

**THESE
POUR LE DIPLOME D'ETAT
DE DOCTEUR EN PHARMACIE**

**Soutenue publiquement le
Par M Julien CORDONNIER**

**CREATION, GESTION, ET INTERROGATION DE BASES DE DONNEES DE
RESONANCE MAGNETIQUE NUCLEAIRE ET DE SPECTROMETRIE DE MASSE POUR LA
DEREPLICATION DE SUBSTANCES NATURELLES EN MELANGE**

Membres du jury :

Président : PU. RIVIERE Céline | UFR3S, Pharmacie

Directeur, conseiller de thèse : MCU. REMY Simon | URCA, Pharmacie

Assesseur(s) : PU. SAHPAZ Sevser | UFR3S, Pharmacie

 	LISTE GREE	LG/FAC/001
FACULTE DE PHARMACIE	Enseignants et Enseignants-chercheurs 2024-2025	Version 2.2 Applicable au 02/01/2022
Document transversal		Page 3/9

REDACTION	VERIFICATION	APPROBATION
Audrey Hennebelle Assistante de direction	Cyrille Porta Responsable des Services	Delphine Allorge Doyen

Université de Lille

Président
Premier Vice-président
Vice-présidente Formation
Vice-président Recherche
Vice-président Ressources humaines
Directrice Générale des Services

Régis BORDET
Etienne PEYRAT
Corinne ROBACZEWSKI
Olivier COLOT
Bertrand DÉCAUDIN
Anne-Valérie CHIRIS-FABRE

UFR3S

Doyen
Premier Vice-Doyen, Vice-Doyen RH, SI et Qualité
Vice-Doyenne Recherche
Vice-Doyen Finances et Patrimoine
Vice-Doyen International
Vice-Doyen Coordination pluriprofessionnelle et Formations sanitaires
Vice-Doyenne Formation tout au long de la vie
Vice-Doyen Territoire-Partenariats
Vice-Doyen Santé numérique et Communication
Vice-Doyenne Vie de Campus
Vice-Doyen étudiant

Dominique LACROIX
Hervé HUBERT
Karine FAURE
Damien CUNY
Vincent DERAMECOURT
Sébastien D'HARANCY
Caroline LANIER
Thomas MORGENROTH
Vincent SOBANSKI
Anne-Laure BARBOTIN
Valentin ROUSSEL

Faculté de Pharmacie

Doyen
Premier Assesseur et
Assesseur à la Santé et à l'Accompagnement
Assesseur à la Vie de la Faculté et
Assesseur aux Ressources et Personnels
Responsable des Services
Représentant étudiant
Chargé de mission 1er cycle
Chargée de mission 2eme cycle
Chargé de mission Accompagnement et Formation à la Recherche
Chargé de mission Relations Internationales
Chargée de Mission Qualité
Chargé de mission dossier HCERES

Delphine ALLORGE

Anne GARAT

Emmanuelle LIPKA
Cyrille PORTA
Honoré GUISE
Philippe GERVOIS
Héloïse HENRY
Nicolas WILLAND
Christophe FURMAN
Marie-Françoise ODOU
Réjane LESTRELIN

 	LISTE GEREE	LG/FAC/001
FACULTE DE PHARMACIE	Enseignants et Enseignants-chercheurs 2024-2025	Version 2.2 Applicable au 02/01/2022
Document transversal		Page 4/9

Professeurs des Universités - Praticiens Hospitaliers (PU-PH)

Civ.	Nom	Prénom	Service d'enseignement	Section CNU
Mme	ALLORGE	Delphine	Toxicologie et Santé publique	81
M.	BROUSSEAU	Thierry	Biochimie	82
M.	DÉCAUDIN	Bertrand	Biopharmacie, Pharmacie galénique et hospitalière	81
M.	DINE	Thierry	Pharmacologie, Pharmacocinétique et Pharmacie clinique	81
Mme	DUPONT-PRADO	Annabelle	Hématologie	82
Mme	GOFFARD	Anne	Bactériologie - Virologie	82
M.	GRESSIER	Bernard	Pharmacologie, Pharmacocinétique et Pharmacie clinique	81
M.	ODOU	Pascal	Biopharmacie, Pharmacie galénique et hospitalière	80
Mme	POULAIN	Stéphanie	Hématologie	82
M.	SIMON	Nicolas	Pharmacologie, Pharmacocinétique et Pharmacie clinique	81
M.	STAELS	Bart	Biologie cellulaire	82

Professeurs des Universités (PU)

Civ.	Nom	Prénom	Service d'enseignement	Section CNU
M.	ALIOUAT	El Moukhtar	Parasitologie - Biologie animale	87
Mme	AZAROUAL	Nathalie	Biophysique - RMN	85
M.	BERLARBI	Karim	Physiologie	86
M.	BERTIN	Benjamin	Immunologie	87
M.	BLANCHEMAIN	Nicolas	Pharmacotechnie industrielle	85
M.	CARNOY	Christophe	Immunologie	87
M.	CAZIN	Jean-Louis	Pharmacologie, Pharmacocinétique et Pharmacie clinique	86

 	LISTE GEREE	LG/FAC/001
FACULTE DE PHARMACIE	Enseignants et Enseignants-chercheurs 2024-2025	Version 2.2 Applicable au 02/01/2022
Document transversal		Page 3/9

M.	CUNY	Damien	Sciences végétales et fongiques	87
Mme	DELBAERE	Stéphanie	Biophysique - RMN	85
Mme	DEPREZ	Rebecca	Chimie thérapeutique	86
M.	DEPREZ	Benoît	Chimie bioinorganique	85
M.	DURIEZ	Patrick	Physiologie	86
M.	ELATI	Mohamed	Biomathématiques	27
M.	FOLIGNÉ	Benoît	Bactériologie - Virologie	87
Mme	FOULON	Catherine	Chimie analytique	85
M.	GARÇON	Guillaume	Toxicologie et Santé publique	86
M.	GOOSSENS	Jean-François	Chimie analytique	85
M.	HENNEBELLE	Thierry	Pharmacognosie	86
M.	LEBEGUE	Nicolas	Chimie thérapeutique	86
M.	LEMDANI	Mohamed	Biomathématiques	26
Mme	LESTAVEL	Sophie	Biologie cellulaire	87
Mme	LESTRELIN	Réjane	Biologie cellulaire	87
Mme	LIPKA	Emmanuelle	Chimie analytique	85
Mme	MELNYK	Patricia	Chimie physique	85
M.	MILLET	Régis	Institut de Chimie Pharmaceutique Albert Lespagnol	86
Mme	MUHR-TAILLEUX	Anne	Biochimie	87
Mme	PERROY	Anne-Catherine	Droit et Economie pharmaceutique	86
Mme	RIVIÈRE	Céline	Pharmacognosie	86
Mme	ROMOND	Marie-Bénédicte	Bactériologie - Virologie	87
Mme	SAHPAZ	Sevser	Pharmacognosie	86
M.	SERGHERAERT	Éric	Droit et Economie pharmaceutique	86
M.	SIEPMANN	Juergen	Pharmacotechnie industrielle	85

 	LISTE GEREE	LG/FAC/001
FACULTE DE PHARMACIE	Enseignants et Enseignants-chercheurs 2024-2025	Version 2.2 Applicable au 02/01/2022
Document transversal		Page 4/9

Mme	SIEPMANN	Florence	Pharmacotechnie industrielle	85
M.	WILLAND	Nicolas	Chimie organique	86

Maîtres de Conférences - Praticiens Hospitaliers (MCU-PH)

Civ.	Nom	Prénom	Service d'enseignement	Section CNU
Mme	CUVELIER	Élodie	Pharmacologie, Pharmacocinétique et Pharmacie clinique	81
Mme	DANEL	Cécile	Chimie analytique	85
Mme	DEMARET	Julie	Immunologie	82
Mme	GARAT	Anne	Toxicologie et Santé publique	81
Mme	GENAY	Stéphanie	Biopharmacie, Pharmacie galénique et hospitalière	81
M.	GRZYCH	Guillaume	Biochimie	82
Mme	HENRY	Héloïse	Biopharmacie, Pharmacie galénique et hospitalière	80
M.	LANNOY	Damien	Biopharmacie, Pharmacie galénique et hospitalière	80
Mme	MASSE	Morgane	Biopharmacie, Pharmacie galénique et hospitalière	81
Mme	ODOU	Marie-Françoise	Bactériologie - Virologie	82

Maîtres de Conférences des Universités (MCU)

Civ.	Nom	Prénom	Service d'enseignement	Section CNU
Mme	ALIOUAT	Cécile-Marie	Parasitologie - Biologie animale	87
M.	ANTHÉRIEU	Sébastien	Toxicologie et Santé publique	86
Mme	AUMERCIER	Pierrette	Biochimie	87
M.	BANTUBUNGI-BLUM	Kadiombo	Biologie cellulaire	87
M.	BERTHET	Jérôme	Biophysique - RMN	85
M.	BOCHU	Christophe	Biophysique - RMN	85
M.	BORDAGE	Simon	Pharmacognosie	86

 	LISTE GEREE	LG/FAC/001
FACULTE DE PHARMACIE	Enseignants et Enseignants-chercheurs 2024-2025	Version 2.2 Applicable au 02/01/2022
Document transversal		Page 3/9

M.	BOSC	Damien	Chimie thérapeutique	86
Mme	BOU KARROUM	Nour	Chimie bioinorganique	
M.	BRIAND	Olivier	Biochimie	87
Mme	CARON-HOUDE	Sandrine	Biologie cellulaire	87
Mme	CARRIÉ	Hélène	Pharmacologie, Pharmacocinétique et Pharmacie clinique	86
Mme	CHABÉ	Magali	Parasitologie - Biologie animale	87
Mme	CHARTON	Julie	Chimie organique	86
M.	CHEVALIER	Dany	Toxicologie et Santé publique	86
Mme	DEMANCHE	Christine	Parasitologie - Biologie animale	87
Mme	DEMARQUILLY	Catherine	Biomathématiques	85
M.	DHIFLI	Wajdi	Biomathématiques	27
Mme	DUMONT	Julie	Biologie cellulaire	87
M.	EL BAKALI	Jamal	Chimie thérapeutique	86
M.	FARCE	Amaury	Institut de Chimie Pharmaceutique Albert Lespagnol	86
M.	FLIPO	Marion	Chimie organique	86
M.	FRULEUX	Alexandre	Sciences végétales et fongiques	
M.	FURMAN	Christophe	Institut de Chimie Pharmaceutique Albert Lespagnol	86
M.	GERVOIS	Philippe	Biochimie	87
Mme	GOOSSENS	Laurence	Institut de Chimie Pharmaceutique Albert Lespagnol	86
Mme	GRAVE	Béatrice	Toxicologie et Santé publique	86
M.	HAMONIER	Julien	Biomathématiques	26
Mme	HAMOUDI-BEN YELLES	Chérifa-Mounira	Pharmacotechnie industrielle	85
Mme	HANNOTHIAUX	Marie-Hélène	Toxicologie et Santé publique	86
Mme	HELLEBOID	Audrey	Physiologie	86
M.	HERMANN	Emmanuel	Immunologie	87

 	LISTE GEREE	LG/FAC/001
FACULTE DE PHARMACIE	Enseignants et Enseignants-chercheurs 2024-2025	Version 2.2 Applicable au 02/01/2022
Document transversal		Page 4/9

M.	KAMBIA KPAKPAGA	Nicolas	Pharmacologie, Pharmacocinétique et Pharmacie clinique	86
M.	KARROUT	Younes	Pharmacotechnie industrielle	85
Mme	LALLOYER	Fanny	Biochimie	87
Mme	LECOEUR	Marie	Chimie analytique	85
Mme	LEHMANN	Hélène	Droit et Economie pharmaceutique	86
Mme	LELEU	Natascha	Institut de Chimie Pharmaceutique Albert Lespagnol	86
M.	LIBERELLE	Maxime	Biophysique - RMN	
Mme	LOINGEVILLE	Florence	Biomathématiques	26
Mme	MARTIN	Françoise	Physiologie	86
M.	MENETREY	Quentin	Bactériologie - Virologie	
M.	MOREAU	Pierre-Arthur	Sciences végétales et fongiques	87
M.	MORGENROTH	Thomas	Droit et Economie pharmaceutique	86
Mme	MUSCHERT	Susanne	Pharmacotechnie industrielle	85
Mme	NIKASINOVIC	Lydia	Toxicologie et Santé publique	86
Mme	PINÇON	Claire	Biomathématiques	85
M.	PIVA	Frank	Biochimie	85
Mme	PLATEL	Anne	Toxicologie et Santé publique	86
M.	POURCET	Benoît	Biochimie	87
M.	RAVAUX	Pierre	Biomathématiques / Innovations pédagogiques	85
Mme	RAVEZ	Séverine	Chimie thérapeutique	86
Mme	ROGEL	Anne	Immunologie	
M.	ROSA	Mickaël	Hématologie	
M.	ROUMY	Vincent	Pharmacognosie	86
Mme	SEBTI	Yasmine	Biochimie	87

 	LISTE GEREE	LG/FAC/001
FACULTE DE PHARMACIE	Enseignants et Enseignants-chercheurs 2024-2025	Version 2.2 Applicable au 02/01/2022
Document transversal		Page 3/9

Mme	SINGER	Elisabeth	Bactériologie - Virologie	87
Mme	STANDAERT	Annie	Parasitologie - Biologie animale	87
M.	TAGZIRT	Madjid	Hématologie	87
M.	VILLEMAGNE	Baptiste	Chimie organique	86
M.	WELTI	Stéphane	Sciences végétales et fongiques	87
M.	YOUS	Saïd	Chimie thérapeutique	86
M.	ZITOUNI	Djamel	Biomathématiques	85

Professeurs certifiés

Civ.	Nom	Prénom	Service d'enseignement
Mme	FAUQUANT	Soline	Anglais
M.	HUGES	Dominique	Anglais
Mme	KUBIK	Laurence	Anglais
M.	OSTYN	Gaël	Anglais

Professeurs Associés

Civ.	Nom	Prénom	Service d'enseignement	Section CNU
M.	DAO PHAN	Haï Pascal	Chimie thérapeutique	86
M.	DHANANI	Alban	Droit et Economie pharmaceutique	86

 	LISTE GEREE	LG/FAC/001
FACULTE DE PHARMACIE	Enseignants et Enseignants-chercheurs 2024-2025	Version 2.2 Applicable au 02/01/2022
Document transversal		Page 4/9

Maîtres de Conférences Associés

Civ.	Nom	Prénom	Service d'enseignement	Section CNU
M.	COUSEIN	Etienne	Biopharmacie, Pharmacie galénique et hospitalière	
Mme	CUCCHI	Malgorzata	Biomathématiques	85
M.	DUFOSSEZ	François	Biomathématiques	85
M.	FRIMAT	Bruno	Pharmacologie, Pharmacocinétique et Pharmacie clinique	85
M.	GILLOT	François	Droit et Economie pharmaceutique	86
M.	MITOUMBA	Fabrice	Biopharmacie, Pharmacie galénique et hospitalière	86
M.	PELLETIER	Franck	Droit et Economie pharmaceutique	86

Assistants Hospitalo-Universitaire (AHU)

Civ.	Nom	Prénom	Service d'enseignement	Section CNU
M.	BOUDRY	Augustin	Biomathématiques	
Mme	DERAMOUDT	Laure	Pharmacologie, Pharmacocinétique et Pharmacie clinique	
Mme	GILLIOT	Sixtine	Biopharmacie, Pharmacie galénique et hospitalière	
M.	GISH	Alexandr	Toxicologie et Santé publique	
Mme	NEGRIER	Laura	Chimie analytique	

Hospitalo-Universitaire (PHU)

	Nom	Prénom	Service d'enseignement	Section CNU
M.	DESVAGES	Maximilien	Hématologie	
Mme	LENSKI	Marie	Toxicologie et Santé publique	

 	LISTE GEREE	LG/FAC/001
FACULTE DE PHARMACIE	Enseignants et Enseignants-chercheurs 2024-2025	Version 2.2 Applicable au 02/01/2022
Document transversal		Page 3/9

Attachés Temporaires d'Enseignement et de Recherche (ATER)

Civ.	Nom	Prénom	Service d'enseignement	Section CNU
Mme	BERNARD	Lucie	Physiologie	
Mme	BARBIER	Emeline	Toxicologie	
Mme	COMAPGNE	Nina	Chimie Organique	
Mme	COULON	Audrey	Pharmacologie, Pharmacocinétique et Pharmacie clinique	
M.	DUFOSSEZ	Robin	Chimie physique	
Mme	KOUAGOU	Yolène	Sciences végétales et fongiques	
M.	MACKIN MOHAMOUR	Synthia	Biopharmacie, Pharmacie galénique et hospitalière	

Enseignant contractuel

Civ.	Nom	Prénom	Service d'enseignement
M.	MARTIN MENA	Anthony	Biopharmacie, Pharmacie galénique et hospitalière
M.	MASCAUT	Daniel	Pharmacologie, Pharmacocinétique et Pharmacie clinique
Mme	NDIAYE-BOIDIN	Maguette	Anglais
M.	ZANETTI	Sébastien	Biomathématiques

CYCLE DE VIE DU DOCUMENT

Version	Modifié par	Date	Principales modifications
1.0		20/02/2020	Création
2.0		02/01/2022	Mise à jour
2.1		21/06/2022	Mise à jour
2.2		01/02/2024	Mise à jour



UFR3S-Pharmacie

L'Université n'entend donner aucune approbation aux opinions émises dans les thèses ; celles-ci sont propres à leurs auteurs.

SERMENT DE GALIEN



En présence des Maîtres de la Faculté, je fais le serment :

- D'honorer ceux qui m'ont instruit(e) dans les préceptes de mon art et de leur témoigner ma reconnaissance en restant fidèle aux principes qui m'ont été enseignés et d'actualiser mes connaissances
- D'exercer, dans l'intérêt de la santé publique, ma profession avec conscience et de respecter non seulement la législation en vigueur, mais aussi les règles de Déontologie, de l'honneur, de la probité et du désintéressement ;
- De ne jamais oublier ma responsabilité et mes devoirs envers la personne humaine et sa dignité. En aucun cas, je ne consentirai à utiliser mes connaissances et mon état pour corrompre les mœurs et favoriser des actes criminels ;
- De ne dévoiler à personne les secrets qui m'auraient été confiés ou dont j'aurais eu connaissance dans l'exercice de ma profession ;
- De faire preuve de loyauté et de solidarité envers mes collègues pharmaciens ;
- De coopérer avec les autres professionnels de santé.

Que les Hommes m'accordent leur estime si je suis fidèle à mes promesses.
Que je sois couvert(e) d'opprobre et méprisé(e) de mes confrères si j'y manque.

Remerciement



Moi ce que j'aime, c'est observer, contempler et chercher à comprendre. L'écriture ce n'est pas mon truc. Je n'ai d'ailleurs jamais été très doué pour ça. Voilà la raison pour laquelle je n'ai pas fait L, comme je l'ai souvent dit à mes parents. Ce travail a duré une éternité, sans cesse accompagnée par une tension entre l'attente et l'action. « *Trois mille six cents fois par heure, la Seconde Chuchote : Souviens-toi !* ». Procrastination ? Trop souvent, « *l'Espoir, Vaincu, pleure, et l'Angoisse atroce, despotique, Sur mon crâne incliné plant[a] son drapeau noir* ».

Mais au contraire de Baudelaire, je préfère me rappeler des instants « *Traversé[s] ça et là par de brillants soleils* » plutôt que par « *Le tonnerre et la pluie* ». C'est tout, c'est comme ça ! Et désormais « *Là, tout n'est qu'ordre et beauté, Luxe, calme et volupté* ».

Je tiens profondément à remercier le Professeur Céline Rivière qui m'a fait l'honneur de juger ce travail et de présider le jury de ma thèse. Veuillez trouver ici toute ma gratitude pour vos enseignements durant mon cursus, que j'espère désormais pouvoir partager à mon tour avec autant d'enthousiasme et de dévouement.

Mes remerciements vont également droit au Professeur Sevser Sahpaz, grâce à qui mon engouement pour la Pharmacognosie s'est renforcé sur les bancs de la Fac. Votre présence et vos paroles empreintes de bienveillance ont su m'encourager et apaiser mes incertitudes. Je tiens à exprimer en ces mots toute ma reconnaissance.

Je remercie le Docteur Simon Remy pour l'exigence, le cadre et les standards élevés, mais extrêmement formateurs qui m'ont permis d'atteindre un haut niveau d'exigence dans la réalisation de cette thèse.

Un très grand MERCI à Noah Alves-Dacosta et au Professeur Georges Massiot qui, à la manière de coéquipiers légendaires dans une quête épique, ont insufflé un second souffle pour atteindre le boss final avec brio.

Je témoigne toute ma sympathie au Professeur Jean-Hugues Renault qui a su, par son calme et sa sagesse créer un espace de travail constructif et bienveillant. Son approche laissera une empreinte durable sur ma manière d'appréhender les défis.

A mes parents, frères, grands-parents et à toutes celles et ceux de « la même tribu » qui, par leur amour, leur compréhension et leur patience, m'ont soutenu tout au long de cette aventure.

À celle qui sait pourquoi je lui dis merci, pour tout et pour ce que les mots ne peuvent pas exprimer.

Allez sur ceux on va aller boire un verre, ou on n'irait pas au bar, hum ? P'tit godet ?



Table des matières



Table des matières

Remerciement	XVII
Table des matières	XXI
Liste des figures	XXV
Introduction générale.....	1
Partie I : Bibliographie	7
I.1. Les substances naturelles.....	9
I.1.1. Définitions et généralités	9
I.1.2. Les alcaloïdes.....	10
I.2. Profilage chimique.....	21
I.2.1. Définitions et généralités	21
I.2.2. Acquisition des données spectrales expérimentales nécessaires à la déréplication.....	22
I.2.3. Analyse et visualisation des données spectrales MS/MS.....	30
I.3. Gestion des données en recherche.....	33
I.3.1. Garantir la qualité et la réutilisation des données de référence	34
I.3.2. Solutions mises en œuvre	36
I.4. Les bases de données	41
I.4.1. Généralités	41
I.4.2. Encodage numérique de la structure moléculaire	43
I.4.3. Rôle des bases de données dans le processus de déréplication	45
I.4.4. Etat général des limites des bases de données existantes	51
Partie 2 : Travaux expérimentaux	55
Partie 3 : Conclusion & Discussion	91
III. Conclusion & Discussion.....	93
Références.....	97



Liste des figures



Figure 1 : Liste des alcaloïdes découverts au début du XIXe siècle, selon Amé Pictet.....	11
Figure 2 : Noyaux des principaux acides aminés précurseurs d'alcaloïdes	14
Figure 3 : Réaction de condensation se produisant entre la tryptamine et un aldéhyde.....	15
Figure 4 : Numérotation biosynthétique des alcaloïdes mono-indoliques, selon LeMen et Taylor (14)	16
Figure 5 : Apport de la partie terpénique dans le cas des alcaloïdes mono-indoliques issus des plantes de la famille taxonomique des Apocynaceae.....	16
Figure 6 : Structures moléculaires de la 3- α (S)-strictosidine, et des alcaloïdes bis-indoliques vinblastine et vincristine issus de <i>Catharanthus roseus</i> (L.) G.Don	17
Figure 7 : Photo de <i>Catharanthus roseus</i> (L.) G.Don (selon Kenriez au Parque Botánico de Maspalomas (Gran Canaria)) et planche d'herbier du MNHN.....	18
Figure 8 : Couplage de l'unité indolique tétracyclique (16-méthoxyvelbamine) et d'une unité dihydroindolique (vindoline) donnant accès aux alcaloïdes de type vinblastine	19
Figure 9 : Structures moléculaires de la longicaudatine (alcaloïde bis-indolique) issue de <i>Strychnos longicaudata</i>	20
Figure 10 : Flux de travail en déréplication	22
Figure 11 : Schéma explicatif du fonctionnement d'un spectromètre de résonance magnétique (RMN), selon(31).....	25
Figure 12 : Schéma explicatif du fonctionnement d'un spectromètre de masse (ESI-Orbitrap), selon (31)	27
Figure 13: Schéma explicatif de la comparaison spectrale via le cosine-score, selon Wan et al. (44). 31	31
Figure 14 : Analyse coûts-avantage des données de recherche FAIR, selon (46)	33
Figure 15 : Encodage d'une structure moléculaire en InChI et InChIKey, selon https://www.inchi-trust.org/	44
Figure 16 : Structure du fichier MDL pour la Leucine, selon (64).....	45
Figure 17. Algorithmes d'attribution (Bleu : spectre expérimental, Rouge : spectre de référence), selon (92) A) Attribution bipartite, B) Attribution directe, C) Attribution inversée	49



Introduction générale



Les substances naturelles (SN) sont des composés organiques produits par le métabolisme d'organismes vivants. Elles résultent d'une longue évolution de la machinerie biologique de l'organisme producteur. Pour ces raisons, les SN sont aussi appelées métabolites spécialisés. Ces substances ont toujours fait l'objet d'usages en médecine traditionnelle au travers de différentes cultures. L'importante diversité structurale des SN et les nombreuses activités biologiques qui leur sont prêtées expliquent qu'environ 80 % de la population mondiale a recours à la médecine traditionnelle, principalement à des remèdes à base de plantes. (1) Près de 75 % des espèces de plantes à fleurs ont été nommées à l'heure actuelle, dont 670 nouvelles espèces décrites en 2021, ce qui suggère que le nombre total d'espèces végétale pourrait se stabiliser autour des 350 000. (2,3) Pourtant, une infime partie de cette diversité a fait l'objet d'un profilage chimique. En effet, moins de 40 000 espèces sont représentées dans la base de données structurales LOTUS dédiée aux SN. (4)

La recherche contemporaine offre un appui scientifique aux usages en médecine traditionnelle de certaines plantes en identifiant les composés responsables de l'activité biologique. La métabolomique est la science qui étudie l'ensemble des métabolites primaires (*i.e.* sucres, acides aminés, acides gras, *etc.*), mais surtout des métabolites spécialisés (*i.e.* polyphénols, flavonoïdes, alcaloïdes, *etc.*) présents dans une cellule, un organe ou un organisme. Aujourd'hui, le profilage chimique est basé le plus souvent sur l'utilisation, combinée ou non, de techniques analytiques sensibles et/ou robustes (*i.e.* spectrométrie de masse (MS), résonance magnétique nucléaire (RMN)). Cette étape est cruciale dans la cartographie chimique des SN présentes au sein d'un échantillon à un moment donné.

Les années 1950 ont marqué un tournant dans la découverte et l'application des alcaloïdes de la pervenche de Madagascar (*Catharanthus roseus* (L.) G.Don), en particulier la vincristine et la vinblastine. (5) Le Professeur et pharmacien Jean LeMen de la faculté de Médecine et de Pharmacie de Reims a entre autres participé à la purification et l'élucidation de nombreux vinca-alcaloïdes et autres dérivés uniques. La majorité d'entre eux sont conservés en l'état physique à l'UMR CNRS 7312 – Institut de Chimie moléculaire de Reims. Cette « alcaloïdothèque » est un atout précieux sur le plan intellectuel, et représente une source de « hits » potentiels pour la recherche de nouveaux candidats-médicaments en cancérologie, microbiologie, virologie, *etc.* La dérégulation, le criblage biologique à haut débit (*High Throughput Screening* – HTS), et les méthodes d'évaluation biologiques *in silico*, comme le docking, sont autant de méthodes qui contribuent d'ailleurs à raviver l'intérêt pour les SN et leurs dérivés en réponse aux défis de santé modernes.

Néanmoins, la conservation de ces molécules naturelles sous forme physique présente plusieurs limites. De nombreuses SN sont sensibles à la lumière, à la chaleur ou à l'oxygène, ce qui

entraîne des modifications chimiques et donc potentiellement une perte d'activité biologique au fil du temps. Les échantillons peuvent être également contaminés par des microorganismes, de la poussière ou d'autres substances, compromettant leur pureté et leur utilité en recherche. Enfin, certaines molécules peuvent être toxiques, nécessitant des précautions particulières pour éviter les risques d'exposition, de manipulation et d'élimination. Ainsi, l'espace et les coûts de stockage liés aux installations spécifiques (*e.g.* réfrigération, atmosphère contrôlée) sont à considérer. La sauvegarde numérique des SN, sous forme de spectres expérimentaux associés à un ensemble de métadonnées constitue une alternative durable pour leur conservation. La numérisation préserve de la perte d'informations précieuses en cas de dégradation ou de disparition des échantillons originaux. Les bases de données numériques favorisent d'autant plus le partage des connaissances entre chercheurs, et renforce la reproductibilité des études. Par ailleurs, l'accès aux spectres expérimentaux enregistrés limite la répétition d'expériences coûteuses, facilite l'identification rapide de molécules déjà connues (*i.e.* déréplication), et accélère le processus de découverte de nouvelles SN bio-actives.

Les travaux présentés dans ce manuscrit ont pour but d'initier la conservation numérique de l'alcaloïdothèque de Reims, tout en soulignant l'importance d'intégrer les connaissances historiques à la recherche scientifique moderne. Un bref rappel sur les SN et plus particulièrement sur les vinca-alcaloïdes sera apporté dans une première partie introductive. Les méthodes modernes de profilage chimique y seront également développées, suivie de recommandations quant à la gestion des données de recherche. L'état général des bases de données spectrales et de leurs limites seront plus particulièrement abordés. Les travaux expérimentaux relatifs à la construction de la « *Digital Alkaloids Collection from Reims - DACR* » seront présentés dans une seconde partie. Une conclusion et discussion viendront clôturer ce manuscrit.





Partie I : Bibliographie

I.1. Les substances naturelles

I.1.1. Définitions et généralités

Le métabolisme est défini comme étant l'ensemble des réactions biochimiques qui se déroulent au sein d'une cellule et qui conduisent à la synthèse de molécules nommées métabolites. Une distinction doit être faite entre les métabolites primaires et les métabolites spécialisés. Les premiers sont produits par les voies biosynthétiques universelles. Ces voies permettent la synthèse de molécules essentielles à la vie à partir de précurseurs basiques, comme le glucose. Les métabolites primaires sont maintenus en équilibre par des systèmes biologiques interconnectés et fonctionnant ensemble de manière harmonieuse pour assurer le bon fonctionnement d'un organisme. (6) Les métabolites primaires peuvent ensuite s'assembler et s'enchaîner selon les règles de la chimie supramoléculaire, pour former des biopolymères tels que les protéines (constituées d'acides aminés), les polysaccharides (constitués d'oses), et les acides nucléiques (constitués de nucléotides). Le métabolisme primaire constitue un cadre primordial pour l'évolution d'un métabolisme plus complexe.

Les métabolites spécialisés ou substances naturelles, sont définis comme des molécules qui *« sont produites par les êtres vivants de façon spécifique et non universelle, par des voies de biosynthèse indépendantes des voies du métabolisme primaire »*. Ces métabolites se caractérisent par leur distribution restreinte au sein du règne du vivant. Par exemple, la morphine n'est produite que chez *Papaver somniferum* (L.) (Papaveraceae). Ils sont également dotés d'une riche diversité et complexité structurale. Le rôle des métabolites spécialisés est de permettre à l'organisme producteur de s'adapter à son milieu de vie. Sa défense, la communication et l'attraction d'autres organismes dépendent de ces métabolites. C'est ce qui est appelé l'écologie chimique. François Jacob qualifiait le métabolisme spécialisé comme étant *« plus adaptatif, plastique »* et ses voies de biosynthèse *« évoluées comme des voies détournées du métabolisme primaire »*. (6)

La fin du XIXe siècle a marqué une évolution significative de la phytochimie, caractérisée par la classification systématique des composés en fonction de leur structure chimique (e.g. flavonoïdes, terpènes, etc.). Au cours de cette période, les chimistes ont adopté une approche méthodique semblable à la classification des espèces par les naturalistes, ce qui a permis de mieux comprendre les substances naturelles. Les alcaloïdes décrivent une des classes établies.

