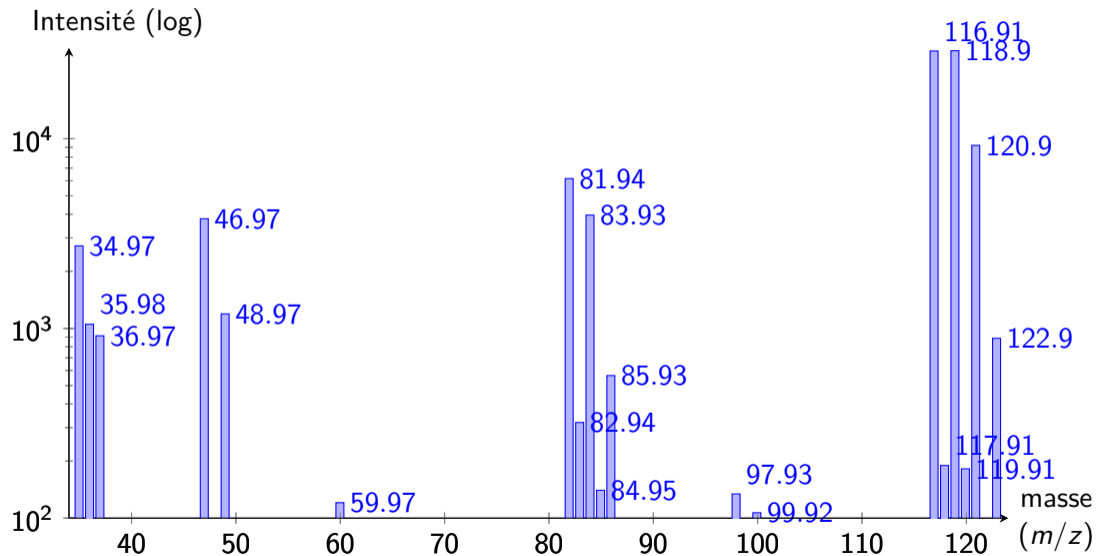


Gas Chromatography
GasPro column

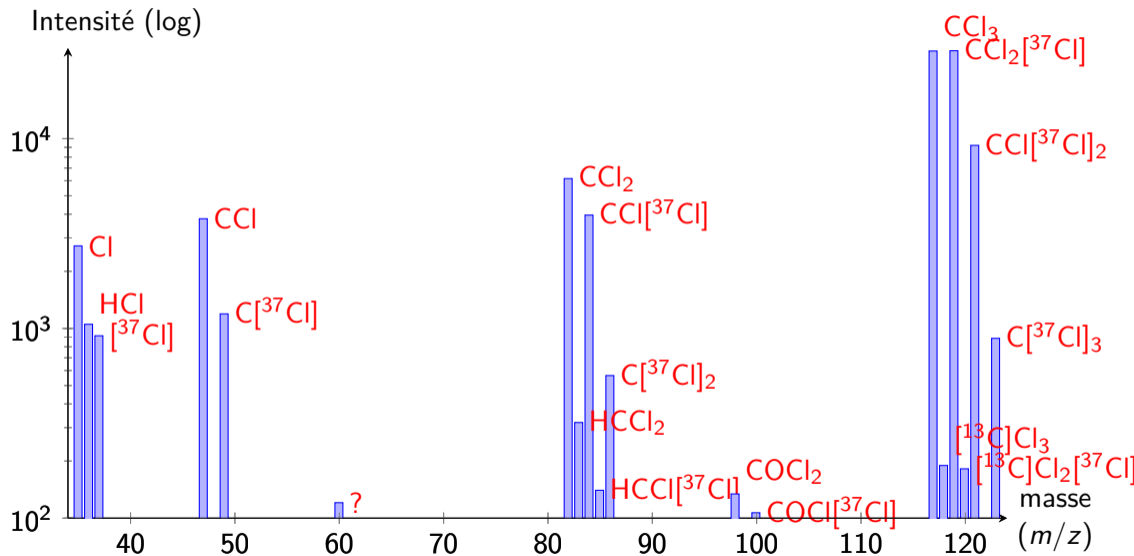
Spectromètre de masse à temps de vol et ionisation par électrons



Données d'entrée : un spectre de masse

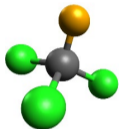


But : annoter la figure

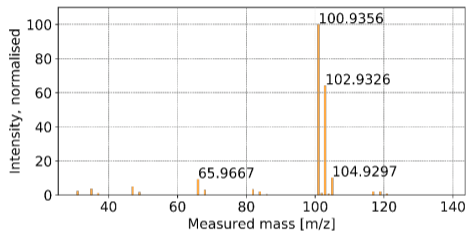


Target vs non-target screening

Target screening:



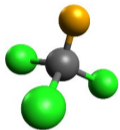
CFC-11
CFCl3



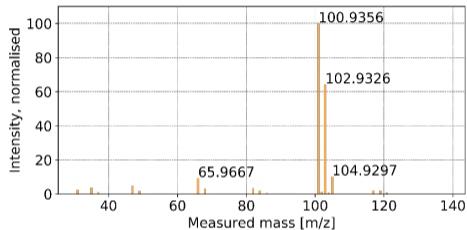
Instrumental fingerprint

Target vs non-target screening

Target screening:



CFC-11
CFCl3



Instrumental fingerprint



Workflow: Knapsack algorithm



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

Which **atoms** can be packed together to match the measured masses, $\pm u$?

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 arrondi : Hydrogène H 1, Bore B 11, Carbone C 12, Azote N 14, Oxygène O 16.

Quelles possibilités ? **Combinatoire**

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

Un exemple : trouver la molécule inconnue

Données d'entrée :

- masse détectée : ≈ 16 m/z
- liste des masses IUPAC des atomes arrondi : Hydrogène H 1, Bore B 11, Carbone C 12, Azote N 14, Oxygène O 16.

Quelles possibilités ? **Combinatoire**

- $O \approx 16$ m/z
- $NH_2 \approx 14 + 2 = 16$ m/z
- $CH_4 \approx 12 + 4 = 16$ m/z
- $BH_5 \approx 11 + 5 = 16$ m/z
- $H_{16} \approx 16$ m/z

Mais c'est "en théorie" !





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

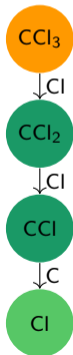
- Les atomes ont une **valence** : une propension à former des molécules ou pas
BH₅ et H₁₆ sont chimiquement impossibles
- Les éléments ont des **isotopes** : connaissez-vous le Carbone 14 ?
- Les spectres de masse sont issus de la fragmentation de molécules identiques :
théorie des graphes

Hiérarchiser les informations

Algorithmes de graphes pour CCl_4

 tout seul,  Maximal,  Nœud,  feuille.

Fragmentation
graph of CCl_4

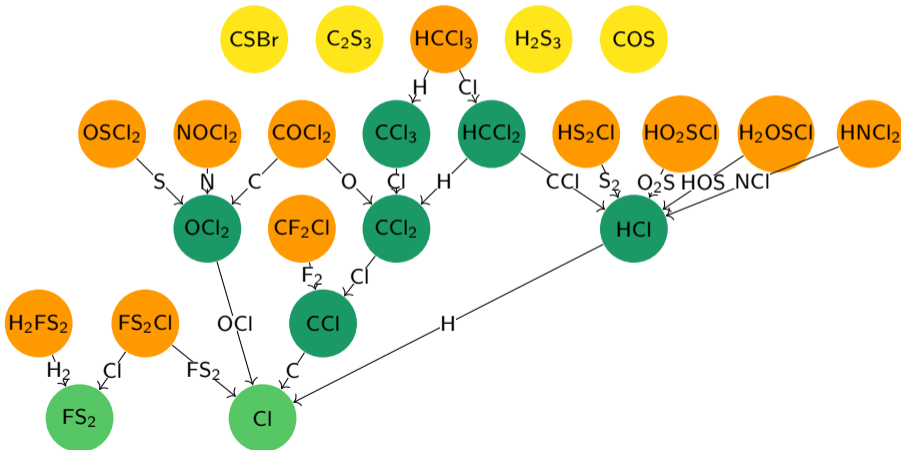
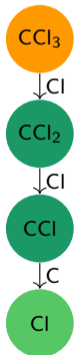


Algorithmes de graphes pour CCl_4

● tout seul,
 ● Maximal,
 ● Nœud,
 ● feuille.

Pseudo-fragmentation graph of knapsack fragments w.r.t. partial order

Fragmentation graph of CCl_4



Un peu de programmation Python

<https://gitlab.inria.fr/guillevi/alpinac/>

<https://members.loria.fr/AGuillevic/files/talks/knapsack.py>

RESEARCH ARTICLE

Open Access



Faire de la recherche :

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

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

Miriam Guillevic^{1*}, Aurore Guillevic², Martin K. Vollmer¹, Paul Schlauri¹, Matthias Hill¹, Lukas Emmenegger¹ and Stefan Reimann¹

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 pre-concentrated 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 ?