I.1.2. Les alcaloïdes

Le terme « alcaloïde » a été introduit pour la première fois en 1819 à l'issu des travaux de W. Meissner. Ce terme peut être décomposé par la racine « alcali » signifiant « base ou à caractère alcalin ou basique » et le suffixe « oïde » signifiant, « semblable à ». (7) La découverte de SN basiques constitua une nouvelle étape en phytochimie, car jusqu'alors, seuls des composés acides ou neutres étaient connus. (7)

I.1.2.1. Définitions et généralités

La définition la plus générale qui puisse être établie est la suivante : « *Un alcaloïde est une substance naturelle azotée, [qui existe] à l'état de sel et biosynthétiquement formé à partir d'un acide aminé. [Il est] plus ou moins basique, de distribution restreinte et douée, fréquemment, de propriétés pharmacologiques parfois puissantes* ». (6,8)

Ces éléments caractérisent ce que l'on peut appeler les alcaloïdes vrais.

L'isolement des alcaloïdes au XIXe siècle a été grandement facilité par l'invention du processus de percolation, développé par le chimiste français M.E. Chevreul pour l'extraction des substances actives. Grâce à cette méthode, les premiers « principes actifs » des plantes ont pu être identifiés, ouvrant ainsi la voie à une utilisation plus rationnelle et scientifique des médicaments. (9)

Parmi les premières grandes découvertes en rapport avec les alcaloïdes, nous pouvons citer la narcotine isolée à partir de l'opium par Robiquet en 1817, puis la caféine à partir de *Coffea arabica* (L.) simultanément avec le chimiste allemand F. F. Runge. Ces molécules étaient les premières d'une longue série d'autres alcaloïdes majeurs. L'apothicaire hanovrien Sertürner a poursuivi ses recherches sur l'extrait d'opium et a réussi à en extraire la morphine, ouvrant la voie à l'isolement systématique d'autres alcaloïdes végétaux. Enfin, les pharmaciens français P.J. Pelletier et J.B. Caventou ont tout autant joué un rôle capital en découvrant des composés clés comme la strychnine et la brucine (*Strychnos nux-vomica*(L.)), l'émétine (*Carapichea ipecacuanha* (Brot. L.Andersson)), la pipérine (*Piper nigrum* (L.)), la colchicine (*Colchicum autumnale* (L.)) la coniine (*Conium maculatum* (L.)), ou encore la quinine (*Cinchona pubescens* (Vahl) et *Cinchona calisaya* (Wedd.)) (c.f. Fig.1).



1817	Narcotine	par	Robiquet
	Éméline		Pelletier et Magendie
1818	Vératrine		Meissner
	Strychnine		Pelletier et Caventou
1819	Brucine		les mêmes
	Pipérine		Oersted
1820	Caféine		Runge
	Cinchonine		Pelletier et Caventou
	Quinine		les mêmes
	[Solanine		Desfosses]
1825	Sinapine		Henry et Garot
1826	Corydaline		Wackenroder
	Berbéline		Chevallier et Pelletan
1828	Nicotine		Posselt et Reimann
1829	Aricine		Pelletier et Corriol
	Sanguinarine		Dana
1830	Curarine		Roulin et Boussingault
1831	Conicine		Geiger
	Atropine		[Brandes (1823)], Geiger et Hesse, Mein
1832	Codéine		Robiquet
	Narcéine		Pelletier
1833	Quinidine		Henry et Delondre
	Aconitine		Geiger et Hesse
	Colchicine		les mêmes
	Hyoscyamine		les mêmes
1835	Thébaïne		Pelletier et Thibouméry

Figure 1 : Liste des alcaloïdes découverts au début du XIXe siècle, selon Amé Pictet

A partir du XIXe siècle, la découverte de nouveaux alcaloïdes a pris une importance économique cruciale. Dès l'extraction de la quinine, Pelletier a initié sa production industrielle pour répondre en « quantité significative » à une grave épidémie de fièvre jaune à Barcelone. (9) Dans la seconde moitié du XXe siècle, les alcaloïdes se sont définitivement inscrits dans l'arsenal thérapeutique moderne, destiné notamment à la lutte contre le cancer. Parmi les réussites notables, nous pouvons citer les alcaloïdes de *Catharanthus* (*i.e.* vinblastine, vincristine). Aujourd'hui, l'intérêt grandit pour la découverte d'autres alcaloïdes possédant des propriétés antinéoplasiques ou antivirales. Le marché des médicaments à base d'alcaloïdes était estimé à 2,78 milliards de dollars en 2022 et devrait en effet passer à 5,65 milliards de dollars d'ici 2030. (10)

Les plantes renfermant des alcaloïdes forment un groupe extrêmement diversifié, tant sur le plan taxonomique (*i.e.* 10-15 % des plantes vasculaires sont alcaloïdofères), que chimique (*i.e.* > 6000 structures connues). (11) Le rôle physiologique de ces SN, leur importance évolutive, leur biogenèse, et surtout leur activité biologique sont souvent mieux définis pour chaque classe d'alcaloïde.

I.1.2.2. Rôle physiologique chez les plantes alcaloïdofères

Le rôle physiologique des alcaloïdes constitue le fondement de nombreuses hypothèses, car la variété de leurs structures ne laisse pas supposer qu'ils remplissent une fonction commune dans toutes les plantes qui les synthétisent, hormis leur caractère basique non spécifique. (11)

Les alcaloïdes se trouvent souvent dans les plantes en association avec des acides caractéristiques. Dans certains cas, les alcaloïdes pourraient servir de moyen de stockage ou de transport sous forme soluble d'un métabolite acide. Une balance entre protides et alcaloïdes a d'ailleurs été observée au sein de graines contenant des bases puriques, et permettrait leur production réciproque pendant la germination. (12) Cette hypothèse est néanmoins fragilisée car les alcaloïdes sont « *trop peu abondants pour justifier pareille mission d'approvisionnement* ». (11)

L'hypothèse qu'il n'existe pas de fonction commune aux alcaloïdes est consolidée par l'étude de leur biosynthèse. La majorité des alcaloïdes est en effet obtenue à partir d'unités facilement disponibles par le biais d'une série de réactions ubiquitaires. Ainsi, leur présence dans la plante pourrait être purement fortuite, et dépendre de la disponibilité des enzymes et des précurseurs. Les alcaloïdes, peu nocifs pour la plante, ne seraient donc pas éliminés par nécessité à travers la sélection naturelle. Les plantes qui normalement ne contiennent pas d'alcaloïdes semblent d'ailleurs ne pas souffrir de réactions indésirables lorsqu'elles sont exposées à des alcaloïdes : « *les feuilles et fruits de [la tomate (*Solanum lycopersicum* (L.))], anormalement chargés d'alcaloïdes toxiques par un processus analogue, se développent tout comme les autres* ». (13) L'absence d'alcaloïdes dans le greffon de plantes qui en accumulent normalement ne semble au contraire pas compromettre son développement. Ces deux cas de figure suggèrent que les alcaloïdes ne sont pas essentiels à la croissance de la plante. (11)

Néanmoins, il est pertinent de considérer que les alcaloïdes peuvent offrir une protection contre les prédateurs en raison de leur concentration dans les tissus périphériques. Les alcaloïdes sont pourtant pratiquement inoffensifs pour les insectes. Les alcaloïdes pourraient plutôt exercer une fonction protectrice en entravant le développement des plantules étrangères. (11) Enfin, les alcaloïdes pourraient intervenir face à d'autres menaces environnementales, comme l'oxygène singulet ($^1\text{O}_2$). Cette substance est nocive pour tous les organismes vivants et est produite dans les tissus végétaux en présence de lumière. Les plantes habitant des régions à forte intensité de lumière ultraviolette accumulent en effet davantage d'alcaloïdes, comme la berbérine chez les plantes du genre *Berberis*. (13)



I.1.2.3. Activités biologiques

Outre le rôle physiologique des alcaloïdes chez la plante, leurs activités biologiques à l'égard de l'Homme sont souvent marquées. C'est pourquoi les plantes alcaloïdofères sont rarement utilisées en phytothérapie, à cause de leur importante toxicité. Elles ont, au contraire, toujours eu une importance notable en médecine allopathique, où le dosage est strictement contrôlé. (13)

Les alcaloïdes ont la capacité d'interagir avec de nombreux systèmes physiologiques du corps humain, notamment le système nerveux central. Ces SN peuvent en l'occurrence provoquer un effet dépresseur, comme avec la morphine qui se lie aux récepteurs opioïdes mu et qui bloque également la libération de neurotransmetteurs excitateurs comme le glutamate et la substance P. A l'inverse, des alcaloïdes comme la caféine vont avoir un effet stimulant. La caféine est un antagoniste compétitif de l'adénosine ce qui inhibe ses effets sédatifs, et entraîne une augmentation de l'éveil et de l'attention. Au même titre, la strychnine est stimulante puisqu'il s'agit d'un inhibiteur compétitif du récepteur de la glycine et entraîne une augmentation excessive de l'excitabilité neuronale.

De plus, les alcaloïdes peuvent interagir avec le système nerveux autonome. Par exemple, l'éphédrine est un précieux hypertenseur car elle provoque des effets sympathomimétiques en se fixant sur les récepteurs adrénergiques α_1 et β_1 . Cela augmente la pression artérielle ainsi que la fréquence cardiaque. L'atropine est quant à elle un parasympatholytique et agit comme antagoniste compétitif des récepteurs muscariniques de l'acétylcholine. Cela provoque par conséquent une augmentation du rythme cardiaque, une diminution des sécrétions et une relaxation des muscles lisses. Les alcaloïdes peuvent au même titre provoquer des effets sympatholytiques et être utilisés comme hypotenseurs (*e.g.* yohimbine, alcaloïdes de l'ergot de seigle, *etc.*), ou des effets parasympathomimétique (*e.g.* pilocarpine, éserine, *etc.*).

D'autres grandes classes thérapeutiques tirent leurs origines d'alcaloïdes. Les premiers myorelaxants sont apparus après la découverte des curares (*e.g.* curarine, toxiferine). Ces molécules inhibent les récepteurs nicotiniques et provoquent une paralysie des muscles squelettiques. Pour continuer, les premiers anesthésiques locaux ont été découverts à partir de la cocaïne. Cette dernière agit en bloquant les canaux sodiques et interrompt la conduction de l'influx nerveux, ce qui engourdit temporairement la région ciblée. Par ailleurs, il n'est pas concevable d'introduire les alcaloïdes sans énoncer ceux issus des écorces du Quinquina rouge. Au côté de la cinchonine, la quinine est connue pour ses propriétés antipaludiques. Elle figure d'ailleurs sur la liste OMS des Médicaments Essentiel pour la prise en charge du paludisme grave. Les isomères de ces deux molécules, la cinchonidine et surtout la quinidine, sont employées pour leur part dans les traitements antiarythmiques.

Pour terminer, les alcaloïdes se distinguent par leur rôle significatif dans la thérapie antinéoplasique. Les alcaloïdes bis-indoliques, tels que la vinblastine, la vincristine et ses dérivés hémisynthétiques, la vinorelbine, la vindésine et la vinflunine, sont aujourd'hui utilisés dans le traitement de divers cancers, notamment les cancers du sein métastatiques, de l'ovaire, de la vessie et des bronches. Ces molécules agissent comme des poisons du fuseau en inhibant la polymérisation de la tubuline. Cela bloque le processus de mitose et entraîne la mort cellulaire.

Ainsi, les alcaloïdes n'ont pas un récepteur spécifique unique. Au contraire, ils interagissent avec divers récepteurs en fonction de leur structure chimique (*i.e.* pharmacophore), qui est déterminée par leur voie de biosynthèse.

I.1.2.4. Biosynthèse

La diversité structurale des alcaloïdes a donné lieu à plusieurs systèmes de classification. Parmi ces systèmes, la classification basée sur leurs précurseurs est la plus largement reconnue. (*c.f.* Fig.2). Seules les voies de biosynthèses conduisant aux alcaloïdes mono et bis-indoliques seront développées.

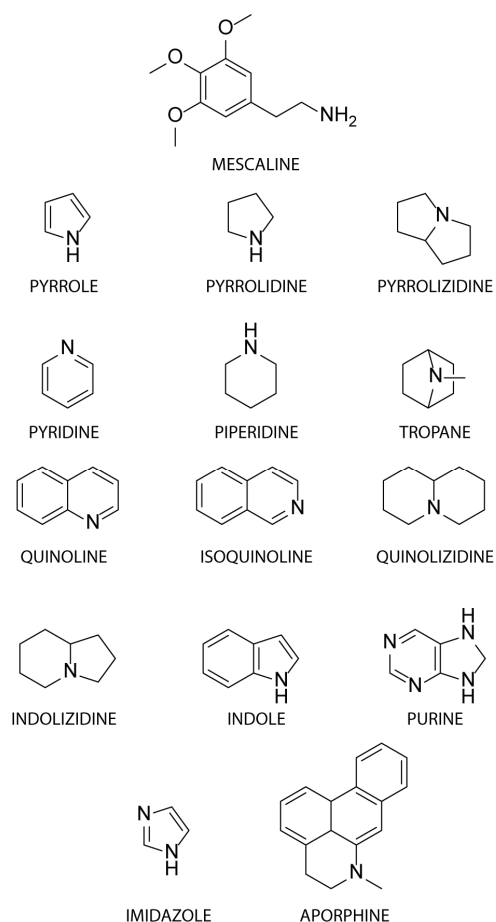


Figure 2 : Noyaux des principaux acides aminés précurseurs d'alcaloïdes



I.1.2.4.a. Biosynthèse des alcaloïdes mono-indoliques

Les alcaloïdes mono-indoliques contiennent généralement deux atomes d'azote, dont l'un est l'azote indolique et l'autre est éloigné de deux carbones de la position bêta du cycle indolique. Deux groupes distincts peuvent être formés selon le type de condensation entre la tryptamine (*i.e.* le noyau indolique) et un aldéhyde ou une cétone. La réaction de Mannich impliquant le carbone alpha du noyau indole produit un dérivé bêta-carboline, tandis que celle impliquant le carbone bêta donne un dérivé indolénine (*c.f.* Fig.3).

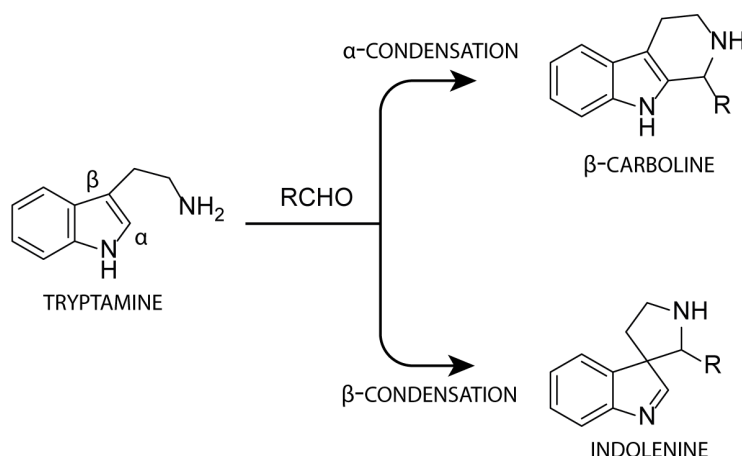


Figure 3 : Réaction de condensation se produisant entre la tryptamine et un aldéhyde

Ainsi, le tryptophane et son produit de décarboxylation, la tryptamine, sont le plus souvent à l'origine de la grande classe des alcaloïdes indoliques. Par ailleurs, des alcaloïdes mono-indoliques plus complexes (*i.e.* alcaloïdes indoliques monoterpéniques) contiennent une partie de la molécule non dérivée du tryptophane. Cette sous-structure est un motif le plus souvent constitué de 10 carbones divisibles en une unité linéaire de 6 carbones (6 C), une unité de carbone (1 C) et une autre unité de 3 carbones (3 C) (*c.f. infra*). Une classification des alcaloïdes indolo-monoterpéniques inspirée de leur origine biosynthétique a donc été proposée (*c.f.* Fig. 4). (14)

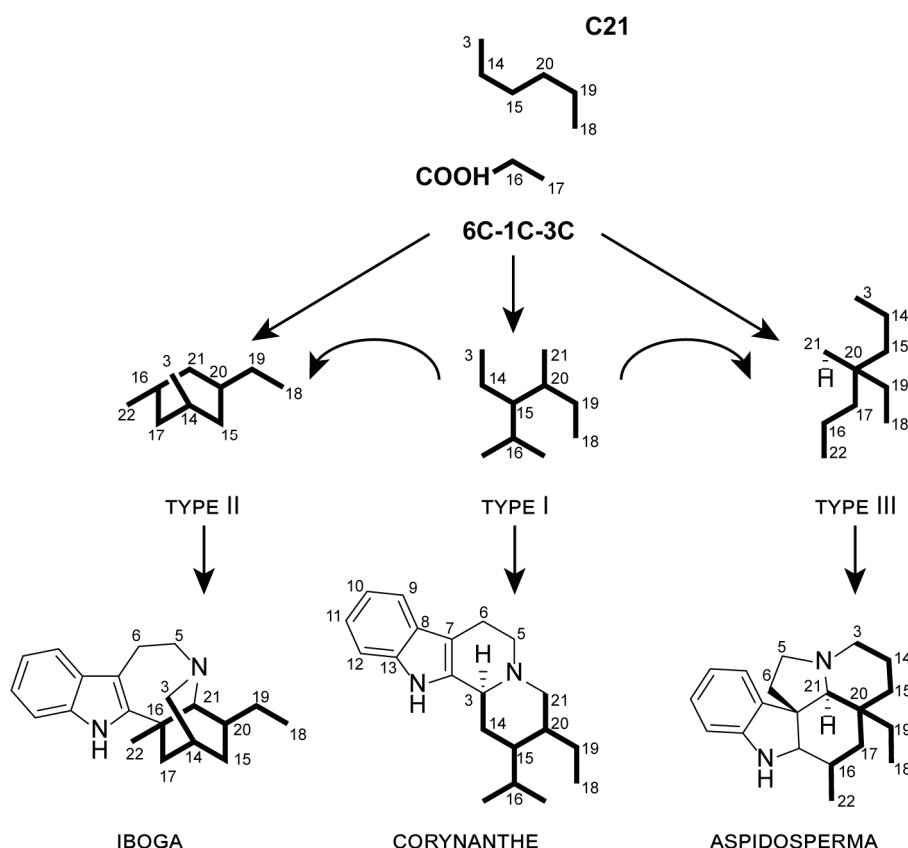


Figure 4 : Numérotation biosynthétique des alcaloïdes mono-indoliques, selon LeMen et Taylor (15)

La partie de la molécule non dérivée du tryptophane est issue de l'acide mévalonique. Il s'agit d'un apport de géraniol en C-10 (*i.e.* monoterpénoïde) dans le cas des alcaloïdes provenant des Apocynaceae (*c.f.* Fig.5).

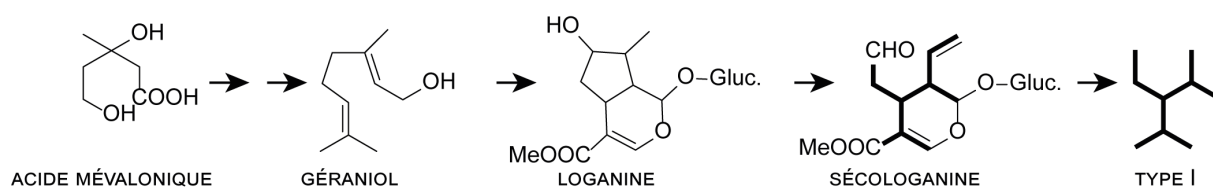


Figure 5 : Apport de la partie terpénique dans le cas des alcaloïdes mono-indoliques issus des plantes de la famille taxonomique des Apocynaceae

La 3- α (S)-strictosidine, (*c.f.* Fig. 6) a été identifiée comme l'intermédiaire clé dans la biogenèse des alcaloïdes indoliques monoterpéniques. Cet intermédiaire est formé par la condensation enzymatique (*i.e.* strictosidine synthase) de la tryptamine et de la sécologanine dérivée du loganoside.

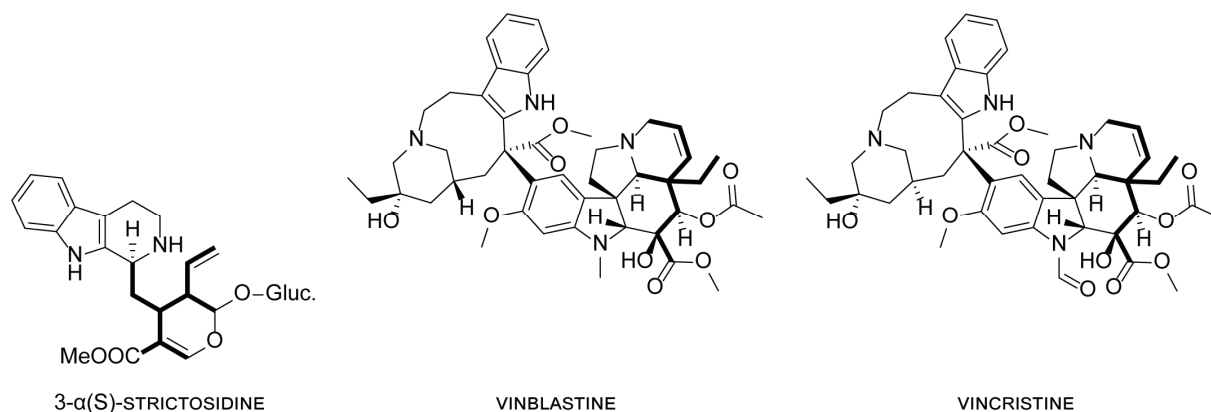


Figure 6 : Structures moléculaires de la 3-α(S)-strictosidine, et des alcaloïdes bis-indoliques vinblastine et vincristine issus de *Catharanthus roseus* (L.) G.Don

I.1.2.4.b. L'étude des alcaloïdes provenant des plantes du genre *Catharantus* : à la découverte des alcaloïdes bis-indoliques

La pervenche de Madagascar, *Catharanthus roseus* (L.) G.Don (syn : *Vinca rosea* (L.), *Lochnera rosea* (L.) K.Schum), est une plante appartenant à la famille des Apocynaceae (c.f. Fig.7). Originnaire de Madagascar, elle est maintenant largement répandue dans les régions où elle est utilisée en tant que plante ornementale.

La pervenche de Madagascar est un sous-arbrisseau herbacé, de 40 à 80 cm de haut, devenant ligneux à la base. Les feuilles sont disposées de manière opposée, oblongues avec une base aiguë pétiolée, un apex arrondi ou mucroné et une marge entière. Les fleurs ressemblent à celles de la grande pervenche (*Vinca major* (L.),) et sont de couleur violette, rose, blanche ou blanche avec un œil rouge. Le fruit est un follicule différencié.



Figure 7 : Photo de *Catharanthus roseus* (L.) G. Don (selon Kenriez au Parque Botánico de Maspalomas (Gran Canaria)) et planche d'herbier du MNHN

La pervenche de Madagascar était autrefois utilisée comme coupe faim par les navigateurs, ce qui a conduit à sa dissémination. La plante était également employée en médecine traditionnelle, notamment en Jamaïque, pour le traitement du diabète. Néanmoins, ces dernières propriétés n'ont jamais été prouvées. En revanche, des extraits de feuilles provoquaient des effets leucopéniques chez les rats. Ainsi, 1 g de feuilles sèches provoquait en une seule injection intrapéritonéale la mort d'un rat de 100 g avec des effets caractéristiques sur les globules blancs. Des études phytochimiques intensives de la plante ont par conséquent été entreprises en vue d'isoler des constituants utiles dans le traitement contre le cancer. (16,17)

Plus de 150 alcaloïdes ont été ainsi isolés à partir des plantes de la famille des Apocynaceae. La vincamine fut isolée pour la première fois à partir de *Vinca minor* en 1953. Elle a d'abord fait l'objet de recherches comme médicament vasodilatateur. Puis, l'intérêt s'est rapidement réorienté vers un groupe de 20 alcaloïdes bis-indoliques issus du genre *Catharanthus*. En 1958, la vinblastine fut en effet découverte à partir de *Catharanthus roseus* (L.) G. Don, puis utilisée en chimiothérapie dès 1961 sous le nom Velbé® pour soigner la maladie de Hodgkin. Peu de temps après, la vincristine (Oncavin®) ainsi que la leurosidine (*i.e.* épimère en C-20' de la vinblastine) ont été isolées. La vincristine est désormais utilisée dans le traitement des leucémies aiguës lymphoblastiques, des lymphomes, du cancer du poumon à petites cellules, du cancer du col de l'utérus et du cancer du sein. La vincristine a une activité antitumorale supérieure à celle de la vinblastine, mais elle est plus neurotoxique. (17)

La vinblastine est tous les autres alcaloïdes de ce type sont notamment constitués de deux parties : une unité indolique tétracyclique A correspondant à la catharantine (*syn* : 16-méthoxycarbonyl-velbamine), et une unité dihydro-indolique B, la vindoline, reliée à la partie A par une liaison C-10-C-16' (c.f. Fig.8).

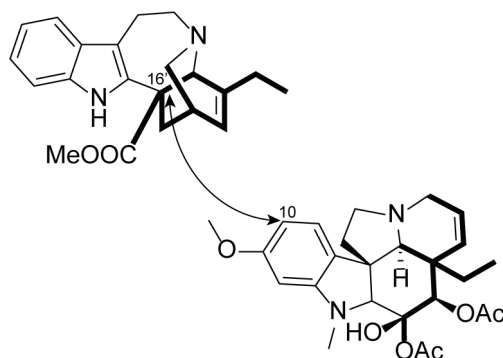


Figure 8 : Couplage de l'unité indolique tétracyclique (16-méthoxyvelbamine) et d'une unité dihydroindolique (vindoline) donnant accès aux alcaloïdes de type vinblastine

Toutefois, la très grande complexité des extraits alcaloïdiques du genre *Catharanthus* ne permettait pas d'isoler la vinblastine et la vincristine en grande quantité. Les alcaloïdes bis-indoliques ne représentent en effet que 1% du poids sec de la plante, et la quantité de vincristine est 8 à 10 fois moindre que celle de la vinblastine.

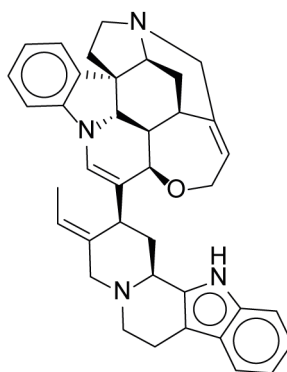
I.1.2.4.c. L'alcaloïdothèque de l'équipe CSN de l'ICMR

Plusieurs pionniers de la Faculté de Pharmacie de Paris (*i.e.* R. Paris, M.-M. Janot, ou encore J. LeMen) se sont intéressés aux alcaloïdes de *Catharanthus roseus* (L.) G.Don avant même la découverte de la vinblastine et vincristine. Pour autant, les propriétés biologiques de ces deux molécules ont stimulé la recherche de nouveaux composés et dérivés à partir d'autres espèces du genre *Catharantus*, notamment à l'Institut de Chimie des Substances naturelles du CNRS (ICSN) de Gif-sur-Yvette.

La partie vindoline de la vinblastine présente plusieurs sites favorables à la fonctionnalisation de la molécule, permettant de moduler son efficacité et sa toxicité : le noyau indolique, le groupement hydroxyle et la liaison ester. (17) Parmi ces réactions, les réactions d'hydrogénation transforment la vinblastine en 14,15-dihydro-vinblastine. Ces modifications ont été réalisées par M.-M. Janot et J. Lemen à la Faculté de Pharmacie de Paris et réduisent la toxicité de la molécule. (17)

De nombreuses tentatives ont été menées pour obtenir des dérivés bis-indoliques à partir de la vindoline (*i.e.* précurseur présent en grande quantité dans la plante). Ainsi, M.-M. Janot et J. LeMen tentaient en 1928 l'hémisynthèse de la vinblastine. Mais leur méthode de couplage C-10-C-16', ne permettait pas d'obtenir des alcaloïdes bis-indoliques de configuration naturelle en C-16' et étaient donc dépourvus d'activité antitumorale. De plus, aucun des composés utilisés comme précurseur de l'unité indolique tétracyclique dans les tentatives de couplage, n'a été identifié parmi les très nombreux alcaloïdes isolés du genre *Catharanthus*. Ainsi, J. LeMen et J. Levy décidèrent, une fois à Reims, d'entreprendre la synthèse biomimétique de la vinblastine jugeant que « *les procédés d'hémisynthèse développés jusqu'alors* [voie de l'alpha-hydroxy-indole, voie de la chloro-indolénine entre autres] *étaient éloignés vraisemblablement de la voie naturelle* ». (17) C'est en 1974 que P. Potier, ancien doctorant de J. LeMen à la faculté de Pharmacie de Paris, et ancien directeur de l'ICSN, réalise le couplage de la vindoline et de la catharantine (*i.e.* unité pentacyclique), par rupture de la liaison C-16-C-21 au moyen de la réaction de Polonovski-Potier, donnant accès aux dérivés nor-vinca alcaloïdes tels que la vinorelbine.

Entre temps, J. LeMen, poursuit ses recherches sur d'autres espèces du genre *Catharanthus* (*Vinca minor* (L.), *Vinca difformis* (Pourr.), *Catharanthus longifolius* (Pichon), *Catharanthus trichophyllus* (Baker, Pichon), *Catharanthus lanceus* (Bojer ex A.DC., Pichon) ainsi que sur les espèces du genre *Strychnos* (e.g. *Strychnos nux-vomica* (L.)), partagé entre les laboratoires de Paris (ICSN) et de Reims (ICMR) dans les années 1960-1970s. Puis, dans les années 1980s et sous l'impulsion de L. LeMen-Olivier et G. Massiot, une vingtaine de nouveaux alcaloïdes ont été identifiés lors du profilage chimique des espèces *Strychnos longicaudata* (Gilg) et *Strychnos ngouniensis* (Pellegr.), dont la nouvelle série de bis-indoles représentée par la longicaudatine (*c.f.* Fig.9). (18) Ces molécules ont été parmi les dernières à être ajoutées à l'alcaloïdothèque de l'ICMR. Cette dernière témoigne de la richesse de cette période propice à la recherche en substances naturelles.



LONGICAUDATINE

Figure 9 : Structures moléculaires de la longicaudatine (alcaloïde bis-indolique) issue de *Strychnos longicaudata*

I.2.Profilage chimique

I.2.1. Définitions et généralités

La structure d'une molécule est fondamentale et détermine toutes ses propriétés (*e.g.* masse, lipophilie, *etc.*) Il existe en revanche un grand nombre d'autres caractéristiques pouvant décrire une molécule : nom vernaculaire ou officiel, sa formule brute, ses propriétés pharmaceutiques, *etc.* Des éléments comme son profil spectral (*c.f.* I.4.3. Rôle des bases de données dans le processus de déréplication) contribuent notamment à faciliter son identification/annotation au sein d'un mélange.

Par ailleurs, la distribution qualitative et quantitative de ces structures varie selon les espèces, les genres ou les familles taxonomiques. La chimiotaxonomie permet ainsi de regrouper les organismes en fonction de leurs similarités biochimiques. Elle complète la taxonomie basée sur la classification morphologique et génétique en taxons (*i.e.* groupes hiérarchisés d'organismes qui partagent des caractéristiques communes).

L'étude des métabolites (*i.e.* la métabolomique) s'inscrit dans une approche générale appelée « omique ». La définition qu'en donne l'INSERM est la suivante : « *[elle] consiste à appréhender la complexité du vivant dans son ensemble, au moyen de méthodologies les moins restrictives possibles sur le plan descriptif. Ces approches peuvent en particulier être utiles pour mettre en évidence et identifier de nouveaux biomarqueurs (d'exposition, d'effet, ou de susceptibilité), générer de nouvelles connaissances sur le plan mécanistique (mode d'action), ou encore élaborer de nouveaux outils de toxicologie prédictive pour aider à l'identification de dangers* ». (19) Ainsi, l'approche « omique » vise à comprendre un organisme dans son ensemble de manière mécanistique. Elle inclue des axes de recherches tels que la génomique (*i.e.* analyse du code génétique), la transcriptomique (*i.e.* expression et régulation des gènes), la protéomique (*i.e.* analyse des protéines) et la métabolomique (*i.e.* étude des métabolites). Plus précisément, la métabolomique a pour objectif d'expliquer la manière dont interagissent les métabolites entre eux, tout en contribuant à la compréhension du vivant dans son ensemble. Cependant, aucun outil analytique ne peut atteindre à lui seul cet objectif.

Le profilage chimique fait partie des études métabolomiques, puisqu'il permet d'obtenir une cartographie chimique des chémotypes (*i.e.* familles chimiques ou métabolites) présents au sein d'un organisme vivant ou d'un extrait à un moment donné, sans pour autant chercher à comprendre les interactions qu'il y a entre eux. La démarche la plus adoptée par les phytochimistes implique l'isolement de molécules (*i.e.* la séparation d'une combinaison chimique à partir d'un mélange complexe). (20,21) L'étape d'extraction (*i.e.* la séparation d'une substance d'une matrice dont elle fait partie (22)) peut quant à elle être plus ou moins sélective, et être menée sur une seule partie de l'organisme. Cette sélectivité est supposée simplifier l'étape de purification et d'isolement des

composés. Ce flux de travail nécessite cependant un investissement considérable, tant sur le plan technique que financier et s'explique par l'importante complexité des matrices naturelles. (23) Un échantillon naturel est estimé contenir des dizaines de composés majoritaires, et des centaines voire des milliers de minoritaires. (23) La cartographie de ces métabolites nécessite donc des méthodes analytiques pouvant à la fois travailler sur une large gamme dynamique et fournir assez d'informations spectrales pour procéder à leur identification ou annotation. (23,24)

Aujourd'hui une étape de déréplication est le plus souvent intégrée au profilage chimique d'un échantillon. (23) La déréplication a été définie pour la première fois en 1990 comme étant « *l'identification rapide au préalable de chémotypes déjà connus au sein des matrices étudiées* ». (25) Elle se base en général sur la comparaison de données analytiques à d'autres données analytiques de référence, présentes dans des bases de données (BD). Ainsi, la déréplication est un travail de mémoire collective.

I.2.2. Acquisition des données spectrales expérimentales nécessaires à la déréplication

Le profilage chimique d'un échantillon, intégrant une étape de déréplication, nécessite l'utilisation de techniques analytiques par spectroscopie et/ou par spectrométrie. En effet, les molécules analysées sont annotées / identifiées par le traitement de leur signature spectrale. Pour des raisons pratiques, seules les principales techniques d'acquisition (*i.e.* RMN et MS) (*c.f.* Fig.10) ainsi que la façon d'analyser et de visualiser les spectres MS/MS seront décrites dans cette partie.

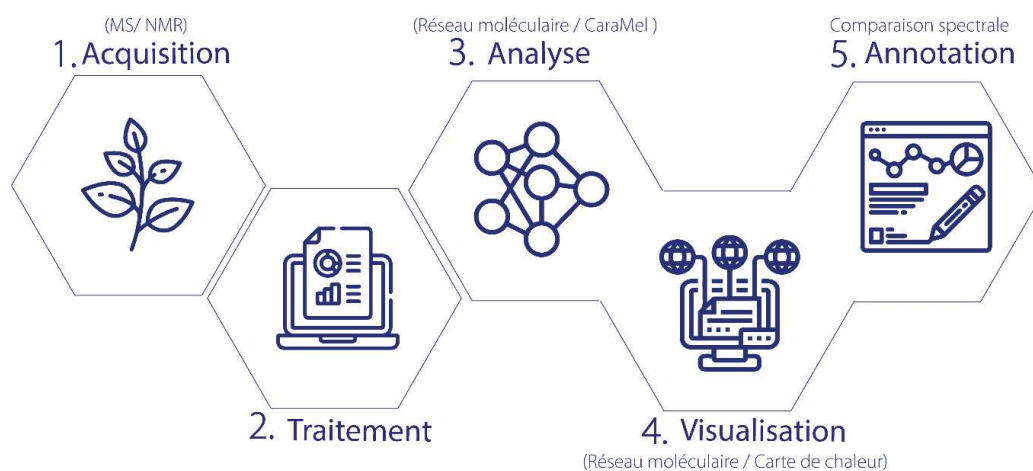


Figure 10 : Flux de travail en déréplication

Les techniques analytiques permettant de produire un spectre peuvent être regroupées en deux groupes : la spectroscopie, et la spectrométrie. La spectroscopie concerne l'ensemble des techniques de caractérisation physico-chimiques qui reposent sur l'interaction entre un rayonnement électromagnétique et la matière. Des transitions d'énergie entre états électroniques, vibrationnels ou rotationnels d'une molécule sont le plus souvent observées. La résonance magnétique nucléaire (RMN), la spectroscopie UV, IR, sont les principales utilisées en chimie des substances naturelles. (21) La spectrométrie de masse (MS) est au contraire une technique qui analyse la molécule au travers du modèle de fragmentation d'ions, plutôt qu'en mesurant directement l'interaction de la molécule avec le rayonnement électromagnétique.

La spectrométrie de masse à haute résolution couplée à une méthode chromatographique (e.g. la LC-HRMS²) a considérablement gagné en intérêt grâce à l'apparition des outils chémo-informatiques de traitements de données, inspirés de la métabolomique (e.g. les réseaux moléculaires - RM). (26,27) En parallèle, la RMN est la technique analytique incontournable pour l'élucidation des structures moléculaires de SN. (21) La Résonance Magnétique Nucléaire est une méthode de spectrométrie d'absorption, car elle mesure l'absorption d'énergie par les noyaux atomiques, procurant ainsi des indications précises concernant leur environnement. Elle est spécialement employée pour déterminer la structure chimique des molécules organiques. La RMN présente une caractéristique non invasive, puisque chaque échantillon analysé peut être récupéré sans altération. Elle autorise aussi la quantification des métabolites présents dans l'échantillon. De plus, la RMN démontre sa robustesse et sa reproductibilité au fil du temps et entre différents laboratoires. (28) Cette technique peut *a contrario* souffrir d'un manque de résolution (i.e. recouvrement de signal pour les analyses du proton (¹H)) et d'un manque de sensibilité (surtout pour la RMN du ¹³C), bien que des améliorations aient été apportées ces dernières années (e.g. sondes cryogéniques, plus hauts champs, etc.). La spectrométrie de masse a quant à elle la capacité de déterminer la masse des molécules présentes dans l'échantillon. La MS est une technique destructive, mais elle nécessite seulement de petites quantités d'échantillon. Sa sensibilité élevée en fait la méthode la plus prédominante pour les investigations métabolomiques, bien que sa capacité à supporter sans conséquence de petites variations des paramètres internes à la méthode est moindre par rapport à la RMN. C'est pourquoi il est essentiel d'associer aux données spectrales leurs paramètres d'acquisition. (29,30)

Les données spectrales fournies par ces deux techniques complémentaires sont aujourd'hui souvent suffisantes pour identifier un composé dont la structure est déjà connue (i.e. spectre renseigné dans les bases de données) ou *de novo*. D'ailleurs, les quantités de composés mineurs isolés par les chimistes au cours du premier quart du siècle dernier, bien qu'insuffisantes à l'époque pour des

investigations plus poussées, seraient désormais largement suffisantes pour une analyse complète de leur structure *via* l'utilisation des méthodes et instruments analytiques modernes ici présentés. (31)

1.2.2.1. Acquisition par RMN

La RMN repose sur l'étude des moments magnétiques des noyaux (*c.f.* Fig. 11). La charge de certains noyaux atomiques tourne autour de l'axe nucléaire. Ce mouvement de rotation génère un dipôle magnétique le long de cet axe. Le moment angulaire de cette charge tournante peut être décrit en termes de nombre de spin (I). Un noyau ayant un spin différent de 0 se comporte comme un petit aimant de moment magnétique μ . Seuls ces noyaux peuvent être observés en RMN. Deux niveaux d'énergie différents (E_1 et E_2) sont générés lorsqu'est appliqué un champ magnétique extérieur (B_0). L'effet Zeeman correspond à la levée de dégénérescence énergétique, et la transition entre E_1 et E_2 constitue le phénomène de RMN. La somme des moments magnétiques individuels est égale à l'aimantation macroscopique (M_0). Dès que B_0 est appliqué, les moments magnétiques μ interagissent avec lui et se mettent à tourner autour de la direction de B_0 comme une toupie. Ce mouvement de précession est appelé la précession de Larmor et sa fréquence est propre à chaque noyau. Une analyse RMN consiste à faire passer le vecteur moment magnétique $\vec{M}$ d'un état magnétique stable $\vec{M0}$ parallèle à $\vec{B0}$ à un état magnétique moins stable $\vec{M1}$ est observé le retour de M à son état magnétique stable. Pour cela, il faut appliquer un champ magnétique $\vec{B1}$ pendant une fraction de seconde, et tournant à la fréquence de Larmor de l'atome étudié. Dès la disparition de ce second champ, $\vec{M}$ évolue jusqu'à retourner à son état d'équilibre $\vec{M0}$. C'est le phénomène de relaxation. Une variation rectiligne sinusoïdale décroissante de la valeur de M_0 est obtenue en fonction du temps, et est appelé Free Induction Delay (FID). La transformée de Fourier permet de traiter ce signal et d'obtenir un spectre RMN conventionnel sur une échelle de fréquence. Un spectre RMN représente des signaux de résonance dont les positions correspondent à des déplacements chimiques, et leurs intensités correspondent à l'intégration du signal.(32) Un spectromètre RMN est ainsi constitué d'un aimant qui fournit un champs magnétique homogène, d'une sonde qui sert à l'excitation et à la détection du signal, et d'une console qui permet de contrôler l'excitation ainsi que l'amplification du signal retour et sa conversion. Un ordinateur permet enfin de recevoir les informations, d'assurer les transformées de Fourier, et visualiser les spectres RMN.



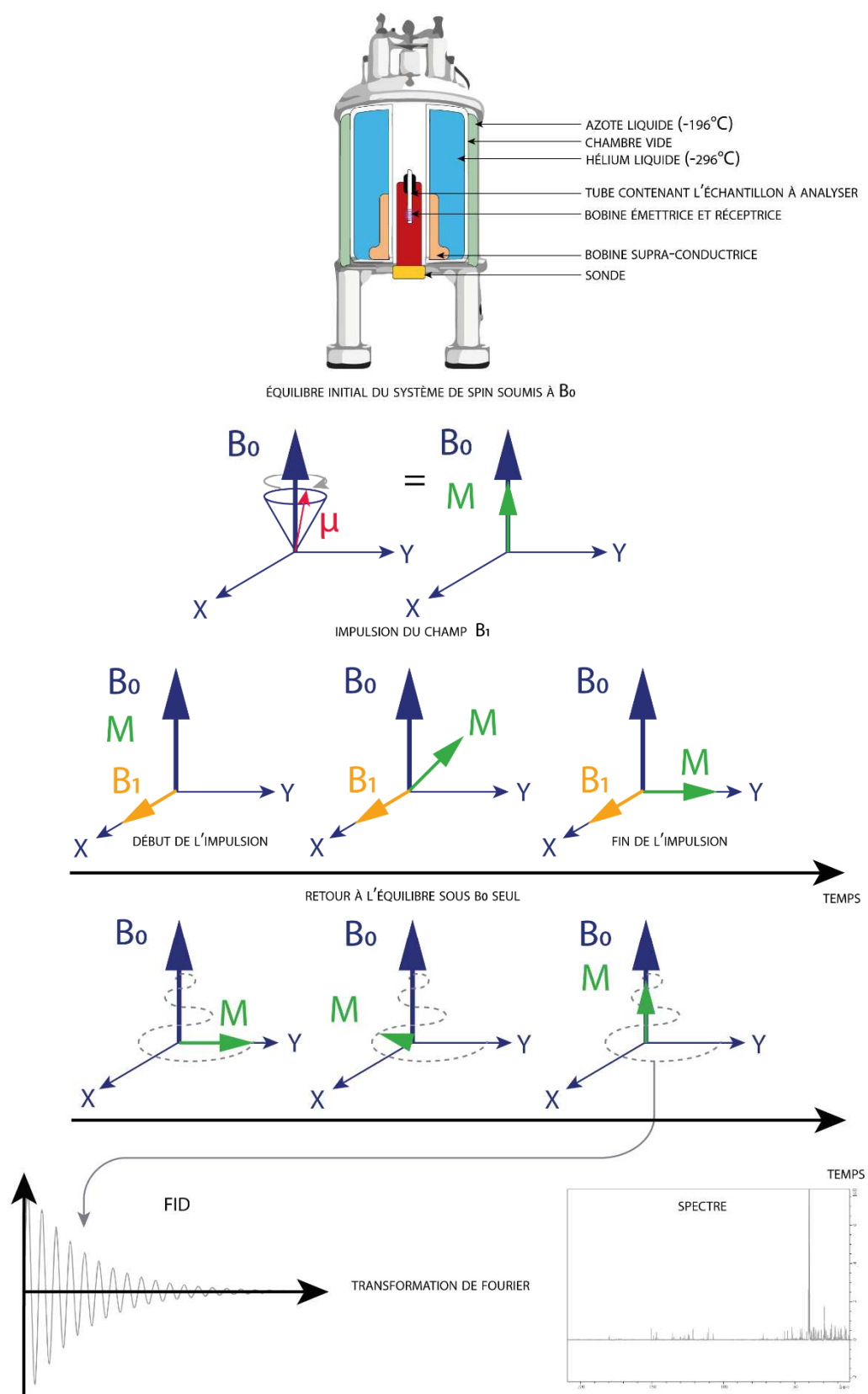


Figure 11 : Schéma explicatif du fonctionnement d'un spectromètre de résonance magnétique (RMN), selon(32)

I.2.2.2. Acquisition par MS

La spectrométrie de masse permet de « peser » les molécules, de corréler les spectres de fragmentation avec les structures moléculaires, et d'expliquer les mécanismes de rupture de ses liaisons. Elle s'appuie sur le déplacement de particules chargées dans des champs électromagnétiques (c.f. Fig.12). Pour cela, les métabolites doivent au préalable passer à l'état d'ions en phase gazeuse. Ces ions sont ensuite séparés en fonction de leur rapport masse sur charge (m/z). Un spectromètre de masse se compose d'une source qui sert à produire les ions à l'état gazeux à partir de l'échantillon à analyser. L'appareil est soumis au vide après la source, pour guider et accélérer les ions (énergie cinétique), et les focaliser vers l'analyseur grâce à un champ électrique *via* un jeu de lentilles, tout en évitant les phénomènes de collision. Cela amène à la notion de libre parcours moyen : la distance moyenne que parcourt une particule entre deux collisions. Elle dépend de la vitesse, de la taille des particules, et de la pression environnante. Ce libre parcours moyen doit être égal à 10 fois la distance réellement parcourue par la particule dans l'appareil d'analyse. Puis, l'analyseur filtre et trie les ions produits par la source selon leur rapport m/z . Le détecteur va permettre de compter les ions préalablement triés et d'envoyer le signal à l'enregistreur. Enfin, un ordinateur doté d'un logiciel d'acquisition permet d'enregistrer et de visualiser les spectres de masse et de traiter les données. Un spectre de masse est constitué de raies dont la position sur le spectre dépend du rapport m/z et la hauteur dépend de l'abondance de ces ions. Lors du traitement des données, les signaux enregistrés peuvent être visualisés en fonction de la durée totale de l'analyse, sous forme d'un courant d'ions total (Total Ion Current – TIC). Il correspond à la somme de l'intensité de tous les ions présents sur chaque spectre de masse acquis. L'abondance d'un ion de rapport m/z donné peut être extrait et visualisé. Il s'agit d'un courant d'ion extrait (Extracted Ion Current – EIC). L'intensité de chaque ion est proportionnelle à sa capacité à s'ioniser, et dans une moindre mesure à la quantité du composé dans l'échantillon. (32)

Plusieurs sources d'ionisation en spectrométrie de masse sont aujourd'hui disponibles (Impact Electronique (IE), Ionisation Chimique (IC), ou encore la Désorption Laser Assistée par Matrice (MALDI), mais la plus utilisée en chimie des substances naturelles est l'Ionisation par Electro Spray (ESI). (23) Il s'agit en effet d'une source à pression atmosphérique, tout comme l'APCI, à pouvoir être couplée à l'HPLC. L'ionisation s'y fait à température ambiante. Cette dernière repose sur l'introduction d'une solution de l'échantillon à analyser par un capillaire métallique très fin, porté à un haut potentiel électrique (2000 – 5000 Volts). Cette tension crée des charges dans la solution, et au sortir du capillaire un spray de fines gouttelettes est formé, favorisé par une nébulisation pneumatique. Dans la source à pression atmosphérique, ces gouttelettes rencontrent un flux d'azote chaud à contrecourant, ce qui



participe à l'évaporation progressive du solvant. La taille des gouttelettes diminue et la densité de charge au sein de la solution augmente. Lorsque la répulsion coulombienne excède la tension superficielle de la gouttelette, celle-ci explose en formant d'autres gouttelettes de deuxième génération, beaucoup plus petites. Après plusieurs étapes de désintégration, la densité de charge dans la gouttelette devient telle que le champ électrique local, très intense, conduit à la désorption par effet de champ. Des ions sont alors formés en phase gazeuse, physiquement intacts, et pouvant porter plusieurs charges en raison de l'attachement de protons aux sites basiques de la molécule. Un ensemble d'ion de masse, voire de charges différentes est produits dans la source, quel que soit le type d'ionisation utilisé.

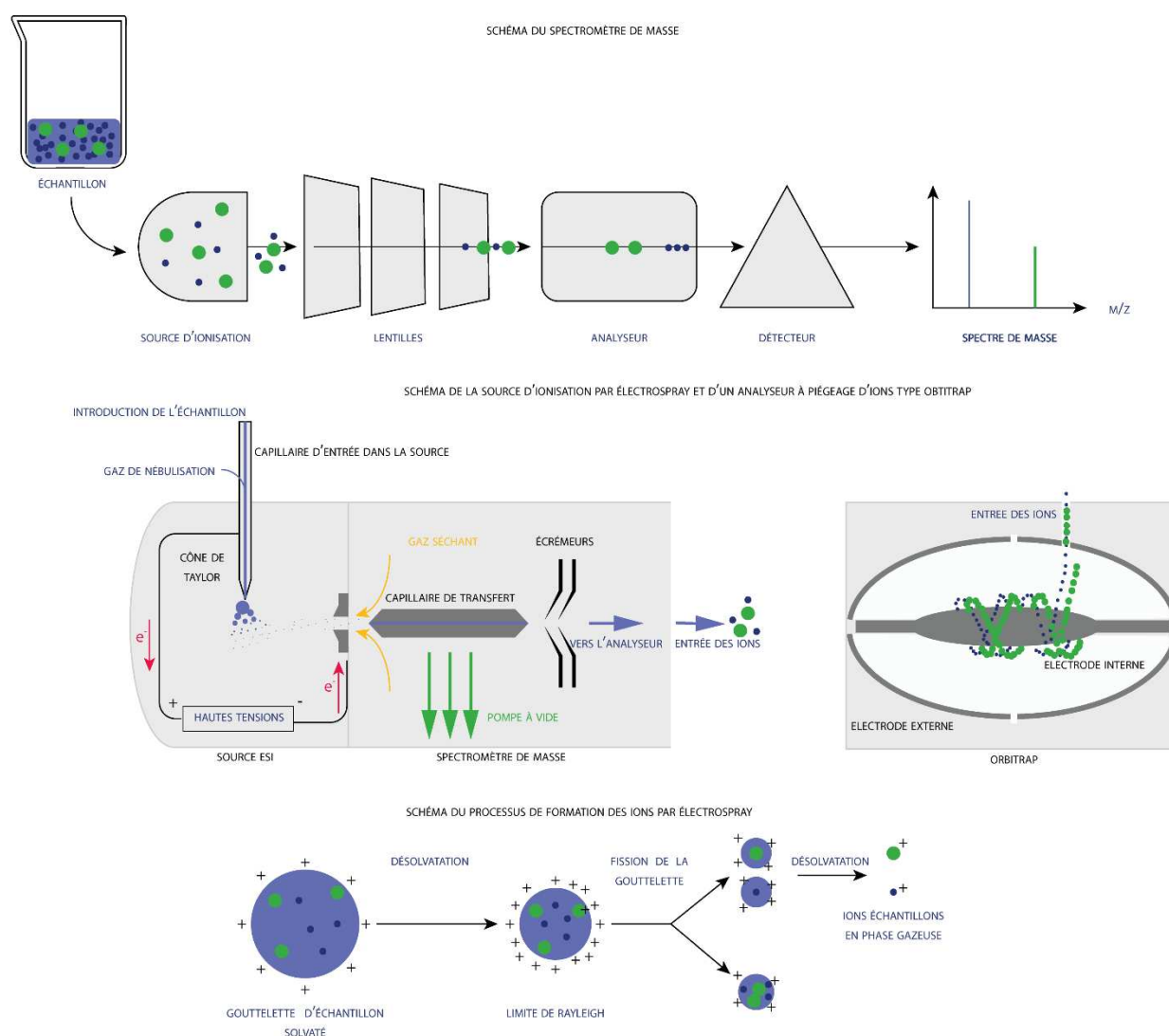


Figure 12 : Schéma explicatif du fonctionnement d'un spectromètre de masse (ESI-Orbitrap), selon (32)

L'analyse du faisceau ionique consiste ensuite à séparer les ions les uns des autres en fonction de leur rapport m/z . Cette information découle d'un phénomène unique : la déviation de ces

ions par des champs électromagnétiques. Il est essentiel de trier les ions selon leur rapport m/z , puisque plusieurs éléments de m/z différents ne peuvent être pesés en même temps. Trois analyseurs sont majoritairement utilisés : le quadripôle, le ToF et l'Orbitrap.

Un quadripôle est constitué de 4 électrodes cylindriques parallèles et raccordées deux à deux. Une tension continue U et une alternative V sont alors appliquées. Les électrodes sont ainsi chargées deux à deux positivement ou négativement à tour de rôle. La trajectoire des ions pour un m/z entre ces 4 électrodes dépend de la fréquence de la tension alternative appliquée. Ainsi, une trajectoire stable est obtenue avec une fréquence appropriée, permettant aux ions de parcourir le quadripôle depuis son entrée jusqu'à sa sortie sans collision. La masse des composés est alors calculée en fonction de U et de V . En balayant rapidement les fréquences, il est possible d'obtenir le spectre de masse pour une gamme de m/z donnée.

Dans l'analyseur à Temps de Vol (Time of Flight – ToF), la relation entre le m/z et la vitesse des ions est utilisée. Les ions sont envoyés par paquet au travers de grilles d'extraction et d'accélération. Les ions ayant un faible m/z seront plus rapides à atteindre le détecteur contrairement aux ions de plus grand rapport.

Le fonctionnement d'un Orbitrap repose quant à lui sur les principes de la dynamique des particules chargées et de la résonance cyclotronique ionique. L'élément central de l'Orbitrap est une électrode centrale cylindrique autour de laquelle se trouve une électrode externe en forme de baril (c.f. Fig.12). Lorsque les ions pénètrent dans le champ électrique créé entre ces deux électrodes, ils sont capturés et se déplacent en spirale autour de l'électrode centrale et oscillent aussi axialement. La fréquence d'oscillation dépend directement du rapport masse/charge (m/z) des ions. Les ions plus lourds se déplacent plus lentement tandis que les ions plus légers oscillent plus rapidement. Le courant électrique induit par ces oscillations est mesuré et converti en un spectre de masse grâce à la transformation de Fourier. Les avantages de l'Orbitrap vis-à-vis du ToF sont une meilleure résolution, ainsi qu'une meilleure sensibilité.

Les premières générations de spectromètres de masse ne permettaient d'étudier que la dimension MS^1 , c'est-à-dire les ions parents. Les études structurales se limitaient alors aux fragments issus de la fragmentation en source de l'ion moléculaire. Cependant, la fragmentation de ces ions, en plus de sa formule brute, aide à l'identification / annotation des petites molécules. (23) C'est pourquoi la spectrométrie de masse en tandem (MS/MS) est une des technique principale pour l'étude des SN. (33) Le fonctionnement de la MS/MS consiste d'abord à isoler l'ion précurseur, puis à l'exciter énergétiquement par collision avec un gaz inerte cible, le faisant passer à un état excité. Ce surplus d'énergie peut être stocké par l'ion sous forme de phénomène de rotations, de vibrations, ou



changement de niveau électronique. Ensuite, deux processus peuvent se produire, selon la pression du gaz cible : la dissociation unimoléculaire de l'ion précurseur ou sa désactivation par transfert d'énergie lors d'une nouvelle collision avec le gaz inerte. Lors de la phase d'activation par collision, il se produit en effet une redistribution rapide de l'énergie interne entre les différents modes vibrationnels de l'ion moléculaire intact par phénomène de résonance. Ce processus se poursuit jusqu'à ce que l'ion précurseur excité atteigne une énergie suffisante et une configuration géométrique propice, permettant, selon le modèle de Rice, Ramsperger et Kassel, la dissociation de l'ion précurseur.(33) Lorsque des ions précurseurs sont activés (*i.e.* changement d'énergie interne de l'ion précurseur) par des collisions simples ou multiples avec un gaz inerte, on parle de dissociation induite par collision (CID). La quantité d'énergie cinétique nécessaire pour provoquer une CID dépend de la masse des partenaires impliqués. Cette énergie est plus importante lorsque la masse du gaz inerte est élevée et que celle de l'ion précurseur est faible. C'est pourquoi des gaz comme le xénon (Xe, 131,293 u) ou l'argon (Ar, 39,948 u) sont couramment utilisés en CID. Toutefois, ces collisions ne sont pas parfaitement élastiques, et une partie de l'énergie apportée est dissipée sous forme de vibrations au sein de la molécule. Par conséquent, l'élan fourni par l'énergie cinétique au système n'est pas entièrement conservé lors d'une collision. Ainsi, l'énergie apportée doit être supérieure au seuil de dissociation de l'ion précurseur pour compenser ces pertes. L'énergie cinétique convertie en énergie intramoléculaire est alors appelée énergie d'endothermicité. (33)

La génération de fragments à partir d'ions précurseurs peut être réalisée physiquement de deux manières. D'une part, un arrangement séquentiel de deux analyseurs ou plus peut être utilisé (*i.e.* tandem in space), où les ions sont isolés, activés et dissociés dans des endroits physiques différents. C'est le cas du triple quadripôle – QqQ, équipé d'une cellule de collision type Collision Induced Dissociation – CID. Cette technique est qualifiée de dissociation par résonance et permet d'accéder à des voies de fragmentation compétitives. D'autre part, un seul analyseur peut suffire, où l'isolation, l'activation et la fragmentation des ions se produisent successivement au même endroit physique appelé *tandem in time*. C'est le cas notamment du Q-Orbitrap, équipé d'une cellule de collision type Higher-energy Collisional Dissociation – HCD). Cette technique est au contraire de la CID, qualifiée de dissociation non résonante. L'HCD permet d'atteindre de plus forte énergie de collision en comparaison avec CID étant donné qu'une grande quantité d'énergie est appliquée sur l'ion en un laps de temps court. En effet, la redistribution de l'énergie interne des ions est fortement réduite. Par conséquent, de plus hautes énergies permettent d'accéder à des voies de fragmentations consécutives en plus des compétitives. (33) Par conséquent, le CID permettra d'atteindre principalement les voies de dissociation à plus basses énergie, tandis que la HCD pourra accéder aux voies nécessitant des niveaux d'énergie les plus élevés. (33,34)

Enfin, la quantité d'information obtenue lors de l'analyse peut être optimisée par l'utilisation de modes d'acquisition automatique. (35) Ces principaux modes sont : l'analyse data-indépendante (data independent analysis - DIA), et l'analyse data-dépendante (data dependent analysis - DDA). La DIA implique que tous les ions d'une gamme de m/z soient fragmentés. (23,36) La DIA assure un bon recouvrement des analytes présents dans un échantillon. L'acquisition des spectres de fragmentation des ions peut soit se faire à un moment donné, soit par multiplexage de toute la gamme des m/z en plus petites fenêtres d'isolement. (23) Elle est cependant limitée par des phénomènes de convolution (*i.e.* perte de la parentalité entre l'ion parent et les fragments, nécessitant un traitement des données plus complexe). Pour cette raison, la DIA est moins appliquée que la DDA en métabolomique. (23) En effet, lors d'une DDA, le spectromètre passe automatiquement du mode MS au mode MS/MS lorsque les ions précurseurs dépassent un seuil d'intensité prédéfini. Ainsi, seuls les N ions les plus intenses du spectre MS sont fragmentés.(37)

I.2.3. Analyse et visualisation des données spectrales MS/MS

Les avancées technologiques en matière d'outils analytiques, et plus particulièrement au sujet de la spectrométrie de masse, permettent aujourd'hui d'obtenir des fichiers de spectres riches en informations. Ainsi, les spectres de plusieurs milliers de *features* peuvent être enregistrés à partir de l'analyse d'un mélange complexe. (38) En spectrométrie de masse, une similarité des schémas de fragmentation de deux *features* est représentative d'une similarité structurale entre eux.(39) Des molécules partageant des structures similaires peuvent ainsi être regroupées dans des groupes appelés familles moléculaires. C'est pourquoi, la comparaison des spectres MS^2 au sein des extraits et entre différents extraits constitue la pierre angulaire de l'analyse des données issues de la MS. (38) Cette comparaison est rendue possible aujourd'hui *via* l'incorporation de méthodes informatiques, notamment grâce à l'utilisation des réseaux moléculaires.

Dans un réseau moléculaire, les spectres MS^2 sont représentés sous forme de nœuds, reliés entre eux par des liens symbolisant le degré de similarité spectrale. Plus deux spectres partagent de pics en commun, plus ils sont similaires. Le degré de similarité entre deux spectres est calculé à partir d'algorithmes développés originellement pour la protéomique (*i.e.* MS-Cluster (40)). L'objectif initial consiste à simplifier chaque spectre en le représentant sous forme de spectre consensus. Les données spectrales sont présentes dans un espace multidimensionnel complexe, défini par les rapports masse sur charge (m/z). L'intensité de chacune des variables correspond à la valeur de coordonnée d'un point dans cet espace multidimensionnel. En d'autres termes, chaque point dans cet espace représente une combinaison unique de rapports m/z et d'intensités (*c.f.* Fig. 13), simplifiant ainsi la représentation des



spectres.(41,42) Dans le cas des réseaux moléculaires, la similarité spectrale est le plus souvent calculée selon un produit scalaire représentant le cosinus de l'angle entre les vecteurs de deux spectres. Un cosinus égal à 0 signifie que les spectres ne partagent aucune similitude. Au contraire, un cosinus égal à 1 signifie que les spectres sont identiques. (43) Une valeur seuil est ensuite établie pour le *cosine-score* (généralement entre 0,5 et 0,7) et pour le nombre de pics en commun à chaque spectre (généralement entre 3 et 7) au-dessus desquels les liaisons sont établies entre les paires de spectres de consensus.(44)

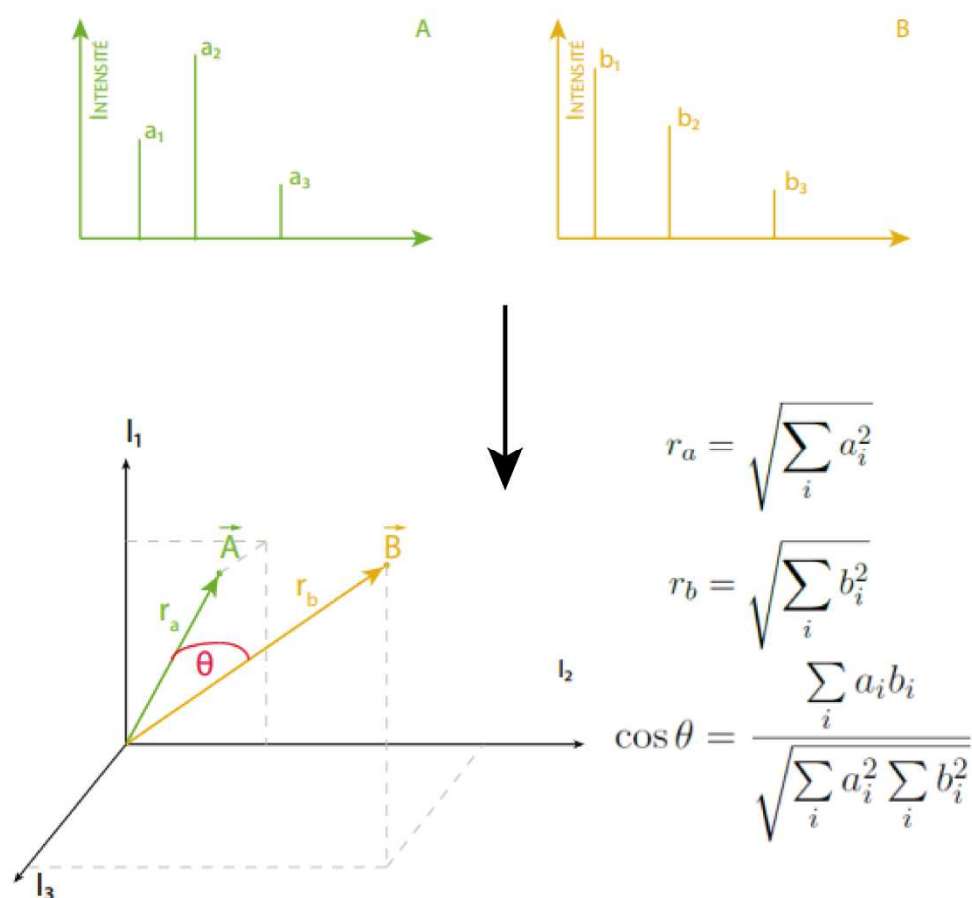


Figure 13: Schéma explicatif de la comparaison spectrale via le cosine-score, selon Wan et al. (45)



I.3. Gestion des données en recherche

La bonne gestion des données de recherche fait encore face à plusieurs obstacles au sein de la communauté scientifique. Nous pouvons citer entre autres le peu de connaissances des chercheurs quant aux plateformes de dépôt disponibles, la réticence de certains à partager leurs données, ainsi que la question économique relative à la mise en œuvre des recommandations. Mais contrairement à une crainte au possible « pillage » des données de recherche, il a été démontré que les scientifiques qui partagent les données associées à leurs articles bénéficiaient de plus de visibilité. En effet, une forte croissance des citations (*i.e.* 25.36%) est observée lorsque des données sont associées à un article. (46) Au contraire, le non-respect de certaines recommandations (*c.f.* 1.3.1.1. *La recommandation FAIR* & 1.3.1.2. *La recommandation TRUST*) peut entraîner des répercussions sur plusieurs aspects de la recherche. La Commission Européenne estime qu'au moins 10,2 milliards d'euros ont été perdus en 2019 par l'ensemble des acteurs de la recherche européenne dû à l'absence de données suivant les recommandations FAIR), telles qu'illustrées en Fig.14 (47)



Figure 14 : Analyse coûts-avantage des données de recherche FAIR, selon (47)

I.3.1. Garantir la qualité et la réutilisation des données de référence

Chaque nouveau composé décrit dans la littérature scientifique devrait idéalement être relié à sa structure moléculaire, à l'origine biologique de l'échantillon au sein duquel la molécule a été identifiée, et aux données spectrales qui ont permis de l'identifier. (29)

L'utilisation de méthodes issues de la métabolomique tels que les réseaux moléculaires nécessitent le développement de nouvelles normes du dépôt pour l'ensembles des données relatives aux substances naturelles, et notamment leurs données spectrales. (48) C'est un principe vrai pour la spectrométrie de masse mais aussi pour la RMN ou toutes autres données analytiques utiles à l'identification d'une SN. D'autant plus que ces données de référence constituent la base du développement d'outils de prédictions de propriétés physico-chimiques des molécules d'intérêt.

L'expansion des plateformes spectrales MS collaboratives a conduit en toute logique à la multiplication des liens d'accès aux données spectrales de référence. Plusieurs liens différents peuvent ainsi mener au même spectre. Le dépôt de données spectrales nécessite donc l'utilisation de codes d'identification des spectres (*e.g.* MSV et CCMSLIB, SPLASH (Spectral Hash)) d'une substance naturelle. Ces codes sont par exemple générés et implémentés lors du dépôt de jeux de données spectrales sur les plateformes MassIVE (Environnement Virtuel Interactif de Spectrométrie de Masse - *Mass Spectrometry Interactive Virtual Environment*), ou les bases de données MassBank, GNPS, *etc.* (48) Ce système d'identifiant unique permet alors d'améliorer le partage, la recherche de données spectrales de masse en évitant les doublons, tout comme avec l'encodage InChIKey, le numéro CAS, ou le PubChemCID pour une structure moléculaire (*c.f.* I.4.2. Encodage numérique de la structure moléculaire).

Cependant, les bases de données « maisons » sont encore nombreuses à être utilisées au laboratoire. Chacune de ces BD suit son propre schéma, et les trois piliers de la déréplication ne sont souvent respectés qu'en partie. Enfin, peu de ces BD sont harmonisées, ce qui rend difficile leur partage entre équipes. De façon générale, la création, la gestion et l'interrogation des bases de données devrait suivre deux grandes recommandations : FAIR et TRUST, comme c'est le cas pour la base de données structurales de substances naturelles LOTUS. (4)



I.3.1.1. La recommandation FAIR

La recommandation FAIR fut introduite par Mark D. Wilkinson en 2016. Ses recommandations regroupent un ensemble de critères qui s'appliquent aux données scientifiques numériques, afin d'améliorer leur qualité, la transparence de leur gestion. Cette démarche permet de faciliter la collaboration et l'interconnectivité entre les scientifiques du monde entier. Ces principes sont les suivants : (49)

- *Findable* (trouvable) : les données doivent être facilement localisables et doivent inclure des métadonnées (*i.e.* données de données) détaillées, des identificateurs uniques (*e.g.* identifiant d'objet digital -DOI). L'utilisateur peut ainsi trouver et sélectionner directement les informations dont il a besoin.
- *Accessible* : les données doivent être accessibles de manière simple, fiable et sécurisée (*e.g.* grâce à une adresse web - *Uniform Resource Locator* – URL). De plus, les licences et conditions d'utilisation des données doivent être clairement définies. Tout utilisateur doit avoir accès à cette donnée sans restriction.
- *Interopérable* : les données doivent suivre une organisation structurée, cohérente et standardisée. Les données peuvent être utilisées efficacement sur différents outils informatiques. Pour cela, les formats de fichiers ouverts et non propriétaires seront privilégiés.
- *Réutilisable* : les données doivent être documentées de manière complète. C'est-à-dire que le plus de métadonnées doivent être reliées (*e.g.* conditions de collecte, traitement, analyse) La compréhension et la réutilisation efficace de ces données par d'autres utilisateurs (humains ou machines) en dépend.

Le principe FAIR met ainsi en lumière le besoin de se conformer aux bonnes pratiques dans la gestion des données. Cependant, ces données doivent être préservées dans le temps, ce qui nécessite des dépôts d'archives numériques fiables (*Trustworthy Digital Repositories* -TDR).

I.3.1.2. La recommandation TRUST

La recommandation TRUST aborde, à la différence de FAIR, la question d'organisation consensuelle et d'infrastructures fiables pour le stockage de données numériques. Cinq points définissent le principe TRUST : (50)

- **Transparence** : les utilisateurs doivent avoir une visibilité claire sur l'origine des données, les transformations appliquées et les règles de gestion associées.
- **Responsabilité** : les créateurs d'une base de données sont responsables de l'authenticité et de l'intégrité des données détenues. De plus, ils doivent gérer la fiabilité et la pérennité du service de stockage de l'information.
- **Utilisateur orienté** : lorsqu'une base de données est créée, il faut veiller à ce que les normes de gestion des données et les attentes des utilisateurs concordent.
- **Soutenable** : le service proposé doit être maintenu, et la préservation des collections de données doit être durable.
- **Technologie** : il faut fournir l'infrastructure et les capacités technologiques adéquates pour soutenir la sécurité, la pérennité et la fiabilité des collections de données.

L'utilisation des deux principes précédemment évoqués pour la construction, la gestion, ou l'interrogation juste et équitable des données spectrales pérennes et de qualités serait profitable à la communauté scientifique des substances naturelles.

I.3.2. Solutions mises en œuvre

Au regard des points précédemment discutés, il est essentiel d'améliorer la qualité et la réutilisation des données en recherche en mettant l'accent sur deux aspects fondamentaux : d'une part, une définition plus rigoureuse des données avant même leur création, et d'autre part, l'établissement de solutions pérennes pour le stockage et la gestion de ces données. Ces deux aspects, permettraient de faciliter l'interprétation et la réutilisation des données par d'autres chercheurs, tout en assurant la conservation, la sécurité et l'accès à long terme à ces données, afin de contribuer à une science ouverte et cumulative.

1.3.2.1. Plan de gestion de données

Le Plan de Gestion de Données (PGD, ou Data Management Plan - DMP) est le document synthétique qui décrit les différentes étapes de collecte, traitement, description, partage, protection et archivage des données d'un projet de recherche. Il décrit dans le détail, en s'inspirant des recommandations FAIR et TRUST, les méthodes et processus de création, de fourniture, de maintenance, de conservation et de protection des données. Sa rédaction permet ainsi d'initier très tôt une réflexion collective sur les bonnes pratiques et d'anticiper les questions relatives à la gestion des données.

Ce document est désormais requis par les agences pour satisfaire aux conditions de financement de projet de recherche. L'Agence Nationale de la Recherche (ANR) demande ainsi une première version six mois après le lancement du projet, puis une version finale, en faveur de l'ouverture des données de recherche pour qu'elles soient accessibles librement et gratuitement. Le DMP est un document dynamique et continuellement mis à jour : introduction d'un nouveau jeu de données, données faisant l'objet d'un dépôt de brevet, changement dans le consortium, *etc.* Le responsable du plan de gestion de données est généralement le coordinateur du projet.

Il n'existe pas de trame unique de DMP. Néanmoins des points sont communs dans tous les modèles, à savoir : (51)

- Informations administratives
- Collections de données
- Documentation et métadonnées
- Ethique et cadre légal
- Stockage, sauvegarde, sécurité
- Sélection et conservation
- Partage de données
- Responsabilité et moyens

1.3.2.2. Solutions de dépôts thématiques

En ce qui concerne les données spectrales MS/MS, le consortium ProteomeXchange (PX) a été créé, afin de normaliser et faciliter la soumission et la diffusion de données issues de la spectrométrie de masse dans le domaine public. Il est associé à des plateformes telles que MassIVE, qui est une

ressource communautaire développée par le *Center for Computational Mass Spectrometry* (NIH) afin de promouvoir l'échange mondial et gratuit de données de spectrométrie de masse. Depuis sa création, plusieurs téraoctets de données spectrales y ont été importés. Cette plateforme permet, en plus de stocker les données, de les parcourir, redistribuer, réanalyser par et pour le public.

De même que dans le domaine de la spectrométrie de masse, l'initiative NMReData propose depuis 2018 des protocoles standards pour partager à la communauté les données spectrales RMN, leurs attributions et les paramètres expérimentaux. (30)

I.3.2.3. Solutions Open-Data

La recherche se trouve aujourd'hui confrontée à un défi de taille : préserver, partager et ouvrir les données pour promouvoir une science ouverte, plus transparente et cumulative, au bénéfice des équipes de recherche et de la société dans son ensemble. La gestion des données de recherche devient ainsi un enjeu stratégique et critique, mais représente également une charge de travail supplémentaire, souvent complexe, pour les équipes de recherche. Malgré leur importance, la valeur des données de recherche reste encore sous-exploitée.

C'est dans ce contexte que le Premier ministre a mandaté le député Éric Bothorel pour élaborer un rapport intitulé « Pour une politique publique de la donnée », publié en décembre 2020, afin d'encourager une meilleure valorisation des données.

En réponse à ces enjeux, et dans le cadre du deuxième Plan national pour la science ouverte, la plateforme nationale Recherche DataGouv a été développée par et pour la communauté de recherche. Ce projet, confié à l'INRAE en raison de son expertise reconnue dans le domaine des entrepôts génériques, en collaboration avec le CNRS et sept autres établissements partenaires, a abouti à l'ouverture de la plateforme le 8 juillet 2022. Cette initiative vise à fournir un entrepôt pluridisciplinaire aux chercheurs ne disposant pas d'une solution souveraine pour le dépôt sécurisé et fiable de leurs données. (52–55)

De plus, une publication scientifique, le data-paper, peut être associé au dépôt d'un jeu de données dans un entrepôt de stockage. Un data-paper a alors pour objectif de décrire précisément un jeu de données, en informant la communauté scientifique de son existence, de ses modalités et de son potentiel de réutilisation. Publié en libre accès, le data-paper prend la forme d'un article évalué par les pairs dans une revue scientifique classique. Contrairement aux articles traditionnels, un data-paper ne discute ni des hypothèses ni des conclusions tirées des données, mais se concentre sur les analyses

techniques et statistiques validant la qualité du jeu de données. La publication d'un data-paper présente de nombreux avantages. Elle permet de valoriser les données, d'en faciliter la réutilisation, de leur apporter une visibilité accrue, et de les rendre plus facilement repérables et citables. En tant que publication citable, le data-paper met en valeur ses auteurs en tant que créateurs de données et assure la traçabilité des citations et des réutilisations de ces données. (56)



I.4. Les bases de données

I.4.1. Généralités

Une base de données (BD) est un système de gestion de données utilisé dans de nombreux domaines, et qui permet de stocker, organiser et accéder à des informations de manière efficace. Il existe différents types de bases de données, mais elles ont toutes pour objectif de faciliter la recherche, la mise à jour et l'analyse des données. Pour utiliser une base de données, il est nécessaire d'utiliser un logiciel de gestion de base de données (*Data Base Management Software* -DBMS). Celui-ci est utile pour créer, gérer et interroger les bases de données (*e.g.* MySQL, SQL Server, Oracle, PostgreSQL, MongoDB, *etc.*). Plusieurs modèles de bases de données peuvent exister :

- Les bases de données en réseaux (également appelées BD graphiques). Les graphes permettent de représenter les relations entre les différents éléments de données sous forme de nœuds (ou sommets) et d'arrêtes (ou liens). Les nœuds représentent les entités (*e.g.* une molécule), et les arrêtes représentent les relations entre les entités (*e.g.* similarité structurale), de la même manière que pour la construction d'un réseau moléculaire. Une arrête possède toujours un nœud de départ, un nœud de fin, un type et une direction. Le nombre et le type de relations qu'un nœud peut avoir sont illimités. Dans les bases de données orientées graphe, traverser des liaisons ou des relations se fait très rapidement. (57)
- Les bases de données relationnelles. (58) Ce type de base de données fut introduit par E.F. Codd en 1970. Une BD relationnelle contient un ensemble de relations stockées sous forme de ligne. Chaque relation est un sous-ensemble du produit cartésien de plusieurs domaines de valeurs. Chaque domaine est un ensemble de valeurs de même type. Leur rôle au sein d'une relation est spécifié par la notion d'attribut. Le schéma d'une relation définit enfin l'ensemble des noms d'attributs constitutifs de la relation. L'extension d'une relation correspond à l'ensemble des valeurs de domaines associées au schéma d'une relation. Elle peut prendre la forme d'une table. Ces bases de données utilisent souvent le langage informatique SQL (*Strured Query Language*) pour être gérées, et suivent un modèle orienté valeur où les éléments d'une relation sont identifiés par des clés (*i.e.* attribut ou un groupe d'attributs qui permet d'établir des liens), où les indices font correspondre les valeurs des colonnes aux numéros de ligne. (58) Toutefois, elles se révèlent souvent inadaptées pour la gestion de données à structure complexe (*e.g.* des données spectrales).

- Les bases de données NoSQL (*Not Only SQL*). Il s'agit de bases de données « hybrides », puisqu'elles ne suivent pas les normes de BD relationnelles. Elles sont souvent utilisées pour stocker des données non structurées ou semi-structurées, telles que des images, des vidéos, des fichiers audios, ou des données spectrales. De plus, elles sont plus adaptées à la gestion de grands volumes de données à grande échelle. Plusieurs DBMS ont été développés pour gérer ces DB, tels que MongoDB (*i.e.* système de DB documentaire qui stocke les informations sous forme de dictionnaires (JSON)) ou Apache Cassandra (*i.e.* les données sont stockées en colonne plutôt qu'en ligne comme dans le modèle relationnel, et ainsi les colonnes font correspondre les numéros de ligne aux valeurs des colonnes).

Les éléments primordiaux à incorporer au minimum dans une base de données de substances naturelles font écho à la section I.1.1. *Définitions et généralités*, et englobent la structure moléculaire, son origine biogénétique et son profil spectral. Ces trois aspects constituent les piliers de la déréplication. (29) Le projet SISTEMAT fut le précurseur dans l'agrégation de ces éléments.(59) Plus récemment, d'autres initiatives facilitant l'accès à ces informations ont vu le jour : KnapSack (60), COCONutDB (61) et LOTUS (4). Néanmoins, aucune d'elles ne renvoie à des données spectrales expérimentales pour le moment.

L'accès aux BD de référence spécialisées aux substances naturelles constitue le socle de la déréplication, et donc de l'annotation de composés avec un haut niveau de confiance. L'engouement des scientifiques pour les substances naturelles lors de ces 20 dernières années a engendré la multiplication des BDs (*e.g.* *SuperNatural*, *Universal Natural Product Database*, *ZINC*, *Natural Product Activity and Species Source Database*, *Riken Natural Product Encyclopedia*, *3DMet*, *KNAPsACK*, *Natural Product Atlas*, *DrugBank*, *FoodB*, *Marine Natural Product Database*, *COCONUT*, *LOTUS*). Entre 2000 et 2019, pas moins de 120 BD propres aux substances naturelles ont été publiées. Toutefois, seules 98 d'entre elles sont encore aujourd'hui accessibles, dont 50 en libre accès. (62) Ces BDs peuvent être classées selon plusieurs critères :

- Publique ou propriétaire
- Générale ou orientée
- Structurale ou spectrale
- Si spectrale, données expérimentales ou prédites (*i.e.* *in silico*)



I.4.2. Encodage numérique de la structure moléculaire

La structure d'une molécule peut être illustrée de différentes manières. Une molécule est le plus souvent matérialisée sous forme de graphique, où chaque atome correspond à des nœuds, reliés entre eux par des liens symbolisant les liaisons covalentes (*e.g.* représentation de Lewis, formule développée plane, *etc.*). Néanmoins, la conformation de la molécule dans l'espace n'est pas conservée, à la différence de la représentation de Cram ou la projection de Newman. La formule topologique reste de loin la plus appréciée pour évoquer les SN. Celle-ci ne fait plus figurer les atomes de carbone ni ceux d'hydrogène portés par ces derniers. Les liaisons carbone-carbone sont dépeintes par des traits obliques et peuvent être discriminés en liaison simple, double et triple. Chaque atome de carbone est par conséquent situé à l'intersection de deux segments.

Au fil de ces 60 dernières années, l'énorme quantité de données et d'informations produites par de nombreux domaines de la chimie a nécessité le recours à des méthodes informatiques afin d'aider à leur traitement et à leur analyse. Les représentations intuitives et lisibles par l'Homme ne l'étaient pas pour un ordinateur, et ont nécessité un codage pour garantir que les données moléculaires puissent être traitées par des algorithmes. (63,64) À leurs débuts, les ordinateurs utilisaient des chaînes alphanumériques pour encoder une structure, moins gourmandes en mémoire et en puissance de calcul. Mais avec l'évolution des ordinateurs et la baisse des coûts de mémoire et de traitement, des formats plus flexibles et moins ambigus sont devenus courants. Le Simplified Molecular Input Line Entry Specification (SMILES) est une des premières notations linéaires sous forme de courtes chaînes de caractères ASCII. (65) Les atomes sont représentés par les mêmes symboles utilisés dans le tableau périodique. Cela rend facile l'écriture de chaînes SMILES syntaxiquement correctes. Cependant, cette nature non canonique peut compliquer les recherches de (sous)structures dans une base de données. (66,67) L'encodage SMARTS est une extension plus générique du SMILES car il inclut des opérateurs logiques, des symboles atomiques et de liaison spéciaux. L'encodage SMART est plus adapté à la recherche de sous-structures dans une base de données. (68) L'International Chemical Identifier (InChI) est un autre identifiant textuel, et est considéré comme un standard d'encodage. Il fut initialement développé par l'Union Internationale de Chimie Pure et Appliquée (IUPAC) et le National Institute of Standards and Technology (NIST). Ses principaux avantages sont sa capacité de canonisation intégrée, son statut open source, et son applicabilité à la plupart des composés organiques. Sa structure hiérarchique contient des couches d'informations sur le composé, chacune séparée par un « / » et précédée d'un préfixe (*c.f.* Fig.15). (65) L'InChIKey est une version abrégée et simplifiée de l'InChI (*i.e.* longueur fixe de 27 caractères). Il a été conçu pour être un identifiant unique et moins ambigu de chaque composé chimique. Son utilisation rend ainsi plus facile

la recherche et l'indexation de structure dans les bases de données. Il est obtenu par un hachage fort de la chaîne de caractères InChI selon un algorithme cryptographique (*i.e. strong hash* - SHA-256), garantissant l'unicité et l'intégrité des données. (29)

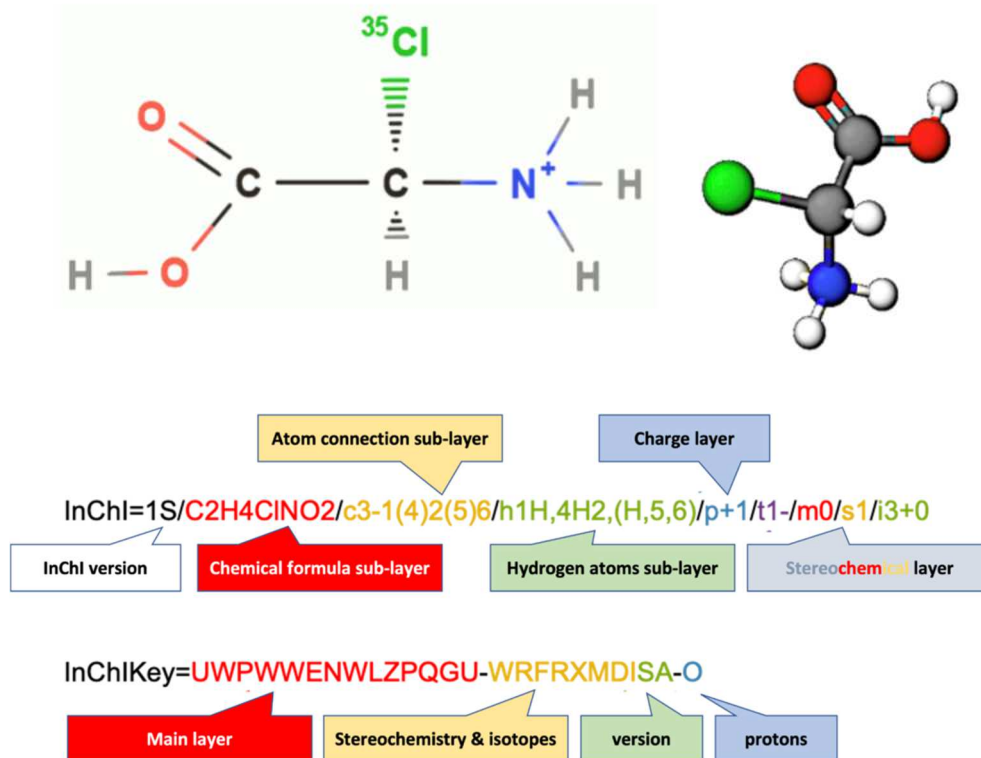


Figure 15 : Encodage d'une structure moléculaire en InChI et InChIKey, selon <https://www.inchi-trust.org/>

Enfin, la structure d'une molécule peut être codée sous forme tableau chimique (chemical table – CT). Un CT fournit les coordonnées spatiales x, y et z de chaque atome dans une table de connexions (CTab), ainsi que les liaisons entre les atomes dans une molécule. Cela facilite la génération de représentations graphiques en 2D ou 3D. Ces tableaux sont souvent utilisés pour représenter des molécules dans des bases de données. Le format le plus utilisé est le fichier MDL molfile. (65) Les fichiers MDL molfile se composent de trois sections principales : un bloc d'en-tête (contenant le titre, l'horodatage et un commentaire facultatif), une table de connexions (CTab) aussi appelé MolBlock, et une ligne de fin qui doit être indiquée par "M END" (*c.f.* Fig.16). Plusieurs profils MDL molfile, représentant différents encodages de molécules, peuvent être regroupés dans un même fichier de données structurales. Le format de fichier le plus courant pour cela est le Structure Data File (SDF). (29,65)

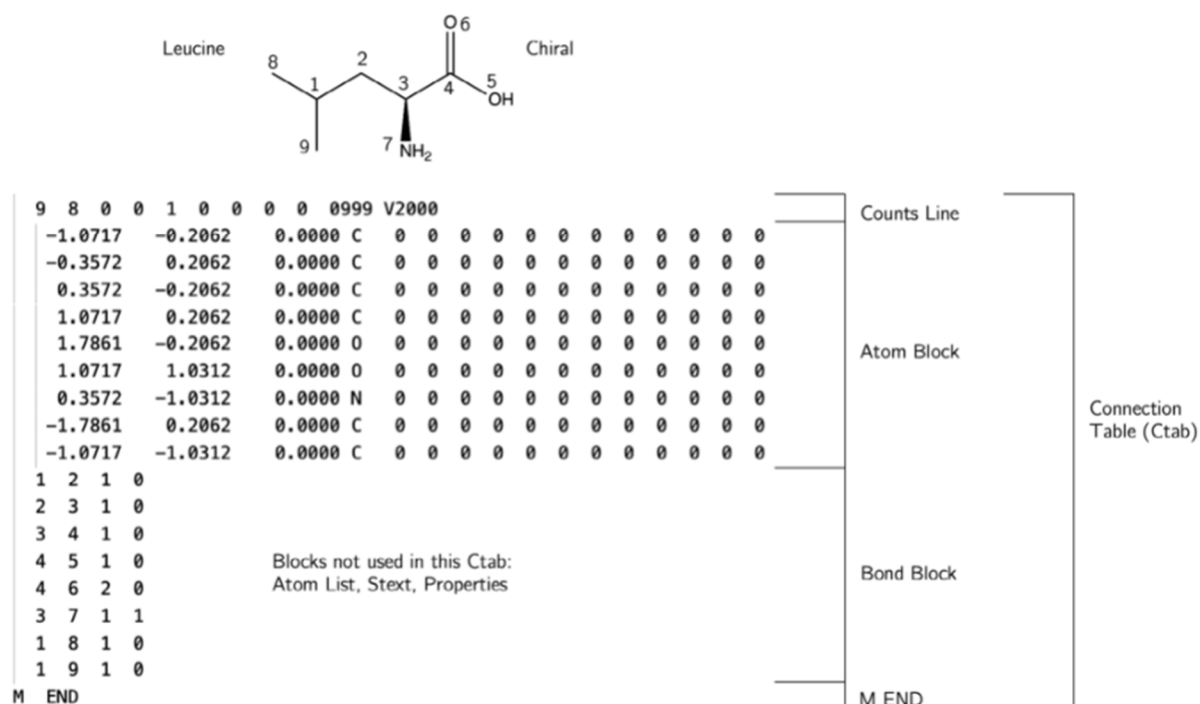


Figure 16 : Structure du fichier MDL pour la Leucine, selon (65)

I.4.3. Rôle des bases de données dans le processus de déréplication

Le profilage chimique dans le cas présent est réalisé de façon non ciblée, c'est-à-dire qu'il a pour objectif d'identifier le plus large panel de métabolites présents dans un échantillon. L'annotation des données analytiques (*i.e.* spectres RMN ou MS²) peut suivre 3 stratégies avancées de déréplication : annotation des chémotypes connus, analyse suspecte, annotation exhaustive (incluant l'utilisation de spectres *in silico*). Pour se faire, des spectres de référence regroupés dans des bases de données sont en généralement comparés aux spectres expérimentaux.(38)

I.4.3.1. Comparaison aux bases de données spectrales

Le fait de travailler avec une base de données spectrales expérimentale pour l'approche de déréplication est l'idéal, et augmente la probabilité d'obtenir de meilleures annotations.

I.4.3.1.a. MS/MS

La comparaison des spectres MS^2 expérimentaux à ceux issus de bases de données spectrales de référence résulte de la même approche que celle utilisée lors de la construction des réseaux moléculaires. Les algorithmes de calcul de similarité spectrale donnent généralement un score de correspondance représentant la probabilité qu'un spectre MS^2 expérimental soit apparié au spectre d'un métabolite référencé. (41) De même que les algorithmes utilisés en RMN, l'attribution des pics d'un spectre peut se faire de façon directe ou inversée. L'attribution inversée permet, au contraire de la directe, de considérer uniquement les pics du spectre expérimental ayant été appariés au spectre de référence. Dans ce cas, le score de similarité n'est donc pas impacté par les fragments pour lesquels aucune correspondance n'est retrouvée. (41,69) De plus, un cosinus-score modifié (*i.e. modified dot product*) (70,71) a été proposé, permettant de faire l'appariement entre les fragments de deux ions précurseurs (MS^1) de valeur m/z décalée, de même qu'avec l'algorithme TREMOLO. (72) Ainsi, deux métabolites de structures similaires peuvent concorder même s'ils diffèrent d'un groupe fonctionnel (*i.e. fonction chimique*). (41) La méthode de comparaison des spectres MS^2 pour le calcul du score de similarité peut prendre en compte, en plus des fragments de chaque spectre, les pertes neutres (*Neutral Loss* - NL), définie comme la différence de masse (m/z) entre l'ion précurseur et chaque ion de fragmentation. La méthode d'appariement spectral METLIN-NL a été la première à exploiter ce principe. Une base de données a été créée, dans laquelle chaque métabolite est associé à un spectre de masse de pertes neutres calculé. Chaque spectre MS^2 expérimental est converti de la même manière en un spectre de pertes neutres, puis comparé à la base de données précédemment décrite, selon l'algorithme dédié. (73,74) Les pertes neutres peuvent également être calculées de façon hypothétique (perte neutre hypothétique – *Hypothetical Neutral Loss* -HNL), telle que Xing *et al.* le proposent. (75) L'HNL est définie comme la différence de masse (m/z) entre les deux pics d'une paire de fragments d'un même spectre MS^2 . Chaque spectre MS^2 (de référence ou expérimental) peut être converti en un spectre de pertes neutres hypothétiques. Puis, l'algorithme *Core-Structure-based Search* (CSS) permet de calculer le degré de similarité entre ces spectres, et ainsi d'annoter des métabolites ayant une structure principale similaire. (75) L'algorithme *Hibrid Similarity Search* (HSS) tient compte du principe de similarité reposant à la fois sur la comparaison des fragments (*cosine-score, modified cosine-score*) et sur la comparaison des pertes neutres (NL, HNL). (76)

D'autres algorithmes de comparaison spectrale ont été développés, avec comme objectif final d'obtenir la structure annotée la plus proche de la réalité. Par exemple le concept d'entropie spectrale, prend en compte le nombre de fragments et leurs intensités. Les spectres MS^2 sont alors comparés en intégrant ce paramètre (*i.e. complexité spectrale ou probabilité de distribution des intensités des*



fragments d'un spectre).(77) Enfin, SIMILE est un algorithme qui repose sur la conversion en matrice Laplacienne de la matrice de similarité de tous les fragments présents dans une paire de spectres MS² permettant d'obtenir un espace de dimensions réduit. L'alignement et la comparaison peuvent ainsi être effectués de manière plus efficace et significative. (78)

I.4.3.1.b. RMN

En RMN, un spectre expérimental correspond à un ensemble de déplacements chimiques (¹H, ¹³C, etc.) et des constantes de couplages (ⁿJ_{1H-1H}, ⁿJ_{1H-13C}, etc.) qui sont propres à chaque molécule. Une molécule possède ainsi une empreinte (*fingerprint*) unique d'un point de vue analytique (ici en RMN, (δ , J)), pour un solvant et à une température donnée, qu'importe que le métabolite soit en mélange avec d'autres.(79) Ces empreintes peuvent ensuite constituer l'intrant des méthodes d'annotation telles que DerepCrude, CaraMel ou encore Mixonat, utilisant des outils statistiques supervisés ou non. (80)

La petite plage spectrale de la RMN ¹H (12 ppm) peut être responsable d'un degré de chevauchement important qui est proportionnel à la complexité du mélange analysé. Cela rend l'analyse des signaux ¹H plus délicate. Pour cette raison, l'utilisation de la RMN 1D ¹³C est considérée comme une manière de contourner le problème de superposition du signal. (38) Le ¹³C est un isotope moins abondant que le ¹²C, et détectable de façon moins sensible que le ¹H. Cependant, tous les composés organiques sont composés de carbone par définition, ce qui rend possible l'analyse des métabolites majoritaires dans un mélange complexe. De plus, le ¹³C présente une plage spectrale beaucoup plus élevée que celle du proton (x 20 – 240 ppm), et une fois les spectres de RMN ¹³C découplés du proton, il n'y a pas de phénomène de multiplet (*i.e.* un signal correspond à un pic). (38,81) Chacune des méthodes mentionnées utilise dans leur processus des empreintes RMN 1D ¹³C. (82,83)

La méthode nommée CaraMel permet de mettre en évidence les corrélations statistiques dans l'ensemble des données produites (déplacements chimiques ¹³C). (84) Une classification hiérarchique ascendante (CHA) non supervisée est réalisée, et permet de regrouper les déplacements chimiques en clusters (*i.e.* groupes de déplacements chimiques), en fonction de la variation relative de l'intensité des pics entre les différentes fractions, c'est-à-dire de similarité du profil chromatographique (*c.f.* Fig. 25). Chaque liste de déplacements chimiques ¹³C propre à un cluster pourra ensuite être comparée, *via* le logiciel ACD/C+H NMR Predictors and DB Workbook Suite, à une base de données associant des déplacements chimiques ¹³C expérimentaux de références ou prédits (85,86), et structures chimiques.

La méthode DerepCrude (79) découle de la méthode CaraMel. A la différence de la précédente, DerepCrude est une méthode supervisée, puisque les intensités relatives entre pics sont utilisées comme paramètres lors de la comparaison des déplacements chimiques ^{13}C expérimentaux à ceux de référence. De plus, DerepCrude permet d'identifier des composés en mélange directement sans passer au préalable par une étape de fractionnement de l'échantillon, en comparant les déplacements chimiques ^{13}C expérimentaux obtenus à ceux d'une DB prédite grâce à l'algorithme développé. (79)

CSEARCH (87) propose quant à lui un autre algorithme complet et modulable qui, en plus d'effectuer une recherche de correspondance, permet l'attribution des pics à chaque carbone d'une structure connue de façon entièrement automatisée. Pour cela, l'algorithme nécessite une structure chimique, la liste de déplacements chimiques ^{13}C expérimentaux associés ou non à leur multiplicité. L'espace de recherche peut être filtré en utilisant une formule moléculaire, ou être orienté par une sous-structure renseignée par l'utilisateur. Les déplacements chimiques de la structure connue sont prédits selon un algorithme de prédiction (*i.e.* HOSE(88)) qui utilise une base de données de référence associant plus de 8 000 structures chimiques et déplacements chimiques expérimentaux. L'information sur la multiplicité des carbones peut servir comme filtre dans ce processus de recherche, afin de limiter les sous-structures à comparer dans la base de données. (87)

Dans la même optique, NMRShiftDB est une méthode très similaire à CSEARCH, puisqu'elle permet à la fois de prédire les déplacements chimiques ^{13}C , mais aussi de faire correspondre une liste de déplacements chimiques à une liste de référence issue d'une base de données spectacles. (89)

MixONat est une méthode d'analyse supervisée, et prend en plus le principe de parité du nombre de protons attachés à chaque atome de ^{13}C (*i.e.* le type de carbone : C_q , CH, CH_2 , CH_3), déduit à partir de l'expérience combinée DEPT-135, et introduit par la méthode INFERCNMR. (90) A l'origine, INFERCNMR était utilisée pour prédire des sous-structures d'un composé inconnu à partir de son spectre de RMN 1D ^{13}C , de la multiplicité de chacun de ses carbones, de la formule brute du composé et d'une base de données spectrale de référence. MixONat ajoute un filtre reposant sur la masse molaire à l'algorithme de recherche. Il est ainsi plus aisé de confirmer ou exclure des propositions d'annotation après comparaison avec la base de données *via* le même logiciel que pour CaraMel. (91)

La recherche de correspondances entre les données expérimentales et celles d'une base de données est très efficace en RMN. En effet les déplacements chimiques expérimentaux 1D ^{13}C des métabolites permet d'identifier les composés avec une précision de 0,1 ppm, si toutes les analyses sont menées dans le même solvant deutéré et avec les mêmes paramètres. (92) Cette comparaison peut être réalisée selon trois algorithmes différents : attribution directe, attribution inversée, et attribution bipartite. L'attribution directe assigne à chaque pic de la liste expérimentale le pic d'un



composé de référence dans la base de données ayant une différence de déplacement chimique la plus faible possible. Ainsi, chaque pic de la liste expérimentale est assigné à un pic de référence, mais le contraire n'est pas vrai (*c.f. Figure 17. B*). C'est le contraire pour l'attribution inversée : ce sont les pics de référence qui sont assignés à partir des pics expérimentaux ayant une différence de déplacement chimique minimale. Ainsi, chaque pic de la liste de pics de référence d'un composé dans la base de données est assigné à un pic expérimental, mais le contraire n'est pas vrai (*c.f. Figure 17. C*). Enfin, l'attribution bipartite permet d'obtenir des assignements non ambigus au regard des deux algorithmes précédents. Dans ce cas, le nombre de pics dans les listes expérimentales et de références pour un composé de la base de données est comparé. L'algorithme fait alors correspondre la plus petite des deux listes de pics avec la plus grande, afin que chaque pic de la plus petite liste soit attribué de manière unique à un pic de la liste plus grande. Ainsi, deux pics de la plus petite liste ne peuvent être assignés à un même pic de la plus grande liste (*c.f. Figure 17. A*). (93)

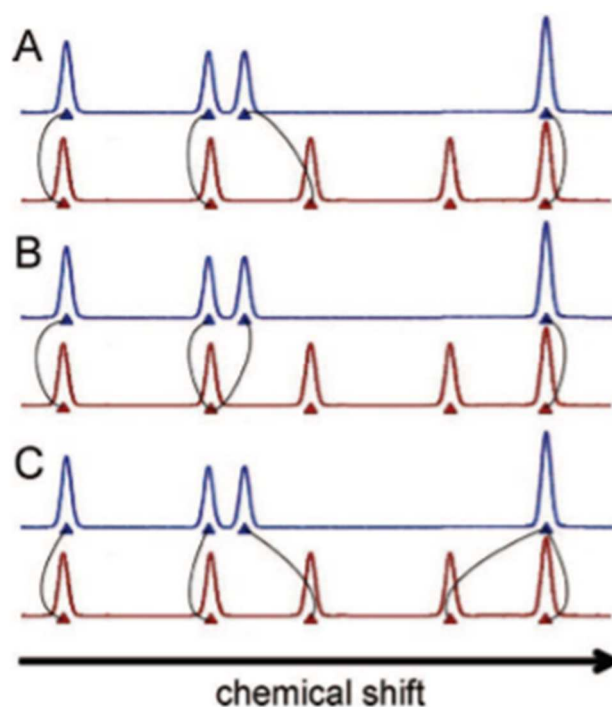


Figure 17. Algorithmes d'attribution (Bleu : spectre expérimental, Rouge : spectre de référence), selon (93) A) Attribution bipartite, B) Attribution directe, C) Attribution inversée

I.4.3.2. Au-delà de la comparaison de spectres expérimentaux

Des faiblesses dans la relation entre les scores de similarité spectrale et les vraies similarités structurales ont été décrites.(94) Le calcul de la similarité entre deux spectres MS^2 a récemment su tirer parti des approches d'intelligence artificielle (IA) par apprentissage profond (*deep learning*). Huber *et al.* ont appliqué des méthode originellement dédiées au traitement automatique du langage

naturel (*Natural Language Processing* – NLP), afin d'apprendre les co-occurrences entre fragments ou pertes neutres dans un spectre MS^2 . Il est ainsi possible de dériver des enchâssements spectraux synthétiques (*i.e.* représentation simplifiée, ou proxy), à partir d'un spectre MS^2 , qui seront alors utilisés pour évaluer les similitudes spectrales. (94) Enfin, la même équipe a utilisé les réseaux neuronaux siamois (*i.e.* réseau neuronal qui travaille en tandem sur deux vecteurs d'entrée différents pour calculer des vecteurs de sortie comparables) pour prédire directement cette fois la similarité structurale à partir de spectres MS^2 . Chaque spectre MS^2 est converti en une représentation vectorielle à 200 dimensions par un modèle entraîné qui est utilisée pour prédire un score de similarité structurale, dit de Tanimoto. (95) Par ailleurs, l'accumulation des connaissances acquises sur les relations entre les structures moléculaires et les spectres expérimentaux (*i.e.* MS/MS et RMN 1D ^{13}C) a rendu possible la conception d'outils de prédiction spectrale.

Plusieurs algorithmes de fragmentation existent pour la simulation des spectres de masse MS^2 , et chacun utilise une approche différente. La fragmentation d'une structure par rupture de liaisons chimiques spécifique peut suivre des règles de fragmentation (*e.g.* MS -FINDER, MAGMa) (96), elles-mêmes construites à partir de spectres de masse expérimentaux. Au contraire de la fragmentation reposant sur des règles, des approches combinatoires sont également disponibles : MetFrag. (97) Cette méthode n'a pas pour but *stricto sensu* de prédire un spectre de fragmentation. Elle a plutôt pour objectif d'expliquer les pics d'un spectre MS^2 en les reliant à des sous-structures elles-mêmes issues de bases de données structurales. (98) D'autres méthodes reposent sur l'apprentissage automatique (*i.e.* machine-learning) grâce à un modèle probabiliste du processus de fragmentation, construit à partir de données expérimentales : CFM-ID. (99) Ces approches sont d'autant plus intéressante dans une stratégie d'annotation exhaustive, car elle permet de générer des hypothèses structurales pour les inconnus. Les bases de données spectrales simulées (*i.e.* *in silico*) représentent une alternative attrayante (*e.g.* UNPD-ISDB, FooDB, AntiBase, Mona, *etc*), et pourraient remplacer les données spectrales expérimentales, du moins pour les premières étapes de déréplication. (62,100) Ainsi, l'exemple de l'ISDB montre qu'en augmentant le nombre de spectres MS^2 (>10 fois par rapport à tous les spectres expérimentaux de MS^2 disponibles à l'époque) l'espace chimique pouvant être exploré par la déréplication par LC-HRMS² est considérablement augmenté. (101) Récemment, cette même démarche a été utilisée pour la simulation des structures présentes au sein de la base de données structurales précédemment décrite : LOTUS. (102) La prédiction s'est faite uniquement en mode d'ionisation positive, et réalisée à l'aide de la version 3 de CFM-ID. Les 3 spectres de niveau d'intensité différent (*i.e.* faible, moyen et élevé) par structure ont été moyennés.



La même démarche a été menée pour la création d'une base de données de déplacements chimiques prédits ^{13}C . Les déplacements chimiques RMN ^{13}C retrouvés dans Acd_lotusv7 ont été prédits selon l'approche HOSE (*Hierarchical Organisation of Spherical Environments*) utilisée par le programme informatique nmrshiftdb2. Cet algorithme repose sur l'utilisation de bases de données spectrales expérimentale et fonctionne de la manière suivante : dans une première étape, le tableau de connectivité entre les carbones de la structure fournie est généré selon le code HOSE allant jusqu'à cinq niveaux (*i.e.* découpe la structure principale en sous-structures où les deux atomes terminaux sont espacés au maximum de cinq liaisons). Pour chaque niveau, les déplacements chimiques correspondant aux carbones de chaque sous-structure générée sont recherchés dans une base de données de référence associant sous-structures et déplacements chimiques. Pour un composé donné, chacun des carbones des sous-structures du premier niveau est comparé à la base de données, puis l'on passe au niveau suivant, jusqu'à ce que tous les carbones de la structure de la requête soient assignés. Un contrôle est réalisé entre chaque cycle. (88)

Il est toutefois crucial de garder à l'esprit que les concordances entre spectres expérimentaux et prédits doivent être interprétées avec prudence et ne sauraient se substituer à la comparaison avec les spectres de référence acquis dans les mêmes conditions d'acquisition (*i.e.* ionisation et fragmentation en MS, solvant en RMN). Il a notamment été démontré que la précision de ces logiciels de prédiction en RMN dépendait fortement de la classe structurale des métabolites naturels, mais pas du type d'hybridation du carbone. (92)

I.4.4. Etat général des limites des bases de données existantes

I.4.4.1. Exhaustivité

Aucune des BD spécialisées aux substances naturelles disponibles ne représente l'ensemble de l'espace chimique que recouvrent les substances naturelles. Certaines se limitent à la localisation géographique d'origine des échantillons analysés, une classification taxonomique, chimique, ou à un usage en particulier. Une cinquantaine de BD relatives aux substances naturelles, encore accessibles et en libre accès, ont été comparées pour étudier leur recouvrement. Les substances naturelles reportées sont majoritairement lipophiles, et dans une moindre mesure hydrophile. De plus, 40 des 50 BD ont un recouvrement d'au moins 50 % avec une autre BD. Sur l'ensemble des BD, 50% des molécules comportent au moins une information stéréochimique, et 70% de ces informations concordent entre les différentes BD. Il est intéressant de noter que seulement 5 composés sont

communément retrouvés parmi 34 des 50 BD : apigénine, quercétine, camphérol, catéchine, naringénine, et sont tous des flavonoïdes. (103)

Pour que l'étape de déréplication soit optimale lors de l'investigation d'un échantillon naturel, il serait nécessaire d'établir une base de données englobant les spectres des quelques 400 000 substances naturelles répertoriées jusqu'à présent. (103) Aussi, en l'absence de spectres de référence, les outils de prédiction / simulation spectrale (*e.g.* CFMID-4.0 (99), NMRShiftDB (89)) peuvent générer des spectres *in silico* (*i.e.* spectres prédits). (29,86) Une telle démarche amène rapidement à des listes de candidats d'annotation de grande taille et dont la majorité des candidats ne sont pas pertinents avec le sujet d'étude. De plus, l'utilisation d'une base de données restreinte aux composés présents dans le taxon d'origine de l'organisme étudié s'avère plus pertinente. (100)

I.4.4.2. Incohérences

A titre d'exemple, un même composé peut avoir différents identifiants chimiques (*e.g.* SMILES, InChI, InChIKey, etc.) au travers des différentes BD, et un même nom de molécule peut mener à différentes structures selon les BD. De telles incohérences augmentent considérablement le temps passé à sélectionner et établir une liste de candidats d'annotation « correcte ».

I.4.4.3. Accessibilité

Pour qu'une donnée soit exploitable par un utilisateur, il faut tout d'abord qu'elle lui soit accessible. Bien qu'il existe des bases de données structurales accessibles gratuitement (*e.g.* LOTUS, COCONUT, etc.), d'autres commerciales le sont en contrepartie d'une rétribution (*e.g.* Dictionary of Natural Products). Enfin, des BD propriétaires ne sont accessibles que par leur créateur ou l'entreprise à laquelle elles appartiennent.

L'accessibilité des données inclue également une notion temporelle. L'information doit être maintenue dans le temps. Or, la durabilité de l'accessibilité à l'information peut en être mise en péril pour des raisons de coûts financiers (*e.g.* Universal Natural Product DataBase – UNPD). La disparition de bases de données amène dramatiquement à la perte d'informations. (103)

I.4.4.4. Versatilité

Chaque BD est construite selon son propre modèle et intègre les données et métadonnées qu'elle juge nécessaire lors de sa création (*i.e.* schéma de l'information). Elles peuvent être de différents supports, tels que des catalogues industriels (*e.g.* Merck), des livres sur les substances naturelles (*e.g.* Bruneton), des compilations de structures (*e.g.* Supplementary materials) ou site



internet spécialisé (*e.g.* LOTUS). Chacune d'entre elles ne permet pas de faire les mêmes requêtes. Par exemple, la recherche de molécules peut se faire suivant le critère de la formule moléculaire dans l'une, ou selon la classe chimique de la molécule d'intérêt dans l'autre. C'est pourquoi la recherche harmonisée au travers de plusieurs BD en même temps apparaît être une tâche complexe.

I.4.4.5. Données spectrales

Cette dernière limite est sans doute l'une des plus impactante sur le choix d'une BD pour la déréplication d'un mélange complexe puisqu'elle touche directement à l'aspect analytique. Premièrement, toutes les BD propres aux substances naturelles ne comportent pas de données spectrales de référence pour une molécule connue, ce qui rejoint le problème d'exhaustivité des BD. Quand bien même ces données sont reportées, elles peuvent varier de façon importante. En effet, les données spectrales MS sont considérablement impactées par le type d'appareil utilisé, l'environnement d'acquisition et les paramètres appliqués. Il en est de même pour la RMN dans une moindres mesure (*i.e.* les données peuvent varier en fonction du solvant utilisé, l'intensité du champ magnétique du spectromètre, *etc.*). Ces considérations techniques apportent en complexité la comparaison des données spectrales.



Partie 2 : Travaux expérimentaux



Preservation for better use: Towards the Digital Alkaloid Collection of Reims

J. Cordonnier^{1*} ; N. Alves Dacosta¹ ; G. Massiot¹ ; J-H. Renault¹ ; S. Remy^{1*}

¹Institut de Chimie Moléculaire de Reims, UMR CNRS 7312, Université Reims-Champagne-Ardenne, UFR Sciences,
BP 1039, CEDEX 2, 51687 Reims, France

Abstract

Natural substances (NS) are molecules derived from the metabolism of living organisms and play a vital role in modern medicine. However, only a small fraction—fewer than 40,000 of the approximately 350,000 known flowering plant species—has been chemically profiled. Metabolomics, specifically chemical profiling, allows for the identification of specialized metabolites from samples, such as the indole alkaloids in the Madagascar periwinkle (*Catharanthus roseus*), using spectral analysis. These spectral data serve as the unique identifiers for NS, making their preservation essential, particularly given challenges in physical storage. Data preservation is crucial for dereplication, a key process that identifies known compounds by comparing experimental spectra to reference spectra. Building robust databases for this information is therefore critical to accelerate research and improve result reliability. Furthermore, the application of FAIR principles (Findable, Accessible, Interoperable, Reusable) ensures that these data are accessible and usable by the scientific community, supporting advances in computational structure elucidation and in silico biological activity assessments. This study initiates the digitization of the alkaloid collection at the UMR CNRS 7312 – Institut de Chimie Moléculaire de Reims (URCA) by creating the Digital Alkaloid Collection of Reims (DACR), a preliminary database for MS/MS and 1D ¹³C NMR spectra. This database is also intended as a model for future digitalization to enhance data interoperability.

Introduction

Seventy-five percent of the world's flowering plant species are currently known, totalling approximately 350,000 species. (1) The exploration of diverse biological resources has led to the identification of numerous bioactive molecules, particularly from plants, which have been integral to traditional and modern medicine. The vast range and complexity of natural products (NP) from plants, fungi, marine organisms, and microorganisms allow these compounds to effectively target biological macromolecules. Indeed, over 70% of the global population relies on plants and derivatives for healthcare. (2) One of the most widely studied NP families is alkaloid, particularly indole-alkaloids.

Many alkaloid-rich plants have a long history in folk medicine. (3) and this group stands out for its powerful biological activities as an antineoplastic agent (*e.g.* vinblastine, vincristine). Consequently, an alkaloids collection has been curated over the past 60 years by J. LeMen *et al.* in collaboration with the Institut de Chimie Moléculaire de Reims (ICMR – UMR 7312 CNRS, Reims), the Institut de Chimie des Substances Naturelles (ICSN – CNRS, Gif-sur-Yvette) and the Faculté de Pharmacie de Paris.

People recognize the intrinsic value of species, but growing evidence shows that the loss of biodiversity raises significant concerns among conservationists, governments, society, and policymakers.(4) As a matter of fact, biodiversity loss will profoundly impact ecosystem functions and diminishing availability of diverse and unique bioactive lead compounds essential for drug discovery. (4) Therefore, the Convention on Biological Diversity (CBD) has prioritized biodiversity conservation and management, particularly in developing countries, since its inception.(4)

Scientists need to turn biodiversity and conservation data into useful knowledge products. This will help improve communication within the scientific community. Making these products more accessible can fill gaps in our understanding related to space, species, and time. This increased understanding will lead to better outcomes for our knowledge of life.(5) Improvements are noteworthy in the quality and accessibility of biodiversity and conservation data over the past decade (*e.g.* World Database on Protected Areas (WDPA), IUCN Red List of Threatened Species, Red List of Ecosystems, *etc.*).

Novel approaches based modern bioinformatics and cheminformatics can benefit data management, analysis, and interpretation of data from current and older natural products research.(6–8) Therefore, high-resolution mass spectrometry and nuclear magnetic resonance spectrometry has already been used as digital scanners to capture the chemical fingerprints of profiled samples. (9) In addition, initiatives like the LOTUS project, NMReData and GNPS aim to create open-access platforms for experimental data sharing related to NP (*i.e.* structural and spectral data) (6,10,11). However, the lack of workflows that centralise in one place diverse natural product experimental is impeding scientific progress.(12) This issue arises from the reliance on non-standardized databases, which complicates data integration, accessibility, community curation, and interoperability across related fields. (6,13)

This study is establishing the prequel of the Reims Digital Collection of Alkaloids and aims to develop robust workflows for digitizing biological resources. The digitalization of this collection was inspired by two recent projects. The first is the MIADB (9) which aggregates Mono Indole Alkaloid mass spectra into a single database. The second is the Digital Botanical Gardens Initiative It ambitions to explore

innovative solutions for the collection, management and sharing of digital information acquired by mass spectrometric approaches on living botanical collections. (14)

The Reims Digital Collection of Alkaloids features MS/MS and 1D/2D NMR data from complex natural products, preserving established knowledge assisting in the identification of new related alkaloids within complex mixtures.(9). Additionally, this collection will support the demand for compounds in high-throughput screening. (15) Researchers could simultaneously use these databases to investigate *in silico* biological phenomena, as therapeutic development required mode of action studies and protein target discovery. For example, the function of the OAT enzyme associated with mitosis was identified using diazonamide A isolated from the marine ascidian *Diazona angulata*. (16)

Results

This work aimed to create the basics for preserving the collection of the 400 alkaloids samples stored in the Natural Product Chemistry laboratory of the University of Reims Champagne-Ardenne, by using modern analytical techniques (*i.e.* LC-HRMS/MS and 1D/2D NMR), and following FAIR recommendations. Chemoinformatic capabilities chemoinformatics were a great help in computer-based database management, curation (*i.e.* Topspin (17), MZMine (18), SIRIUS (19)), and visualization (*i.e.* Cytoscape (20)). These databases should primarily assist in dereplicating complex mixtures, as this process often relies on spectral comparison.

Locally developed NMR dereplication workflow, known as CaraMel (21), committed us to include in the NMR database only ^{13}C chemical shifts from pure compounds. The structure must have previously been fully elucidated *via* 2D NMR experiments (*i.e.* COSY, HSQC, HMBC) and every carbon signals must have been detected and attributed. When this database is next updated, we may consider including the 1D ^1H and/or 2D data that are already available. These considerations will be developed further below.

MS/MS spectral database consists of MS^2 spectra acquired with a soft ionisation source to provide spectra containing specific fragments and pseudo-molecular ion, following standardized parameters (*c.f.* Fig.1). Included MS^2 spectra must enable SIRIUS software to build a detailed fragmentation tree and to suggest the correct structure candidate. Consequently, the impact of different normalized collision energies (30, 50 and 70 eV) on the richness of the fragmentation patterns (*i.e.* MS^2 spectra) was evaluated (*c.f.* Fig.1). Collision energy equal to 50 % was the most appropriate for studying the alkaloids under study, as showed for voacangine sample ($m/z=369.22$, $[\text{M}+\text{H}]^+$). Low rate of fragmentation do not enable SIRIUS to build suitable fragmentation trees. Finally, metadata (*e.g.*

example, the spectra of brucine sample showed the product is 95% pure (*i.e.* ratio = 0.05) and consequently considered as pure product (c.f. Fig.2).

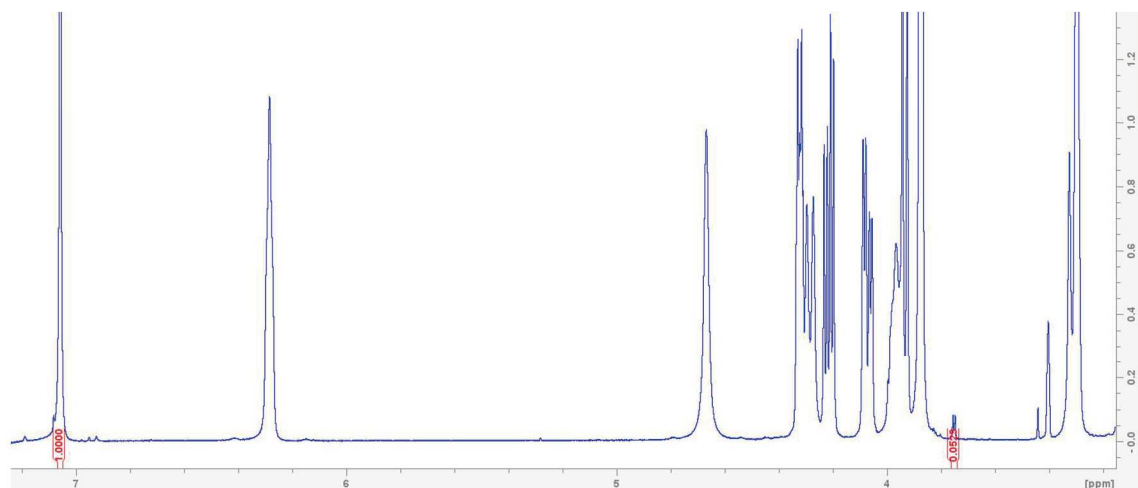


Figure 2: Focus on the ^1H spectrum of brucine sample. The signal of an aromatic proton (left) and a proton belonging to an impurity (right) have been integrated.

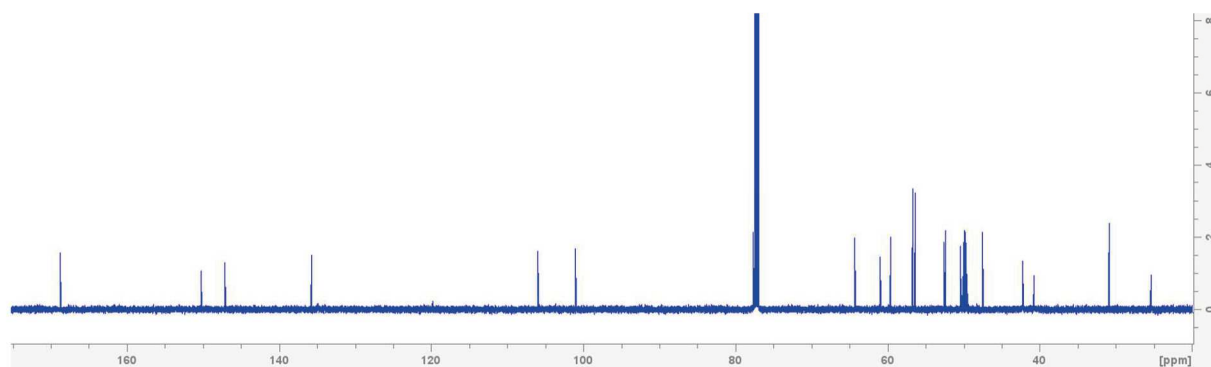


Figure 3: ^{13}C NMR spectrum from brucine (CDCl_3)

Purity is then confirmed if only one major peak is visualized on HPLC-HRMS² chromatogram (c.f. Fig.4), and its parent mass (m/z) corresponds to the molecular mass of the supposed structure (c.f. Fig.5) SIRIUS *de novo* annotation software (<https://boecker-lab.github.io/docs.sirius.github.io/>) was then used to confirm structure annotation. It must first generate a fragmentation tree, which will allow for the annotation of the correct structure from the structural database containing all the molecular structures analyzed in the alkaloid library. (c.f. Fig.6) Consequently, the compatibility with advanced

annotation tool (*i.e.* SIRIUS) allows experimental spectra to be included in training datasets for deep learning algorithms that can be developed for spectral prediction and other applications.

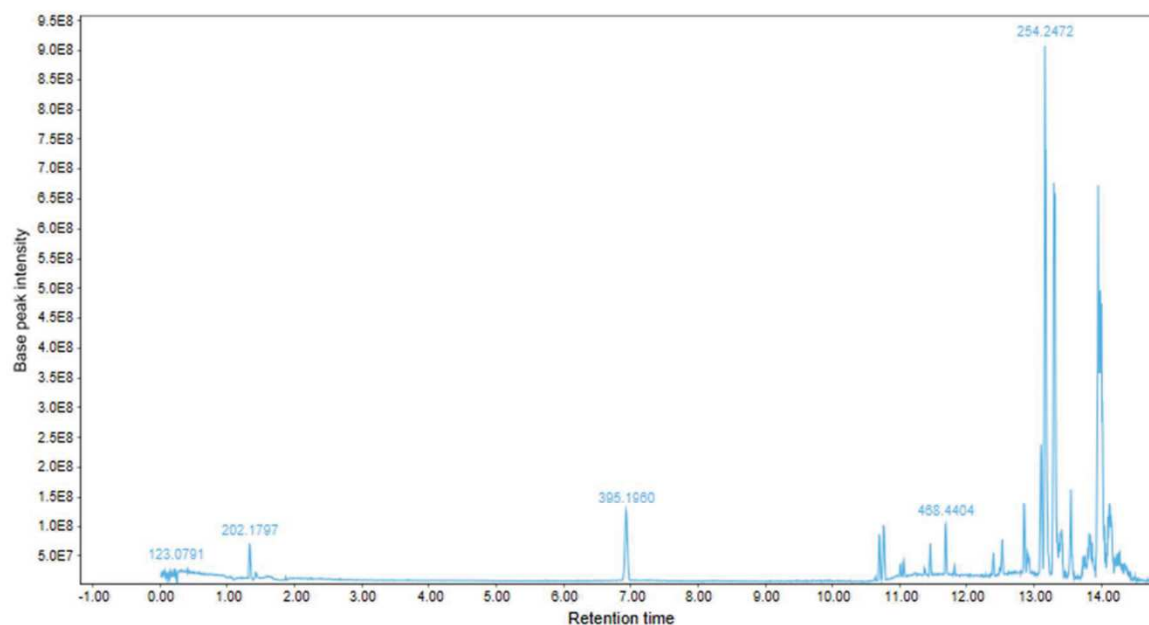


Figure 4: Total Ion Current chromatogram from brucine sample

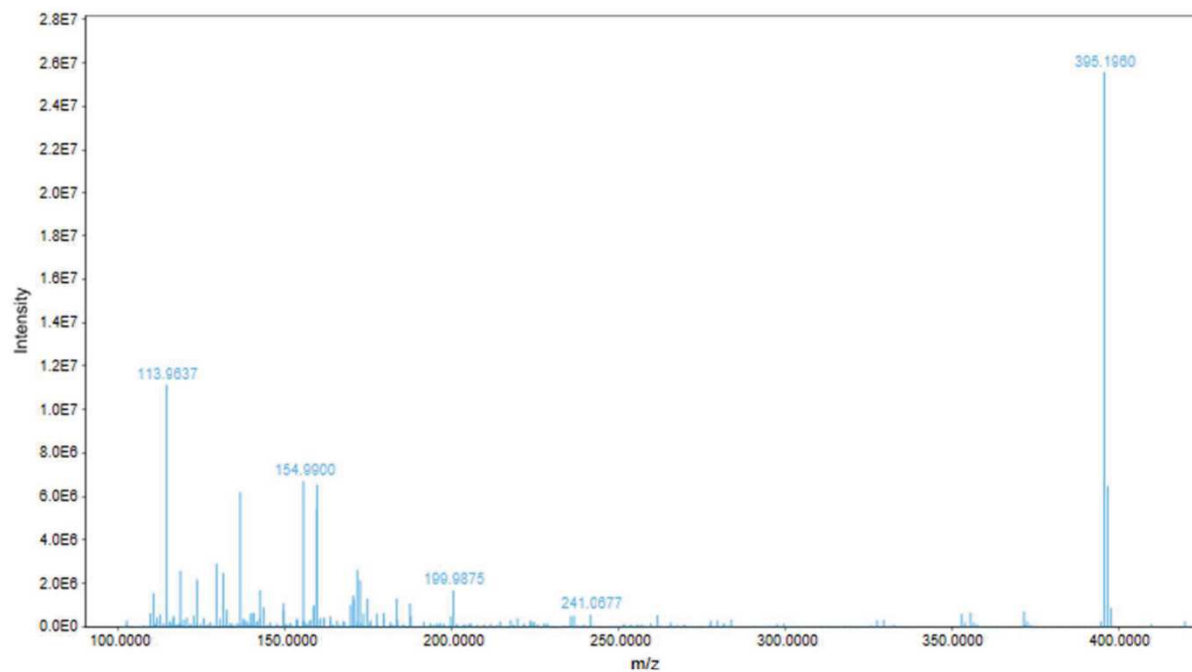


Figure 5: MS¹ spectrum from chromatogram peak of brucine ($m/z = 395.20$)

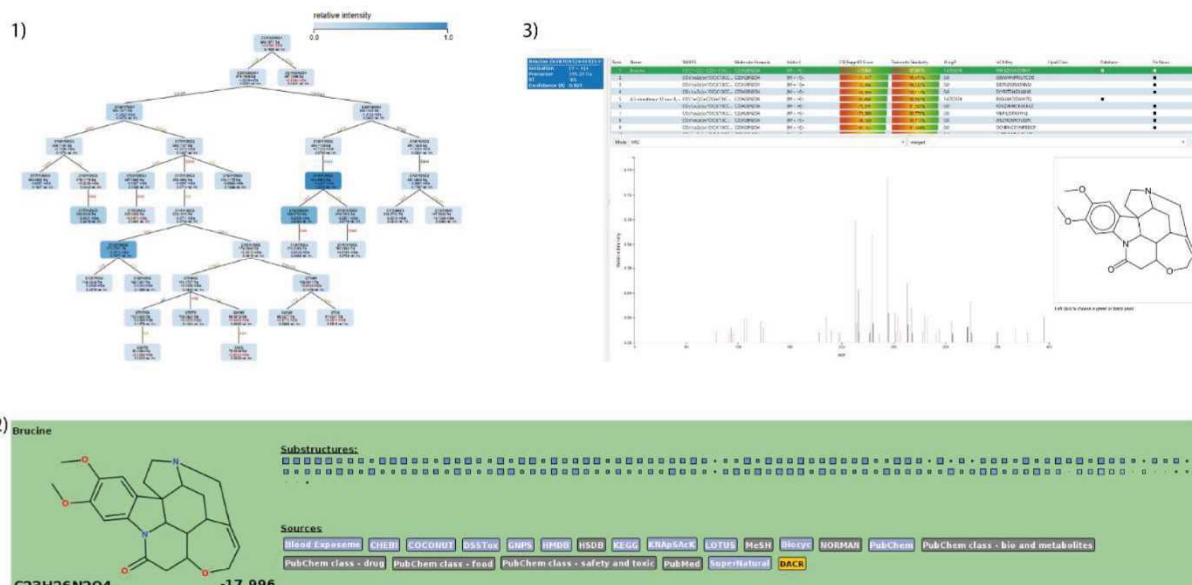
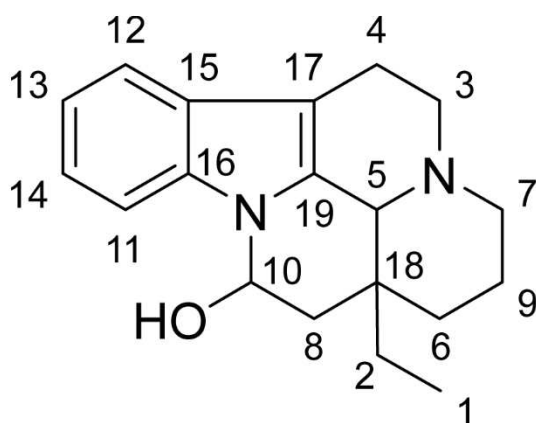


Figure 6: SIRIUS fragmentation tree prediction(1) and structure annotations (3) for brucine ion [M+H]⁺ (m/z = 395.20) using DACR structural database (highlighted in yellow in 2)

The second class, the mislabelled products, corresponds to samples, as for the dihydro-corynantheol (corynanthe alkaloid type), for which the purity was assessed via ¹H NMR, but a difference was noticed for parent mass between the experimental value (m/z=297.20, ESI⁺ mode) and the expected one (m/z=299.43, [M+H]⁺). In this case, 2D NMR spectra were acquired to elucidate the right structure, often one of those proposed during the annotation of the MS/MS spectrum by SIRIUS, as carbon signals

from ^{13}C NMR experiment were unable to be attributed to the supposed one. In this example, the correct structure turned out to be eburnamine, an aspidosperma alkaloid type (c.f. Fig.7).



Shift [ppm]	# H's	Correlation table
7.71	3	C1
16.69	2	C4
20.09	2	C9
24.68	2	C6
28.74	2	C2
37.27	0	C18
43.25	2	C8
44.36	2	C7
51.09	2	C3
59.16	1	C5
76.63	1	C10
105.28	0	C17
112.46	1	C11
118.22	1	C12
120.69	1	C13
122.12	1	C14
128.1	0	C15
136.58	0	C16

Figure 7: Eburnamine ^{13}C NMR chemical shifts assignment table

The last class, the mixed products, corresponds to samples where purity criteria were not achieved (*i.e.* low ^1H integration ratio, large ^1H signal if structures are close, not any major peak on HPLC-HRMS² chromatogram (c.f. Fig.8), several m/z ratio), as is the case for reserpiline sample. This class will require further processing if they are to be considered for inclusion in the NMR and MS/MS spectral databases.

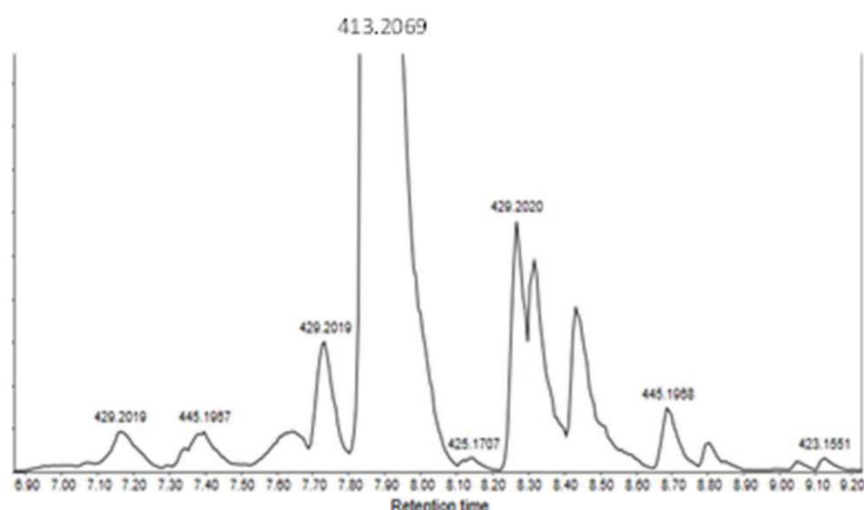


Figure 8: Total Ion Current chromatogram from reserpiline sample

However, when it comes to pure products, experimental spectra (MS² and ^{13}C) have been integrated into dedicated databases. For mislabelled products, the inclusion of ^{13}C spectral data in the database

was only considered if the correct molecular structure could be elucidated and if all the carbon signals could be detected and attributed to this new structure. If the MS² spectrum acquired from the same sample meets the necessary conditions (*i.e.* spectrum with structure-specific fragments, pseudomolecular ion is present, a fragmentation tree can be constructed by SIRIUS and, the structure of the molecule is included in the annotation proposals), then it can be included in the MS/MS spectral database. Finally, in the case of mixed products, only the MS² spectra which meet the above conditions can be included in the MS² spectral database, because the chromatographic dimension is sufficient to separate the different compounds. However, this is not the case for ¹³C NMR spectra, where a prior purification step will facilitate the detection and attribution of signals from each molecule.

Building DACR MS/MS database and ¹³C NMR database

To date, 75 unique alkaloid samples have been analysed by HPLC-HRMS² and NMR in parallel. Among these samples, 49 were included to MS/MS spectral database, while only 25 were included to the ¹³C NMR database.

MS/MS database visualization and comparison with MIADB

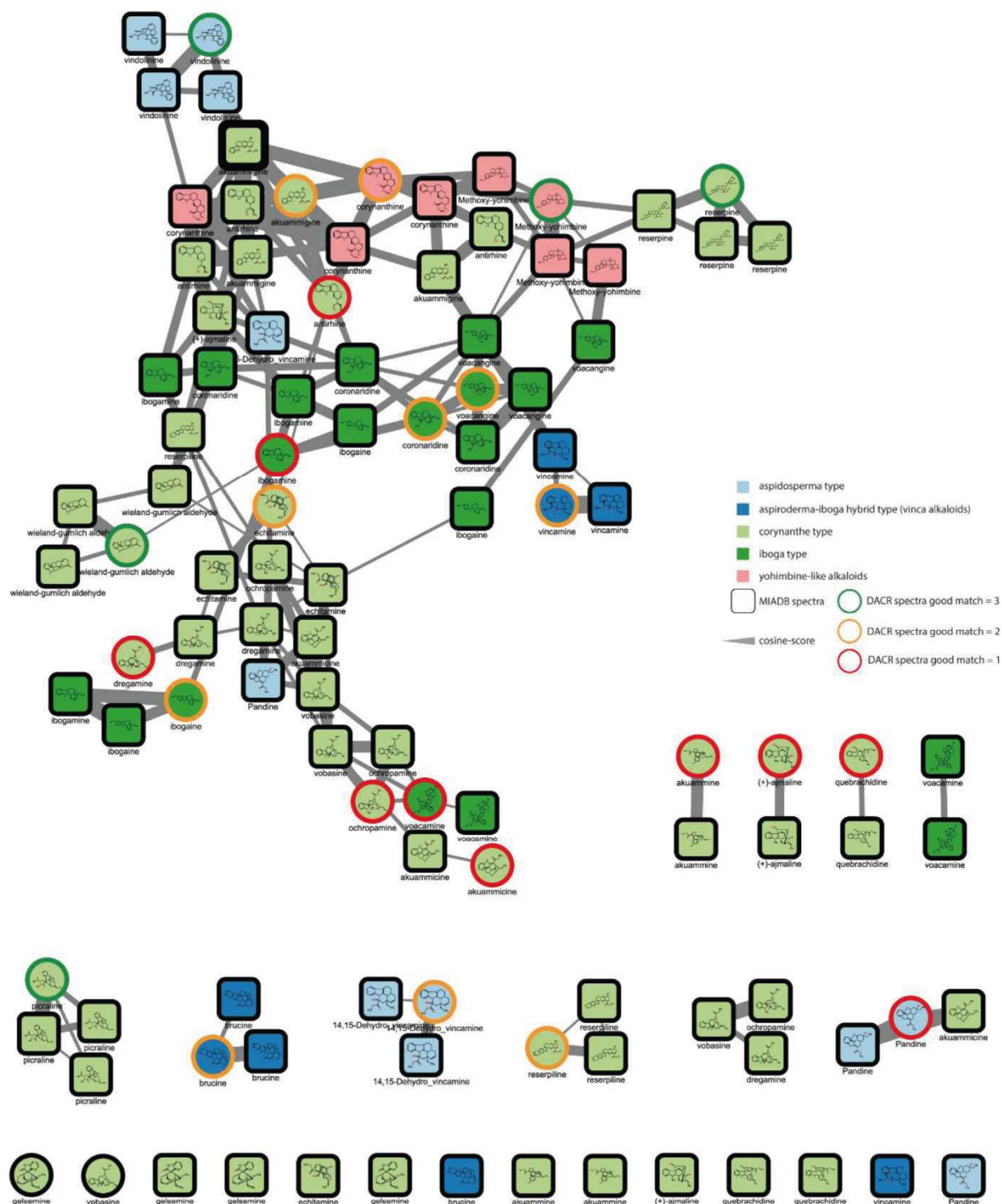
A Feature Based Molecular Network has been realized ($\cos > 0.6$) *via* MetGem software (<https://metgem.github.io/>) to organise and visualise DACR MS/MS spectral database, as shown in Fig.10. The 49 MS² spectra were distributed in 4 main alkaloid classes, according to NPClassifier ontology: 12 aspidosperma, 8 iboga, 3 aspidosperma-iboga hybrids (*i.e.* vinca alkaloid), 23 corynanthe and 3 yohimbine-like alkaloids.

DACR was then compared to the already known Mono-Indole Alkaloid DataBase (MIADB) MS² database. The aim of this approach was to assess how much this initial spectral database encompassed existing information in terms of data quantity and quality, as well as the extent to which it provided new insights. As a result, more than half of the compounds (27/49) contained in DACR had a spectrum that was also found in MIADB and are visualised by square on Fig.10. Consequently, the MS² spectra

for 22 compounds were only found in DACR, and most of them are corynanthe or aspidosperma types of alkaloids, which shown an initial distinction between our database and the MIADB.

The spectral matching algorithm from matchms package (<https://github.com/matchms/matchms>) was then used to compare the spectra from the two databases one by one for these 27 molecules, knowing that each molecule within MIADB has 3 fragmentation spectra corresponding to 3 collision energies applied. Two different cosine-score functions were finally applied: the classic cosine-score function and the modified cosine-score. Whatever the score used, almost three quarters of the molecules commonly found in DACR and MIADB have at least one spectrum that corresponds. Molecules that did not matched for both cosine-score function are: gelsemine, anthurine, which are 2 corynanthe alkaloids and voacamine, which is iboga type alkaloid. Spectral matching results can be found in SI.4 and SI.5. Spectral comparison between DACR and MIADB common molecules was also carried out by constructing a molecular network (cos-score >0.6, MetGem) as shown in Fig.9. The latter corroborates the previous observations. As visualized, 5 DACR spectra matched with 3 MIADB spectra for the same molecule (highlighted with green circles), 10 with 2 MIADB spectra for the same molecule (highlighted with

orange circles) and 10 with at least 1 MIADB spectra for the same molecule (highlighted with red circles).



209

210 *Figure 9: Feature Based Molecular Network of DACR-MIAD common molecules (cosine-score > 0.6)*

211

212 This work also highlighted limitations in how MS² spectra are organized and visualized. The feature-
213 Based Molecular Network, shown in Fig.10 organizes and visualizes the 49 DACR spectra based on their

spectral similarity (*i.e.* cosine score > 0.6, minimum matched-peaks of 4), following the principle that structurally related molecules fragment in similar ways. Additionally, metadata, such as their chemical family, have been mapped onto this network.

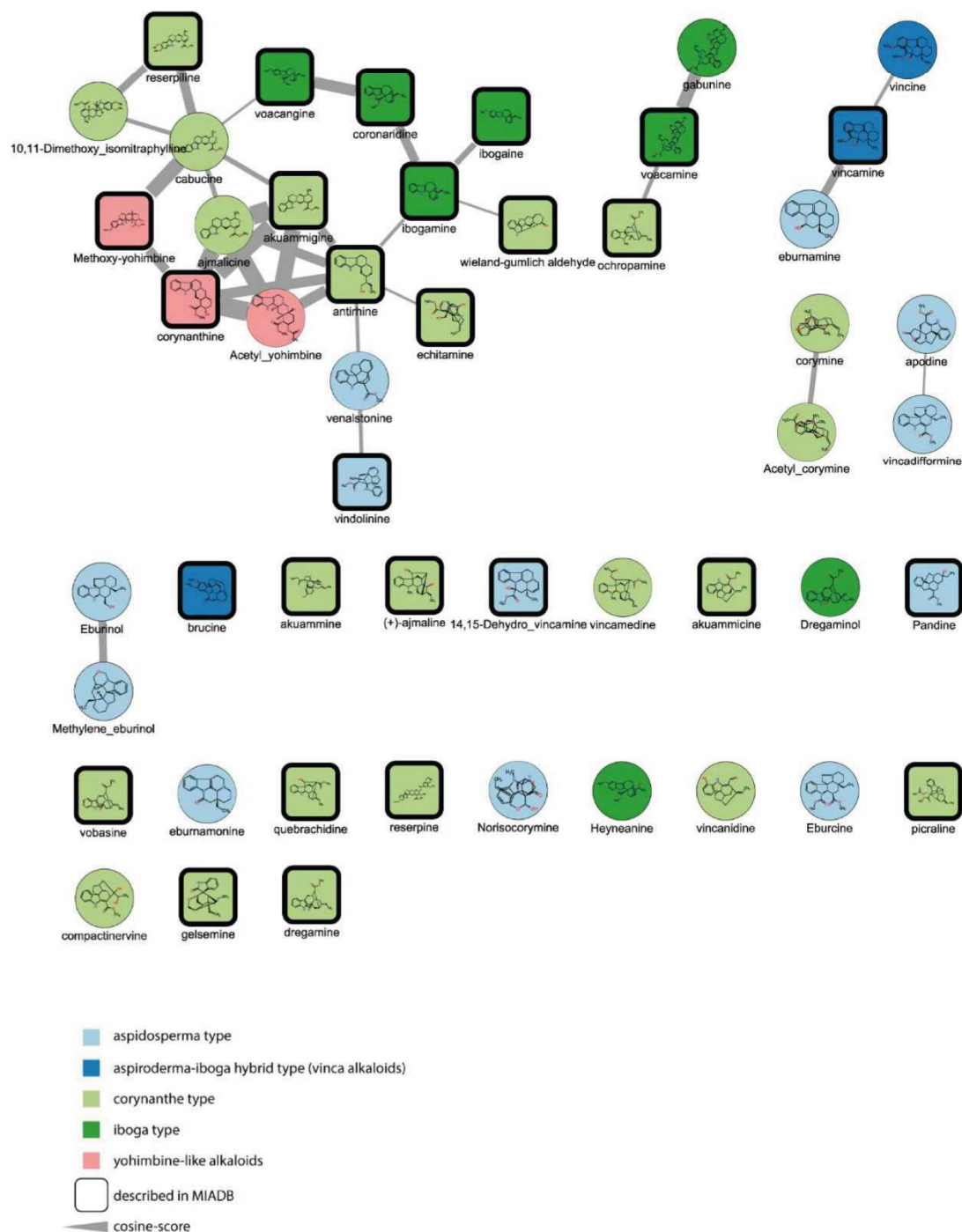


Figure 10: Feature Based Molecular Network of DACR spectra (cosine-score > 0.6)

However, it appears that spectra from molecules belonging to the same chemical family—specifically, aspidosperma alkaloids (*i.e.* eburnamine, eburinol, methylene eburinol, ebucine, and eburnammonine)—do not cluster into a single molecular family (except eburinol and methylene eburinol). Instead,

spectra are quite distinct (cf Fig.11). A t-SNE (t-distributed stochastic neighbor embedding) visualization is increasingly utilized in the context of natural substances. This method is designed to preserve the local structure of data, meaning the relationships between neighboring points within high-dimensional data. The MS data complexity reduction allow for clearer visualization of chemical similarities. Points that are close to each other in the reduced space are likely also close in the original high-dimensional space, indicating chemical similarity. Clusters observed in a t-SNE visualization can reveal underlying alkaloid sub-classes, categories, or patterns in fragmentation. (22,23) For this reason, we realized t-SNE from DACR spectra in comparison to MIADB spectra. Results are shown in SI.3 and confirm the similarity of our spectra with those of MIADB, and to a lesser extent their ability to discriminate between alkaloid subclasses. One more time, spectra from eburinol, methylene-eburinol, eburcine, eburnamine and

eburnamonine did not cluster together although these molecules belong to the same subclass of alkaloids (*i.e.* aspidosperma type).

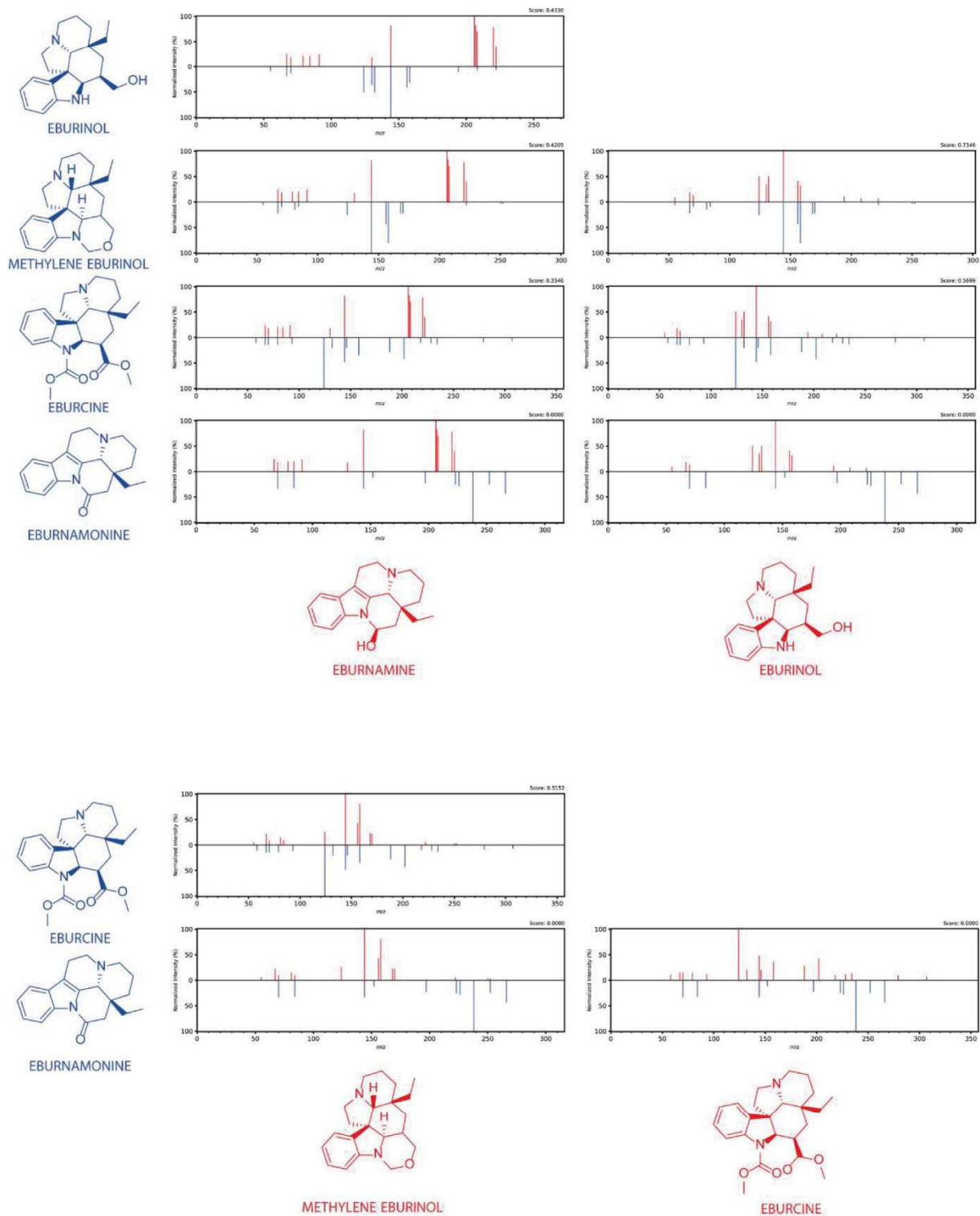


Figure 11: Spectral comparison for a selection of aspidosperma alkaloids

These molecules logically share similar structures, as confirmed by structural comparison using the Tanimoto score (*i.e.* no score below 0.5, c.f Fig.12). This suggests that, despite their structural similarity, these molecules have distinct fragmentation pathways.

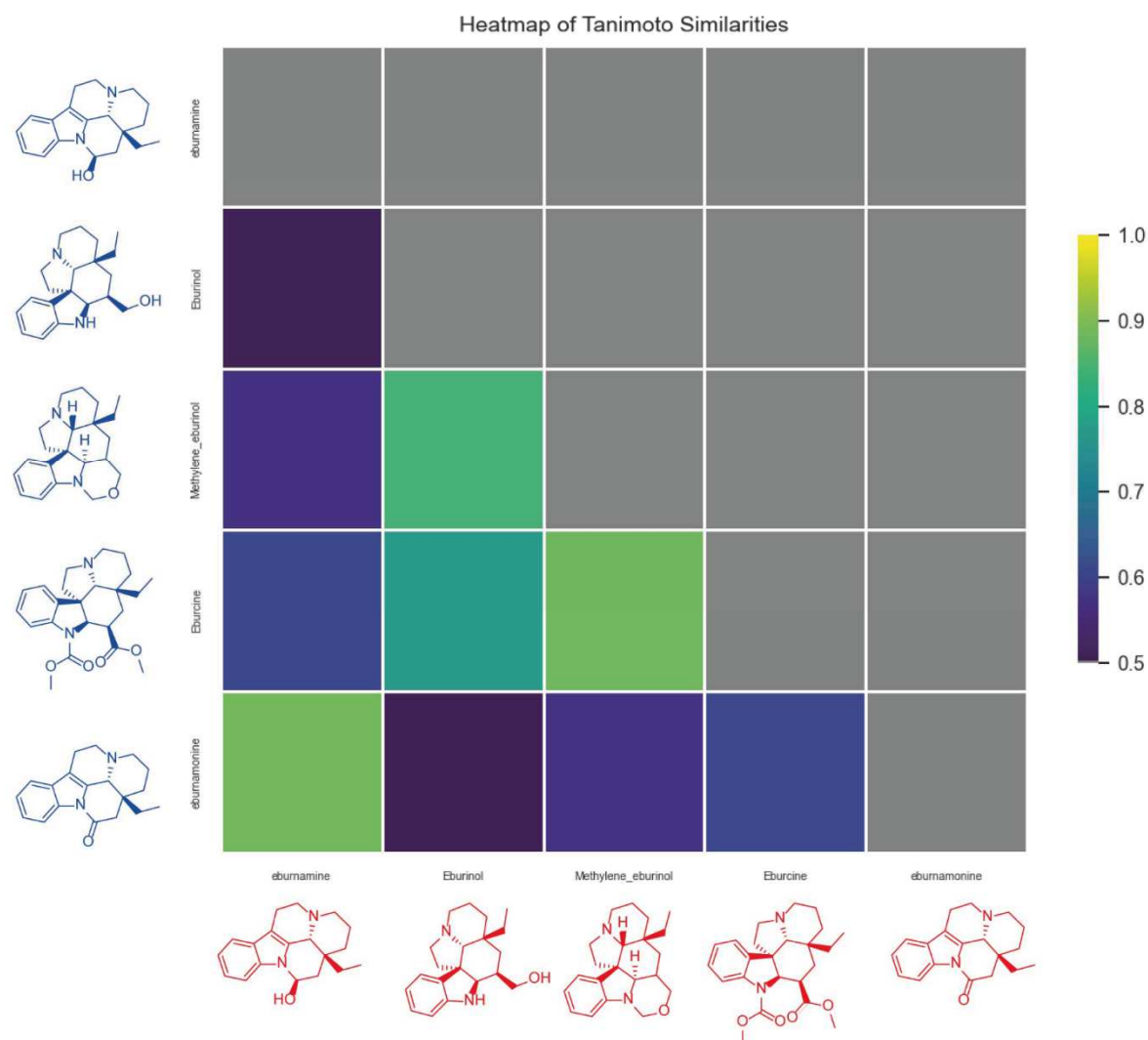


Figure 7: Tanimoto comparison for a selection of aspidosperma alkaloids

¹³C NMR database validation

the validation of the ¹³C NMR database has not yet been carried out in terms of spectral comparison with another database, but this could be considered using a reference database such as NP-MRD. However, due to the unique nature of certain molecules in the DACR, the lack of reference spectra would make this step impossible and, therefore, unnecessary. In any case, NMR data validation poses fewer issues than mass spectrometry, given the universality of this analytical technique. The main

limitation currently lies in the lack of standardization in acquisition conditions, particularly in terms of solvent choice for sample dissolution. Indeed, ^1H and ^{13}C chemical shifts shift depending on the solvent used, as demonstrated by the overlay of two ^1H and ^{13}C spectra of yohimbine in different solvents (*i.e.* DMSO- d_6 in blue, MeOD in red) (cf. Fig.13 and 14). Additionally, a singlet is observed at 10.75 ppm on the blue ^1H spectrum but is absent on the red spectrum. This difference arises because, unlike MeOD, there is no proton exchange between the molecule and DMSO- d_6 does. Therefore, signals corresponding to amine or alcohol groups appear in spectra acquired using DMSO- d_6 , unlike with MeOD. However, because DMSO- d_6 is more viscous than MeOD, the signals obtained with DMSO- d_6 are slightly broader than those observed with MeOD.

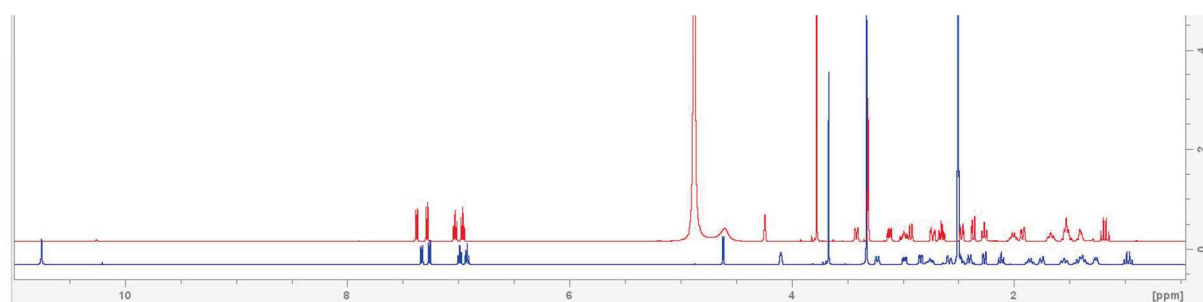


Figure 13: ^1H NMR spectra from yohimbine in DMSO- d_6 (blue) and MeOD (red)

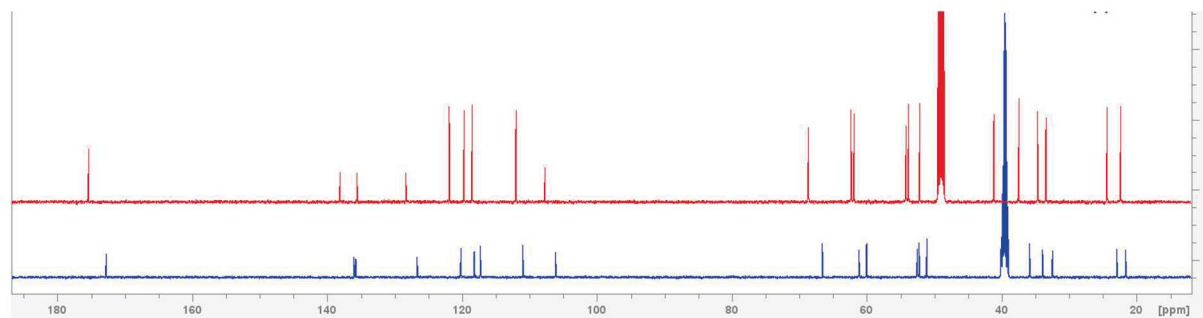


Figure 14: ^{13}C NMR spectrum from yohimbine in DMSO- d_6 (blue) and MeOD (red)

Compliance with FAIR principles

Firstly, the Reims Digital Collection of Alkaloids meet the Findable and Available criteria, as both MS/MS, ^{13}C NMR spectral databases and metadata will be uploaded on the French recherche.data.gouv plateforme. Therefore, data (*i.e.* spectra) and metadata (acquisition parameters, *etc.*) will be preserved in a secure data center following TRUST principles, will be associated to a DOI and access will be free of charge. Secondly, the formats chosen for the MS/MS and ^{13}C NMR spectral databases will be respectively *.mgf (Mascott Generic Format) and .SDF (Structure Data File), to enable exchanges between computer systems (*i.e.* interoperable) and increase the capacity of the metadata to be

274 combined. Thirdly, the databases built are intended to be as reusable as possible. This is why every
275 study should provide as much information as possible about the source of the data, including
276 standardization of the acquisition parameter (c.f material & methods section) and, on the other, the
277 conditions for reuse.

278

279

Conclusion & Discussion

This study is paving the way to the future Digital Alkaloids Collection from Reims (DACR) spectral (^{13}C NMR and HPLC- MS^2) databases by following as closely as possible the FAIR recommendations. The purity of 75 alkaloid samples from the physical collection was first assessed *via* chromatographic (HPLC) and spectroscopic methods (^1H and ^{13}C NMR, MS^1 , MS^2). Three different categories were consequently obtained: pure products, mislabelled products, and mixed products. Only spectra from pure products were included, in accordance with data availability, to ^{13}C NMR and/or MS^2 spectral libraries. To date, 49 among 75 different alkaloid samples were included to MS^2 spectral database. In contrast, 25 alkaloid samples provided a decent quality ^{13}C NMR spectrum to be included in ^{13}C NMR spectral database.

Several of the samples analyzed were qualified as “mislabelled-products”. This kind of error may be due to the operator having entered the wrong product name on the jar at the time. It may also be a product resulting from the degradation of the original product. However, this would mean that all the product in the jar reacted in the same way, which is statistically unlikely. In any case, these samples provided spectra of sufficient quality to be included in spectral databases. It is more likely in the “mixed products” category that the deterioration in sample quality may have played a role over the time, working on chemical functions that are more fragile than others (*e.g.* esters, *etc.*). In this case, further isolation and purification steps are required. These product mixtures are not a drawback in terms of knowledge, as the resulting chemical diversity will result in numerous distinct spectra that can be added to the databases.

During validation of the MS^2 spectral database, 27 compounds showed MS^2 spectra commonly found between DACR and MIADB. The comparison highlighted that $\frac{3}{4}$ of these compounds had at least one spectral matching, assuming the validity of acquired spectra. However, it is important to realise the spectra that constitute these two spectral databases do not come from the same spectrometer (*i.e.* Q-ToF for MIADB, Q-Orbitrap for DACR). The key difference between these two devices is the technology used in the collision cell, respectively hexapole collision cell and HCD (Higher energy collision dissociation). While hexapole cells are primarily used for interference reduction and ion storage, HCD is more focused on enhancing fragmentation efficiency for complex analytes. (24) In contrast, while hexapole collision cells excel in interference management and ion storage, HCD provides a more versatile approach to fragmentation, particularly beneficial for analysing complex biomolecules. Indeed, hexapole collision cells serves to enhance ion fragmentation and reduce interferences, distinguishing it from simple collision-induced dissociation (CID) methods. However, as high amount of energy distribution of ions is deposited on the ions in a short time by using HCD, the internal energy

distribution of ions is expected to be sharply truncated, giving access to higher energy pathways, plus consecutive decomposition.(25,26) Therefore, hexapole collision cells and HCD will provide different fragmentation pathways, however they complement each other. This makes it essential to consider collision cell technology in dereplication using tandem mass spectrometry, as kinetic shifts are crucial when the study aims to compare dissociation pathways (*i.e.* MS² spectra). The comparison between DACR spectra and MIADB here proposed should therefore make it possible to judge whether the collision cell technology influences the matching performance in dereplication workflow. Preliminary results did not allow a hypothesis to be validated. One might think that DACR MS² spectra for molecules with no match, irrespective of the cosine score function, wouldn't contain enough peaks (between 15 and 69 peaks). However, vocangin and ibogaine, which have only 21 and 23 peaks on their respective spectrum, still match up to >0.9 regardless of the function chosen. Acquiring MS² spectral data for the same molecule across instruments with different collision cells can nonetheless help reduce analytical inconsistencies and expand the chemical space covered. Collision cell technology also appears not to influence spectral mismatches between alkaloid subclasses (*e.g.* 1 corynanthe, 2 iboga, 1 aspidosperma types) as described in SI.3 and SI.4.

In contrast, there are 5 mismatches between alkaloids of the same family and 4 between alkaloids of different families using the "modified cosine" function, and 3 mismatches between different families and 1 between the same family for cosine. In relation to the number of errors, cosine score function seems to make more mismatches between different families than modified cosine score function, which suggests a possible influence of the cosine score function chosen to compare the spectra. But beware, this is a very small sample size to draw such conclusions, and the results must be treated with caution.

Finally, the comparison of DACR MS² spectra against those from MIADB also highlights the importance of obtaining a set of spectra for the same molecule at multiple collision energies, regardless of the collision cell technology. Indeed, certain collision energies do not provide common fragmentation pattern of the same molecule, regarding MIADB (*cf.* Fig.9), as for example gelsemine with 3 spectra that do not form a single cluster. This should allow more spectral diversity to be acquired from the same molecule, and more spectral matches to be generated probabilistically during the dereplication workflow. Therefore, given the arguments regarding the differences in experimental conditions, the DACR MS² spectra are not intended to compete with those of the MIADB, but rather to complement them. It could further benefit from acquiring MS² spectra in ESI- mode, which is also well-suited for alkaloids. Similarly, on the NMR side, the collected 2D NMR data could contribute to expanding DACR NMR database. Indeed, 2D NMR data play a crucial role in metabolite annotation within complex mixtures by providing structural insights that aid spectral comparison. Tools like MADByTE (27) and

SMART (28) harness 2D NMR, specifically HSQC and HMBC spectra. The first one allows to map and differentiate chemical compositions within mixtures, enabling rapid identification of compound families or unique structures without purified samples. The second one (*i.e.* SMART - *Small Molecule Accurate Recognition technology*) includes convolutional neural networks (CNNs) algorithms to enhance comparison capabilities. However, challenges persist in developing comprehensive 2D NMR databases due to the need for accurate spin system identification (29), and creating robust spectral libraries. (30,31) Advancements in automated analysis and deep learning techniques offer promising avenues for improving the accuracy and efficiency of 2D NMR data processing, potentially transforming the field

Material & Method

Purification, conversion of an alkaloid from salt to base form

Some mixed product samples were included in the alkaloid sample collection. For this reason, they underwent prior purification on silica column. Approximately 10 mg of sample were dissolved in CHCl_3 and placed on column. The eluent phase consisted of $\text{MeOH}/\text{CHCl}_3$ (v/v: 90/10). Thin layer chromatography was realized on each fraction to determine which should be considered.

In addition, a few samples are preserved in salty form and therefore unsolvable in organic solvents, unlike basic forms. Consequently, 20 mg of salty form samples was dissolved in 5 ml NH_3 3%. The mixture was then placed in separating funnel with 5 ml CH_2Cl_2 . The organic phase was dried by adding Na_2SO_4 . Finally, the organic phase was filtered and evaporated, to obtain basic form of alkaloid.

Data acquisition & database constitution

NMR

All samples were not analysed using the same acquisition and processing parameters (*i.e.* solvent, spectral window, number of scans, *etc.*), partly due to solubility issues. The solubility of aliquots of 5 mg for each alkaloid sample was first assessed in 600 μl CDCl_3 (Innova-Chem, Angervilliers, France), or else in MeOD (Sigma-Aldrich, Missouri, USA), or even in $\text{DMSO}-d_6$ (Eurisotop, Saint-Aubin, France). CDCl_3 has been given priority because of its low cost, inert nature and, its ease of evaporation makes it easier to recover the sample. Soluble samples were then analysed by nuclear magnetic resonance (^1H , ^{13}C , HSQC, HMBC, HMBC ^{15}N - ^1H , COSY, NOESY) at 298 K on a Bruker Avance AVIII-600 MHz spectrometer (Karlsruhe, Germany) equipped with a 5 mm TCI cryoprobe (^1H 600.16 MHz; ^{13}C 150.9 MHz), or on a Bruker Avance Neo 500 MHz spectrometer (Karlsruhe, Germany) equipped with a SP BB&19F/1H probe

(¹H 500.28 MHz; ¹³C 125.8 MHz). Table. 1 describe analysis parameters for each compound of DACR ¹³C NMR database.

Table 1: Parameters for ¹³C NMR spectra acquisition from alkaloids included in DACR13C NMR database

compound name	carbon acquisition frequency	solvent
gelsemine	150	CDCl ₃
echitamine	125	DMSO- <i>d</i> ₆
brucine	150	CDCl ₃
eburnamine	125	CDCl ₃
quebrachidine	150	CDCl ₃
vincamine	125	CDCl ₃
norisocorymine	125	CDCl ₃
ajmalicine	125	CDCl ₃
methylene eburninol	150	CDCl ₃
methoxy yohimbine	125	DMSO- <i>d</i> ₆
venalstonine	125	MeOD
compactonervine	150	CDCl ₃
coronaridine	150	CDCl ₃
apodine	150	CDCl ₃
vincamedine	125	CDCl ₃
akuamicine	125	CDCl ₃
vincadifformine	150	CDCl ₃
eburnamonine	150	CDCl ₃
ibogaine	150	CDCl ₃
voacangine	150	CDCl ₃
strychnine	150	CDCl ₃
corynantheidal	125	DMSO- <i>d</i> ₆
deacetyl vindoline	125	CDCl ₃
epi vindolinine	150	CDCl ₃
yohimbine	125	DMSO- <i>d</i> ₆

The names and SMILES of molecules whose ¹³C spectra met the above conditions for inclusion in the current ¹³C NMR database were first compiled in a *.csv file. Structure Data File (*.sdf) was then created using VersaDB software developed in ICMR, Reims.(32) This file contains MolBlock (*i.e.* a way to compute 2D+ coordinates for a molecule) linked to predicted ¹³C chemical shifts and metadata. To do this, the *.csv file was first placed in LOTUS_input folder path of VersaDB software, and « Predict ¹³C NMR chemical Shifts » function was applied. As a result, the file « 13C_NMR_Database.sdf » was obtained. Finally, the predicted ¹³C chemical shift were replaced by experimental ones.

All samples were analyzed using the same acquisition and processing parameters, at 1 μ M concentration in HPLC-MS grade MeOH (Carlo Erba Reagents, Val de Reuil, France). HPLC analyses were performed with Thermoscientific Vanquish, equipped with an UPTISPHERE Strategy C18 column (2.2 μ m x 150 mm x 2.1 μ m, Interchim). The eluent consisted of H₂O + 0.1% formic acid (A) and MeCN (B), following a gradient 5-60 % B in 7 min, then 60-100 % B in 1 min, then then maintaining 100 % B for 8 min at a flow rate of 0.4 ml/min.

Mass data were acquired with a ThermoFisher Scientific QExactive Plus (Orbitrap) mass spectrometer. The electrospray ionisation source was set as follow: positive mode, source temperature 380°C, capillary 4.10 kV, desolvation temperature 400°C, nebuliser gas flow 5 Bar, desolvation gas flow 700 l/h. MS scans were performed in full-scan mode from m/z 100 to 1500 (scan time 0.1 sec) with a resolution of 35 000 (FWHM). An MS¹ scan was followed by MS² scans of the three most intense ions above a threshold of 8000 counts (exclusion window 3 sec). The selected parent ions were fragmented according to normalized collision energy 50 % (c.f. Fig.1). Prior to the data acquisition, the Orbitrap was calibrated for mass accuracy using Pierce™ LTQ Velos ESI Positive Ion Calibration Solution (mix of cafein, MRFA, Ultramark 1621 and n-butylamine in ACN, MEOH, Acetic acid) (Product number 88323, Thermo Scientific). The raw MS² files were finally processed with MZMine (version 4.0.8). A list of individual compounds (*i.e.* feature) for each sample was generated from a manual MS/MS data mining process on every single file. Parameters were not normalised. The spectra obtained were then submitted to SIRIUS (version 5.8.6) to predict fragmentation trees and to obtain a list of candidate structures. MS² spectra that met all the conditions mentioned above (c.f. results section) were manually compiled in a *.mgf file. Each MS² spectrum was associated with metadata using HEADER descriptors: name of molecule, InChI, InChIKey, isomeric SMILES generated by PubChem, IUPAC name, NPClassifier ontology, *etc.*

Reference

1. Bebber DP, Wood JRI, Barker C, Scotland RW. Author inflation masks global capacity for species discovery in flowering plants. *New Phytologist*. 2014 Jan;201(2):700–6.
2. Newman DJ, Cragg GM. Natural Products as Sources of New Drugs over the Nearly Four Decades from 01/1981 to 09/2019. Vol. 83, *Journal of Natural Products*. American Chemical Society; 2020. p. 770–803.
3. Tarkowská D. A fast and reliable UHPLC-MS/MS-based method for screening selected pharmacologically significant natural plant Indole alkaloids. *Molecules*. 2020 Jul 1;25(14).
4. McCay SD, Lacher TE. National level use of International Union for Conservation of Nature knowledge products in American National Biodiversity Strategies and Action Plans and National Reports to the Convention on Biological Diversity. Vol. 3, *Conservation Science and Practice*. Blackwell Publishing Inc.; 2021.
5. Weatherdon L V., Appeltans W, Bowles-Newark N, Brooks TM, Davis FE, Despot-Belmonte K, et al. Blueprints of effective biodiversity and conservation knowledge

446 products that support marine policy. Vol. 4, *Frontiers in Marine Science*. Frontiers Media
 447 S. A; 2017.

448 6. Rutz A, Sorokina M, Galgonek J, Mietchen D, Willighagen E, Gaudry A, et al. The LOTUS
 449 initiative for open knowledge management in natural products research. *Elife*. 2022 May
 450 1;11.

451 7. Jarmusch AK, Wang M, Aceves CM, Advani RS, Aguirre S, Aksenov AA, et al. ReDU: a
 452 framework to find and reanalyze public mass spectrometry data. *Nat Methods*. 2020 Sep
 453 1;17(9):901–4.

454 8. Mongia M, Mohimani H. Repository scale classification and decomposition of tandem
 455 mass spectral data. *Sci Rep*. 2021 Dec 1;11(1).

456 9. Fox Ramos AE, Le Pogam P, Fox Alcover C, Otego N’Nang E, Cauchie G, Hazni H, et al.
 457 Collected mass spectrometry data on monoterpene indole alkaloids from natural
 458 product chemistry research. *Sci Data*. 2019 Dec 1;6(1).

459 10. Pupier M, Nuzillard JM, Wist J, Schlörer NE, Kuhn S, Erdelyi M, et al. NMReDATA, a
 460 standard to report the NMR assignment and parameters of organic compounds.
 461 *Magnetic Resonance in Chemistry*. 2018 Aug 1;56(8):703–15.

462 11. Wang M, Carver JJ, Phelan V V., Sanchez LM, Garg N, Peng Y, et al. Sharing and community
 463 curation of mass spectrometry data with Global Natural Products Social Molecular
 464 Networking. *Nat Biotechnol*. 2016 Sep 8;34(8):828–37.

465 12. Ballesti S, Mäs M, Helbing D. On disciplinary fragmentation and scientific progress. *PLoS*
 466 *One*. 2015 Mar 19;10(3).

467 13. Van Santen JA, Jacob G, Singh AL, Aniebok V, Balunas MJ, Bunsko D, et al. The Natural
 468 Products Atlas: An Open Access Knowledge Base for Microbial Natural Products
 469 Discovery. *ACS Cent Sci*. 2019 Nov 27;5(11):1824–33.

470 14. The Digital Botanical Gardens Initiative [Internet]. [cited 2024 Oct 28]. Available from:
 471 <https://www.dbgi.org/>

472 15. Kingston DGI. Modern natural products drug discovery and its relevance to biodiversity
 473 conservation. Vol. 74, *Journal of Natural Products*. 2011. p. 496–511.

474 16. Carlson EE. Natural products as chemical probes. Vol. 5, *ACS Chemical Biology*. 2010. p.
 475 639–53.

17. TopSpin | NMR Data Analysis | Bruker [Internet]. [cited 2024 Oct 28]. Available from: https://www.bruker.com/en/products-and-solutions/mr/nmr-software/topspin.html?utm_source=Advertising&utm_medium=GoogleAd&utm_campaign=BBIO-Software-Cross-All-TopSpin-H2-2024&gad_source=1&gclid=Cj0KCQjw7Py4BhCbARIsAMMx-_l3i3KyxpO9pnxkdMHAnGcOY77dE9-l36b9xmXyMaq2zVbinTRGnsYaAoAAEALw_wcB
18. Schmid R, Heuckeroth S, Korf A, Smirnov A, Myers O, Dyrland TS, et al. Integrative analysis of multimodal mass spectrometry data in MZmine 3. Vol. 41, Nature Biotechnology. Nature Research; 2023. p. 447–9.
19. Dührkop K, Fleischauer M, Ludwig M, Aksenov AA, Melnik A V., Meusel M, et al. SIRIUS 4: a rapid tool for turning tandem mass spectra into metabolite structure information. Nat Methods [Internet]. 2019 Apr 1 [cited 2021 Jun 17];16(4):299–302. Available from: <https://www.nature.com/articles/s41592-019-0344-8>
20. Create a network graph with Cytoscape – Digital Humanities 201 [Internet]. [cited 2023 Jan 12]. Available from: <http://miriamposner.com/classes/dh201w19/tutorials-guides/network-analysis/create-a-network-graph-with-cytoscape/>
21. Hubert J, Nuzillard JM, Purson S, Hamzaoui M, Borie N, Reynaud R, et al. Identification of natural metabolites in mixture: A pattern recognition strategy based on ¹³C NMR. Anal Chem [Internet]. 2014 Mar 18 [cited 2021 Jun 18];86(6):2955–62. Available from: <https://pubs.acs.org/sharingguidelines>
22. Andronov MG, Fedorov M V, Sosnin S. Exploring Chemical Reaction Space With Reaction Difference Fingerprints and Parametric t-SNE.
23. Olivon F, Elie N, Grelier G, Roussi F, Litaudon M, Touboul D. MetGem Software for the Generation of Molecular Networks Based on the t-SNE Algorithm. Anal Chem. 2018 Dec 4;90(23):13900–8.
24. Remes PM, Schwartz JC, Mcalister G, Horner J, Zhang T, Kiyonami R, et al. Higher-Energy Collisional Dissociation in the Existing High Pressure Multipole Region of a Stand-Alone Ion Trap Mass Spectrometer. JC US Patent. 1999;6(2):474–6.
25. Bayat P, Lesage D, Cole RB. TUTORIAL: ION ACTIVATION IN TANDEM MASS SPECTROMETRY USING ULTRA-HIGH RESOLUTION INSTRUMENTATION. Mass Spectrom

Rev [Internet]. 2020 Sep 1 [cited 2024 Oct 27];39(5–6):680–702. Available from:
<https://onlinelibrary.wiley.com/doi/full/10.1002/mas.21623>

26. Comstock K, Huang Y. Multiple Fragmentation Methods for Small Molecule Characterization on a Dual Pressure Linear Ion Trap Orbitrap Hybrid Mass Spectrometer.
27. Egan JM, van Santen JA, Liu DY, Linington RG. Development of an NMR-Based Platform for the Direct Structural Annotation of Complex Natural Products Mixtures. Vol. 84, *Journal of Natural Products*. American Chemical Society; 2021. p. 1044–55.
28. Reher R, Kim HW, Zhang C, Mao HH, Wang M, Nothias LF, et al. A Convolutional Neural Network-Based Approach for the Rapid Annotation of Molecularly Diverse Natural Products. *J Am Chem Soc*. 2020 Mar 4;142(9):4114–20.
29. Sgouralis I, Akwataghibe K, Mebrat M, Van Horn WD. Fully automated analysis of NMR data with Bayesian nonparametric statistical learning. *Biophys J*. 2023 Feb;122(3):141a.
30. Shchukina A, Kasprzak P, Urbańczyk M, Kazimierczuk K. The Non-uniform Sampling of 2D NMR Data. In: *Fast 2D Solution-state NMR*. The Royal Society of Chemistry; 2023. p. 115–53.
31. Chen B, Fang Z, Zhang Y, Guan X, Lin E, Feng H, et al. Two-Dimensional Laplace NMR Reconstruction through Deep Learning Enhancement. *J Am Chem Soc*. 2024 Aug 7;146(31):21591–9.
32. Cordonnier J, Remy S, Renault J, Nuzillard J. Versa DB: Assisting ¹³C NMR and MS/MS Joint Data Annotation Through On-Demand Databases. *Chemistry–Methods* [Internet]. 2024 Jan 17;4(1). Available from: <https://chemistry-europe.onlinelibrary.wiley.com/doi/10.1002/cmt.202300020>

530 Supplementary Information. 1.

Spectrum ID for DACR	
ID	Molecule
0	Wieland-gumlich_aldehyde
1	Gelsemine
2	Echitamine
3	Dregamine
4	Brucine
5	Vindolinine
6	Akuammine
7	Ajmaline
8	14,15-Dehydro_vincamine
9	Antirhine
10	Vobasine
11	Vincamine
12	Reserpiline
13	Quebrachidine
14	Ibogamine
15	Ochropamine
16	Picraline
17	Methoxy-yohimbine
18	Corynanthine
19	Voacamine
20	Coronaridine
21	Akuammigine
22	Akuammicine
23	Pandine
24	Ibogaine
25	Voacangine
26	Reserpine

531

532

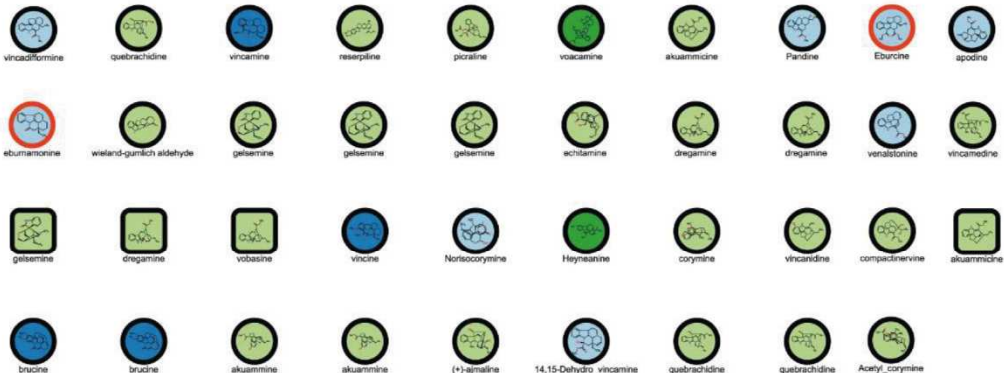
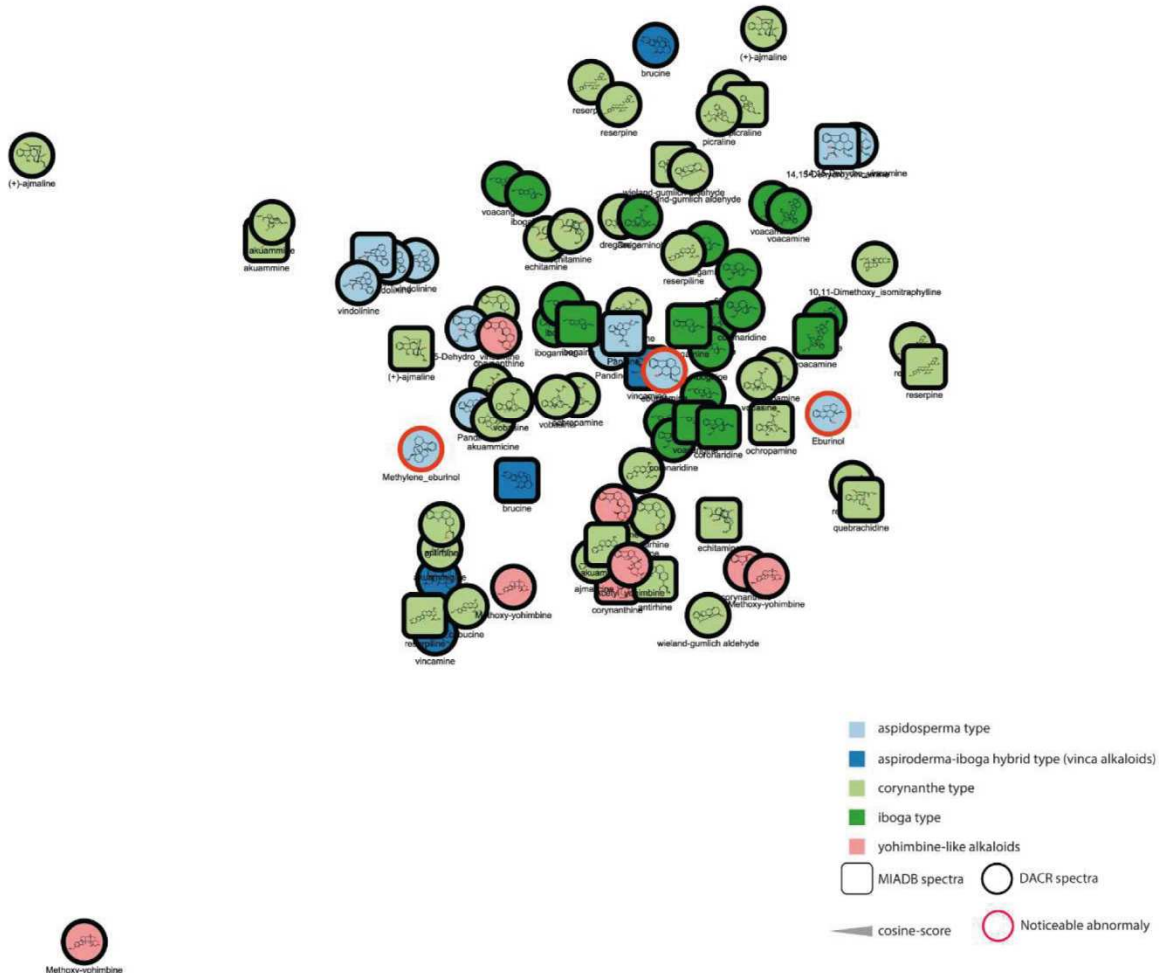
533 Supplementary Information. 2.

Spectrum ID for MIADB		Spectrum ID for MIADB	
ID	Molecule		
0	Wieland-Gumlich aldehyde	21	Ajmaline
1	Wieland-Gumlich aldehyde	22	Ajmaline
2	Wieland-Gumlich aldehyde	23	Ajmaline
3	Gelsemine	24	14,15-dehydrovincamine
4	Gelsemine	25	14,15-dehydrovincamine
5	Gelsemine	26	14,15-dehydrovincamine
6	Echitamine	27	Antirrhine
7	Echitamine	28	Antirrhine
8	Echitamine	29	Antirrhine
9	Dregamine	30	Vobasine
10	Dregamine	31	Vobasine
11	Dregamine	32	Vobasine
12	Brucine	33	Vincamine
13	Brucine	34	Vincamine
14	Brucine	35	Vincamine
15	Vindoline	36	Reserpiline
16	Vindoline	37	Reserpiline
17	Vindoline	38	Reserpiline
18	akuammine	39	Quebrachidine
19	akuammine	40	Quebrachidine
20	akuammine		

534

Spectrum ID for MIADB		Spectrum ID for MIADB	
41	Quebrachidine	61	Coronaridine
42	Ibogamine	62	Coronaridine
43	Ibogamine	63	Akuammigine
44	Ibogamine	64	Akuammigine
45	Ochropamine	65	Akuammigine
46	Ochropamine	66	Akuammicine
47	Ochropamine	67	Akuammicine
48	picraline	68	Akuammicine
49	picraline	69	Pandine
50	picraline	70	Pandine
51	11-methoxyyohimbine	71	Pandine
52	11-methoxyyohimbine	72	Ibogaine
53	11-methoxyyohimbine	73	Ibogaine
54	Corynanthine	74	Ibogaine
55	Corynanthine	75	Voacangine
56	Corynanthine	76	Voacangine
57	Voacamine	77	Voacangine
58	Voacamine	78	reserpine
59	Voacamine	79	reserpine
60	Coronaridine	80	reserpine

535



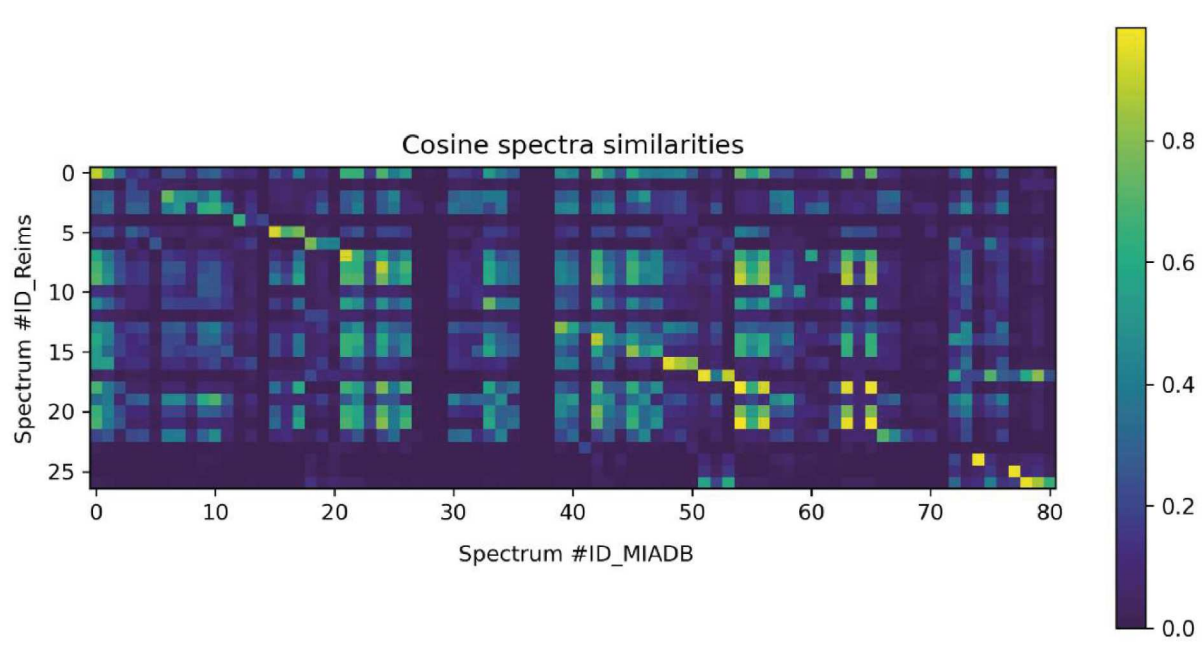
537

538

539

540

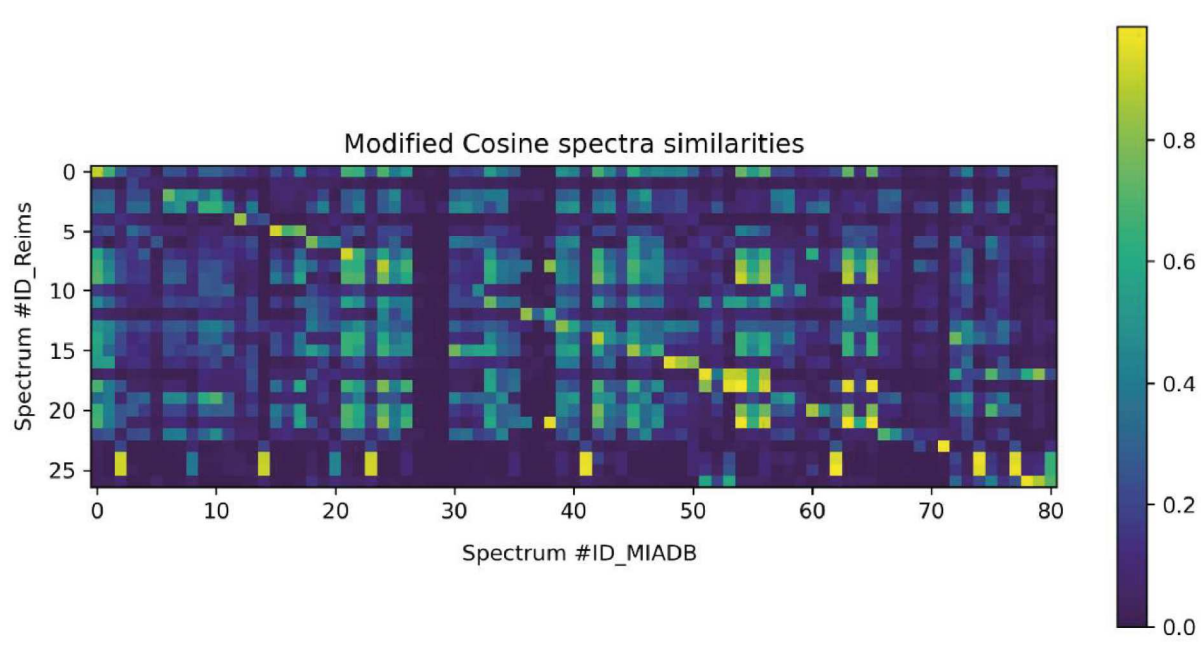
Supplementary Information. 4.



Spectral comparison heatmap (cosine-score) between DACR and MIADB spectra

20 of the 27 molecule spectra found in DACR matched at least one of the 3 spectra of the corresponding molecule in MIADB, using classic cosine-score ($\cos \geq 6$, min matched peaks = 6). 11 molecule spectra from DACR matched 1 time with a spectrum from corresponding molecule in MIADB, 6 molecule spectra from DACR matched 2 times, and 3 molecule spectra from DACR matched 3 times. There were only 4 wrong mismatches at the cosine-score level fixed.

Supplementary Information. 5.



Spectral comparison heatmap (modified cosine-score) between DACR and MIADB spectra

In contrast, 24 molecule spectra from DACR match at least one time with a spectrum of the corresponding molecule in MIADB using the modified cosine-score function ($\cos \geq 6$, min matched peaks = 6), 14 molecule spectra from DACR matched 1 time with a spectrum from corresponding molecule in MIADB, 6 molecule spectra from DACR matched 2 times, and 4 molecule spectra from DACR matched 3 times. There were more wrong mismatches (number=7) at the modified cosine-score level fixed.



Partie 3 : Conclusion & Discussion



III. Conclusion & Discussion

La sauvegarde des connaissances dans des bases de données est un élément essentiel des politiques de financement de la recherche, en raison de ses impacts majeurs. De nombreux domaines scientifiques autrefois très théoriques, comme la chimie des substances naturelles, qui étudie les molécules d'organismes vivants (*i.e.* science basée sur des objets physiques), se transforment désormais en sciences axées sur les données et objets numériques. Cette dématérialisation en phytochimie est notamment facilitée par des outils dérivés de la métabolomique.

Les universités et les laboratoires de recherche qui y sont rattachés ont toujours été des centres dynamiques d'innovation, de collaboration et de découvertes scientifiques et techniques. De nombreuses données sont ainsi utilisées ou produites pour soutenir l'effort de recherche. Notamment, le besoin de connaissances et de données expérimentales pour entraîner des outils d'intelligence artificielle (comme le machine learning et le deep learning) est plus que jamais un enjeu crucial. (104,105) A titre d'exemple, la pandémie de COVID-19 a mis en évidence le besoin urgent pour les chercheurs d'accéder à des ensembles de données pour le développement rapide de vaccins et de traitements, et a montré à quel point le partage de données en temps opportun peut amener à des avancées significatives. (106)

Les bases de données sont donc aujourd'hui la pierre angulaire de notre approche en recherche. Elles ont pour objectifs de rassembler, capturer, stocker, et archiver toutes les données produites dans le cadre de projets et d'expériences scientifiques, non seulement pour la publication des résultats, mais aussi pour faciliter leur partage et leur réutilisation. Bien que les avantages de conserver les connaissances dans des bases de données soient évidents, des défis tels que les préoccupations en matière de confidentialité et les pratiques de gestion des données qui varient selon les disciplines peuvent freiner une mise en œuvre efficace. (107) Des concepts organisationnels pour la gestion des données de recherche doivent donc être mis en place. L'Open Science et les principes FAIR – rendant les données Findable, Accessible, Interoperable, et Reusable – s'imposent comme une base commune, favorisant la sauvegarde du savoir produit, sa réutilisation pour amener à de nouvelles découvertes. (106,108)

Comme nous l'avons vu, les réseaux moléculaires offrent un moyen relativement efficace de gérer et visualiser les données spectrales MS². Cependant, l'exploration de grands ensembles de données reste difficile avec les seuls logiciels de visualisation (*e.g.* Cytoscape). Les limites d'exploration sont rapidement atteintes lorsqu'on tente de superposer plusieurs couches d'informations supplémentaires (*e.g.* bioactivité, origines taxonomiques, classes chimiques, *etc.*). (109) Ce constat que

nous pouvons étendre aux données issues de la RMN, nous amènent à chercher une approche plus globale et versatile pour la préservation des données.

La solution récemment proposée par Arnaud Gaudry *et al.*, nommée « Experimental Natural Products Knowledge Graph (ENPKG) », introduit une nouvelle approche, passant d'une focalisation sur les ensembles de données spectrales à une approche centrée sur les échantillons associés à des informations chimiques et une multitude de métadonnées pertinentes. Ainsi, les données spectrales, accompagnées de métadonnées pour chaque échantillon étudié, sont d'abord standardisées puis intégrées dans un graphe de connaissances unique (Knowledge Graph – KG) à l'image d'une mémoire collective. Comme le détaille l'article *A Sample-Centric and Knowledge-Driven Computational Framework for Natural Products Drug Discovery*, le KG offre un avantage significatif sur les bases de données spectrales classiques en pouvant stocker, organiser et relier des données expérimentales complexes et hétérogènes à d'autres connaissances disponibles. Cette architecture permet une gestion et une interprétation plus fluides des informations. En fin de compte, l'utilisation des graphes de connaissances devrait permettre aux chercheurs de contextualiser plus rapidement leurs recherches, accélérant ainsi le processus de découverte. (109)

Des anomalies de regroupement ont été observées entre les spectres MS² des alcaloïdes suivants : éburinol, méthylène-éburinol, éburcine, éburnamine et éburnamonine (type aspidosperma). Comme indiqué dans l'article précédent, ces composés ont des structures proches (avec un score de Tanimoto d'au moins 0,5). Des fragments communs à faible rapport m/z dans les schémas de fragmentation de ces cinq molécules sont pourtant retrouvés, notamment les fragments à 144,08 m/z, 124,11 m/z et 70,7 m/z, que SIRIUS attribue à la fragmentation du noyau indole. Leur nombre est néanmoins trop faible pour établir une corrélation entre les nœuds des spectres correspondants dans les réseaux moléculaires (nombre minimum de pics correspondant = 4, cosine-score = 0,6). Des voies de fragmentation communes, comme la réaction rétro Diels-Alder déjà observée pour les alcaloïdes de type aspidosperma, peuvent toutefois être relevées. (110) Or, cette réaction génère des motifs de fragmentation qui sont influencés par les substituants des cycles aromatiques et la nature des groupes fonctionnels présents. (111) La difficulté à regrouper les sous-classes d'alcaloïdes selon leur spectre MS² semble donc liée aux variations des groupes fonctionnels et substituants propres à chaque molécule (*i.e.* éburinol avec le méthylène-éburinol et l'éburcine, ou encore l'éburnamine et l'éburnamonine). Les fragments de rapport m/z important et directement issus de l'ion précurseur résultent de la réaction de rétro-Diels-Alder, et auront des m/z différents en raison des légères différences de structure de l'ion précurseur. Ces sous-fragments se fragmentent à leur tour, générant finalement les sous-sous-fragments à 144,08 m/z, 124,11 m/z et 70,06 m/z caractéristiques de la fragmentation du noyau indole.





Références



1. Farnsworth NR. INTERNATIONAL PERSPECTIVES REGARDING THE USE OF FOOD/NATURAL PRODUCTS AS DRUGS. Vol. 21, Drug Informarion Journal. 1987.
2. Costello MJ, Houlding B, Joppa LN. Further evidence of more taxonomists discovering new species, and that most species have been named: Response to Bebbet et al. (2014). Vol. 202, New Phytologist. Blackwell Publishing Ltd; 2014. p. 739–40.
3. Agosti D, Benichou L, Penev L, Hyam R. A Possible Workflow from New and Legacy Publications to keep the World Flora Online up to date with New Species and Augmenting Taxonomic Treatments. Biodiversity Information Science and Standards. 2022 Aug 23;6.
4. Rutz A, Sorokina M, Galgonek J, Mietchen D, Willighagen E, Gaudry A, et al. The LOTUS initiative for open knowledge management in natural products research. Elife. 2022 May 1;11.
5. Minev BR. Cancer Management in Man: Chemotherapy, Biological Therapy, Hyperthermia and Supporting Measures (Cancer Growth and Progression, Volume 13) [Internet]. 1970. Available from: <http://www.springer.com/series/5721>
6. Erwan P, Sabrina B, Corine G, Thierry H, Elisabeth S. Pharmacognosie - Obtention et propriétés des substances actives médicamenteuses d'origine naturelle. Elsevier Masson, editor. 2020.
7. ALCALOÏDES - Universalis Edu [Internet]. [cited 2024 Oct 25]. Available from: <https://www.universalis-edu.com/encyclopedie/alcaloides/15-interet-therapeutique/>
8. Bruneton J. Pharmacognosie - phytochimie - plantes médicinales (5^e éd.). Tec & Doc Lavoisier. Paris; 2016.
9. Fournier J. Alkaloids discovery, markers for the history of organic chemistry. Rev Hist Pharm (Paris). 2001;49(331):315–32.
10. Taille Du Marché, Part, Demande Et Moteurs De Croissance Des Médicaments à Base D'Alcaloïdes D'Ici 2030 [Internet]. [cited 2024 Oct 25]. Available from: <https://www.databridgemarketresearch.com/fr/reports/global-alkaloid-based-drugs-market>
11. Bézanger-Beauquesne L. Les alcaloïdes dans les plantes. Bulletin de la Societe Botanique de France [Internet]. 1958 [cited 2024 Oct 25];105(5–6):266–91. Available from: <https://www.tandfonline.com/doi/abs/10.1080/00378941.1958.10837879>
12. Addoms RM. Fifty Years of Plant Physiology. Th. Weevers. Amsterdam, Holland: Scheltema & Holkema's, 1949. 308 pp. Science (1979) [Internet]. 1950 Apr 21 [cited 2024 Oct 25];111(2886):427–427. Available from: <https://www.science.org/doi/10.1126/science.111.2886.427.b>
13. Trease and Evans Pharmacognosy.
14. 508 Br&ves COInnmnieations-Kurze Mitteilungen EXPEaIENTIA XXI[9].
15. Mohammed AE, Abdul-Hameed ZH, Alotaibi MO, Bawakid NO, Sobahi TR, Abdel-Lateff A, et al. molecules Chemical Diversity and Bioactivities of Monoterpene Indole Alkaloids (MIAs) from Six Apocynaceae Genera. 2021 [cited 2024 Oct 26]; Available from: <https://doi.org/10.3390/molecules26020488>
16. Noble RL. The discovery of the vinca alkaloids--chemotherapeutic agents against cancer. Biochem Cell Biol. 1990 Dec;68(12):1344–51.

17. Nicole Langlois, Yves Langlois. La découverte de la Navelbine: histoire vécue. Langlois, editor. 2015.
18. Massiot G, Zeches M, Mirand C, Men-Olivier L Le, Delaude C, Baser KHC, et al. Occurrence of Longicaudatine, a New Type of Bis-Indole Base and Bisnor-C-alkaloid H in Strychnos Species. *Journal of Organic Chemistry* [Internet]. 1983 [cited 2024 Oct 26];48(11):1869–72. Available from: <https://pubs.acs.org/doi/abs/10.1021/jo00159a015>
19. Les enjeux et les perspectives de l'épigénétique dans le domaine de la santé [Internet]. [cited 2023 Aug 23]. Available from: <https://www.senat.fr/rap/r16-033-1/r16-033-1.html#fnref11>
20. LeRobert. Isoler [Internet]. [cited 2023 Jul 17]. Available from: <https://dictionnaire.lerobert.com/definition/isoler>
21. Salem MA, De Souza LP, Serag A, Fernie AR, Farag MA, Ezzat SM, et al. Metabolomics in the context of plant natural products research: From sample preparation to metabolite analysis. Vol. 10, *Metabolites*. MDPI AG; 2020.
22. LeRobert. Extraire [Internet]. [cited 2023 Jul 17]. Available from: <https://dictionnaire.lerobert.com/definition/extraire>
23. Wolfender JL, Nuzillard JM, Van Der Hooft JJJ, Renault JH, Bertrand S. Accelerating Metabolite Identification in Natural Product Research: Toward an Ideal Combination of Liquid Chromatography-High-Resolution Tandem Mass Spectrometry and NMR Profiling, in Silico Databases, and Chemometrics. *Anal Chem*. 2019;91(1):704–42.
24. van der Hooft JJJ, de Vos RCH, Ridder L, Vervoort J, Bino RJ. Structural elucidation of low abundant metabolites in complex sample matrices. Vol. 9, *Metabolomics*. 2013. p. 1009–18.
25. Beutler JA, Alvarado AB, Schaufelberger DE, Andrews P, McCloud TG. Dereplication of phorbol bioactives: *Lyngbya majuscula* and *croton cuneatus*. *J Nat Prod* [Internet]. 1990 Jul [cited 2019 Nov 25];53(4):867–74. Available from: <https://pubs.acs.org/doi/abs/10.1021/np50070a014>
26. Nothias LF, Petras D, Schmid R, Dührkop K, Rainer J, Sarvepalli A, et al. Feature-based Molecular Networking in the GNPS Analysis Environment. *bioRxiv*. 2019 Oct;812404.
27. Wang M, Carver JJ, Phelan V V., Sanchez LM, Garg N, Peng Y, et al. Sharing and community curation of mass spectrometry data with GNPS. *Nat Biotechnol*. 2017;34(8):828–37.
28. Ward JL, Baker JM, Miller SJ, Deborde C, Maucourt M, Biais B, et al. An inter-laboratory comparison demonstrates that [1H]-NMR metabolite fingerprinting is a robust technique for collaborative plant metabolomic data collection. *Metabolomics*. 2010 Jun;6(2):263–73.
29. Lianza M, Leroy R, Rodrigues CM, Borie N, Sayagh C, Remy S, et al. The three pillars of natural product dereplication. Alkaloids from the bulbs of *urceolina peruviana* (c. presl) j.f. macbr. as a preliminary test case. *Molecules*. 2021 Feb 1;26(3).
30. Pupier M, Nuzillard JM, Wist J, Schlörer NE, Kuhn S, Erdelyi M, et al. NMReDATA, a standard to report the NMR assignment and parameters of organic compounds. *Magnetic Resonance in Chemistry*. 2018 Aug 1;56(8):703–15.
31. Letertre MPM, Dervilly G, Giraudeau P. Combined Nuclear Magnetic Resonance Spectroscopy and Mass Spectrometry Approaches for Metabolomics. Vol. 93, *Analytical Chemistry*. American Chemical Society; 2021. p. 500–18.



32. Bourderioux A, Bourjot M, Lordel-Madeleine S, Valois L. Tout en fiches - Mémo visuel de chimie analytique. DUNOD. Vols. 978-2-10-080165-7. 2020.
33. Bayat P, Lesage D, Cole RB. TUTORIAL: ION ACTIVATION IN TANDEM MASS SPECTROMETRY USING ULTRA-HIGH RESOLUTION INSTRUMENTATION. *Mass Spectrom Rev* [Internet]. 2020 Sep 1 [cited 2024 Oct 27];39(5-6):680-702. Available from: <https://onlinelibrary.wiley.com/doi/full/10.1002/mas.21623>
34. Comstock K, Huang Y. Multiple Fragmentation Methods for Small Molecule Characterization on a Dual Pressure Linear Ion Trap Orbitrap Hybrid Mass Spectrometer,.
35. Allwood JW, Goodacre R. An introduction to liquid chromatography-mass spectrometry instrumentation applied in plant metabolomic analyses. Vol. 21, *Phytochemical Analysis*. 2010. p. 33-47.
36. Fenaille F, Barbier Saint-Hilaire P, Rousseau K, Junot C. Data acquisition workflows in liquid chromatography coupled to high resolution mass spectrometry-based metabolomics: Where do we stand? Vol. 1526, *Journal of Chromatography A*. Elsevier B.V.; 2017. p. 1-12.
37. Raterink RJ, Lindenburg PW, Vreeken RJ, Ramautar R, Hankemeier T. Recent developments in sample-pretreatment techniques for mass spectrometry-based metabolomics. Vol. 61, *TrAC - Trends in Analytical Chemistry*. Elsevier B.V.; 2014. p. 157-67.
38. Wolfender JL, Nuzillard JM, van der Hooft JJJ, Renault JH, Bertrand S. Accelerating Metabolite Identification in Natural Product Research: Toward an Ideal Combination of Liquid Chromatography-High-Resolution Tandem Mass Spectrometry and NMR Profiling, in Silico Databases, and Chemometrics. *Anal Chem* [Internet]. 2019 Jan 2 [cited 2019 Nov 26];91(1):704-42. Available from: <https://pubs.acs.org/doi/10.1021/acs.analchem.8b05112>
39. Demuth W, Karlovits M, Varmuza K. Spectral similarity versus structural similarity: Mass spectrometry. *Anal Chim Acta*. 2004 Jul 19;516(1-2):75-85.
40. Frank AM, Bandeira N, Shen Z, Tanner S, Briggs SP, Smith RD, et al. Clustering millions of tandem mass spectra. *J Proteome Res*. 2008 Jan;7(1):113-22.
41. Cai Y, Zhou Z, Zhu ZJ. Advanced analytical and informatic strategies for metabolite annotation in untargeted metabolomics. Vol. 158, *TrAC - Trends in Analytical Chemistry*. Elsevier B.V.; 2023.
42. Stein SE, Scott DR. Optimization and Testing of Mass Spectral Library Search Algorithms for Compound Identification.
43. Yang JY, Sanchez LM, Rath CM, Liu X, Boudreau PD, Bruns N, et al. Molecular networking as a dereplication strategy. *J Nat Prod*. 2013 Sep 27;76(9):1686-99.
44. Perez De Souza L, Alseekh S, Brotman Y, Fernie AR. Network-based strategies in metabolomics data analysis and interpretation: from molecular networking to biological interpretation. Vol. 17, *Expert Review of Proteomics*. Taylor and Francis Ltd; 2020. p. 243-55.
45. Wan KX, Vidavsky I, Gross ML. Comparing similar spectra: From similarity index to spectral contrast angle. *J Am Soc Mass Spectrom* [Internet]. 2002 Jan [cited 2019 Dec 12];13(1):85-8. Available from: <https://pubs.acs.org/doi/10.1021/jasms.8b01687>
46. Colavizza G, Hrynaskiewicz I, Staden I, Whitaker K, McGillivray B. The citation advantage of linking publications to research data. *PLoS One* [Internet]. 2020 Apr 1 [cited 2024 Oct

- 27];15(4):e0230416. Available from: <https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0230416>
47. Union PO of the E. Cost-benefit analysis for FAIR research data : cost of not having FAIR research data. 2019 Jan 16 [cited 2024 Oct 27]; Available from: <https://op.europa.eu/en/publication-detail/-/publication/d375368c-1a0a-11e9-8d04-01aa75ed71a1/language-en>
48. Fox Ramos AE, Evanno L, Poupon E, Champy P, Benididir MA. Natural products targeting strategies involving molecular networking: Different manners, one goal. Vol. 36, Natural Product Reports. Royal Society of Chemistry; 2019. p. 960–80.
49. Wilkinson MD, Dumontier M, Aalbersberg IJ, Appleton G, Axton M, Baak A, et al. Comment: The FAIR Guiding Principles for scientific data management and stewardship. Sci Data. 2016 Mar 15;3.
50. Lin D, Crabtree J, Dillo I, Downs RR, Edmunds R, Giaretta D, et al. The TRUST Principles for digital repositories. Vol. 7, Scientific Data. Nature Research; 2020.
51. Plan de gestion de données – DoRANum [Internet]. [cited 2024 Oct 27]. Available from: <https://dorum.fr/plan-gestion-donnees-dmp/>
52. Ouvrir la Science - Recherche Data Gouv : plateforme nationale fédérée des données de la recherche [Internet]. [cited 2024 Oct 27]. Available from: <https://www.ouvrirlascience.fr/recherche-data-gouv-plateforme-nationale-federee-des-donnees-de-la-recherche/>
53. Pourquoi Recherche Data Gouv ? [Internet]. [cited 2024 Oct 27]. Available from: <https://recherche.data.gouv.fr/fr/page/pourquoi-recherche-data-gouv>
54. Implication du CNRS dans Recherche Data Gouv - Science ouverte [Internet]. [cited 2024 Oct 27]. Available from: <https://www.science-ouverte.cnrs.fr/le-cnrs-dans-recherche-data-gouv/>
55. Lancement de la plateforme nationale Recherche Data Gouv | Science ouverte [Internet]. [cited 2024 Oct 27]. Available from: <https://scienceouverte.univ-rennes.fr/actualites/lancement-de-la-plateforme-nationale-recherche-data-gouv>
56. Dedieu L. Formation Rédiger un Data paper.
57. Amazon. Qu'est-ce qu'une base de données orientée graphe ? [Internet]. [cited 2023 Aug 2]. Available from: <https://aws.amazon.com/fr/nosql/graph/>
58. Chrisment C, Luguët J, Pujolle G, Zurfluh G. Bases de données relationnelles. Techniques de l'ingénieur. 1997;
59. Gastmans JP, Furlan M, Lopes MN, Borges JHG, Emerenciano VP. A Inteligência Artificial aplicada a Química de Produtos Naturais. O Programa Sistemático. Quim Nova [Internet]. 1990 [cited 2023 Jul 25];13(1):10–6. Available from: https://quimicanova.s bq.org.br/detalhe_artigo.asp?id=2719
60. Afendi FM, Okada T, Yamazaki M, Hirai-Morita A, Nakamura Y, Nakamura K, et al. KNApSACK Family Databases: Integrated Metabolite-Plant Species Databases for Multifaceted Plant Research. [cited 2023 Jan 12]; Available from: <http://kanaya.naist.jp/KNApSACK/>.
61. Sorokina M, Merseburger P, Rajan K, Yirik MA, Steinbeck C. COCONUT online: Collection of Open Natural Products database. J Cheminform. 2021 Dec 1;13(1).



62. Wolf S, Schmidt S, Müller-Hannemann M, Neumann S. In silico fragmentation for computer assisted identification of metabolite mass spectra [Internet]. 2010. Available from: <http://www.biomedcentral.com/1471-2105/11/148>
63. Bharati D, Jagtap R, ... KKAJ of, 2009 undefined. Chemo informatics: Newer approach for drug development. indianjournals.comD Bharati, RS Jagtap, KG Kanase, SA Sonawame, VR Undale, AV BhosaleAsian Journal of Research in Chemistry, 2009•indianjournals.com [Internet]. [cited 2024 Oct 27];2(1). Available from: <https://www.indianjournals.com/ijor.aspx?target=ijor:ajrc&volume=2&issue=1&article=001>
64. Nguyen-Vo TH, Teesdale-Spittle P, Harvey JE, Nguyen BP. Molecular representations in biocheminformatics. Memet Comput. 2024 Sep 20;16(3):519–36.
65. Wigh DS, Goodman JM, Lapkin AA. A review of molecular representation in the age of machine learning. Vol. 12, Wiley Interdisciplinary Reviews: Computational Molecular Science. John Wiley and Sons Inc; 2022.
66. Weininger D. SMILES, a Chemical Language and Information System: 1: Introduction to Methodology and Encoding Rules. J Chem Inf Comput Sci. 1988 Feb 1;28(1):31–6.
67. 1.5.3.0.0.0.0.Daylight.Simplified_molecular-input_line-entry_system.
68. 4. SMARTS - A Language for Describing Molecular Patterns.
69. Stein SE, Scott DR. Optimization and testing of mass spectral library search algorithms for compound identification. J Am Soc Mass Spectrom. 1994;5(9):859–66.
70. Watrous J, Roach P, Alexandrov T, Heath BS, Yang JY, Kersten RD, et al. Mass spectral molecular networking of living microbial colonies. Proc Natl Acad Sci U S A. 2012;109(26).
71. Wang M, Carver JJ, Phelan V V., Sanchez LM, Garg N, Peng Y, et al. Sharing and community curation of mass spectrometry data with Global Natural Products Social Molecular Networking. Nat Biotechnol. 2016;34(8):828–37.
72. Wang M, Bandeira N. Spectral library generating function for assessing spectrum-spectrum match significance. J Proteome Res. 2013 Sep 6;12(9):3944–51.
73. Xue J, Guijas C, Benton HP, Warth B, Siuzdak G. METLIN MS2 molecular standards database: a broad chemical and biological resource. Vol. 17, Nature Methods. Nature Research; 2020. p. 953–4.
74. Aisporna A, Benton HP, Chen A, Derks RJE, Galano JM, Giera M, et al. Neutral Loss Mass Spectral Data Enhances Molecular Similarity Analysis in METLIN. J Am Soc Mass Spectrom. 2022 Mar 2;33(3):530–4.
75. Xing S, Hu Y, Yin Z, Liu M, Tang X, Fang M, et al. Retrieving and Utilizing Hypothetical Neutral Losses from Tandem Mass Spectra for Spectral Similarity Analysis and Unknown Metabolite Annotation. Anal Chem. 2020 Nov 3;92(21):14476–83.
76. Moorthy AS, Wallace WE, Kearsley AJ, Tchekhovskoi D V., Stein SE. Combining Fragment-Ion and Neutral-Loss Matching during Mass Spectral Library Searching: A New General Purpose Algorithm Applicable to Illicit Drug Identification. Anal Chem. 2017 Dec 19;89(24):13261–8.

77. Li Y, Kind T, Folz J, Vaniya A, Mehta SS, Fiehn O. Spectral entropy outperforms MS/MS dot product similarity for small-molecule compound identification. *Nat Methods*. 2021 Dec 1;18(12):1524–31.
78. Treen DGC, Wang M, Xing S, Louie KB, Huan T, Dorrestein PC, et al. SIMILE enables alignment of tandem mass spectra with statistical significance. *Nat Commun*. 2022 Dec 1;13(1).
79. Bakiri A, Hubert J, Reynaud R, Lanthony S, Harakat D, Renault JH, et al. Computer-Aided ¹³C NMR Chemical Profiling of Crude Natural Extracts without Fractionation. *J Nat Prod*. 2017 May 26;80(5):1387–96.
80. Leroy R. Analyse de mélanges complexes par RMN en conditions de diffusion de spin. [Reims]: URCA; 2023.
81. Bakiri A, Plainchont B, de Paulo Emerenciano V, Reynaud R, Hubert J, Renault JH, et al. Computer-aided Dereplication and Structure Elucidation of Natural Products at the University of Reims. *Mol Inform*. 2017;36(10).
82. Laude DA, Wilkins CL. Identification of Organic Mixture Components without Separation: Quantitative and Edited Carbon-13 Nuclear Magnetic Resonance Spectrometry Data for the Analysis of Petroleum Distillates [Internet]. Vol. 58. UTC; 1986. Available from: <https://pubs.acs.org/sharingguidelines>
83. Tomi ABF, Casanova J. COMPUTER-AIDED CARBON-13 NMR STUDY OF PHENOLS CONTAINED IN LIQUIDS PRODUCED BY PYROLYSIS OF BIOMASS. Vol. 6, Biomass omf Bioewrg.v. 1994.
84. Hubert J, Nuzillard JM, Purson S, Hamzaoui M, Borie N, Reynaud R, et al. Identification of natural metabolites in mixture: A pattern recognition strategy based on ¹³C NMR. *Anal Chem* [Internet]. 2014 Mar 18 [cited 2021 Jun 18];86(6):2955–62. Available from: <https://pubs.acs.org/sharingguidelines>
85. Kuhn S, Nuzillard J. Easy Structural Dereplication of Natural Products by Means of Predicted Carbon-13 Nuclear Magnetic Resonance Spectroscopy Data**. *Chemistry–Methods* [Internet]. 2023 Apr 7 [cited 2023 Feb 16];3(4). Available from: <https://chemistry-europe.onlinelibrary.wiley.com/doi/10.1002/cmt.202200054>
86. Cordonnier J, Remy S, Renault J, Nuzillard J. Versa DB: Assisting ¹³C NMR and MS/MS Joint Data Annotation Through On-Demand Databases. *Chemistry–Methods* [Internet]. 2024 Jan 17;4(1). Available from: <https://chemistry-europe.onlinelibrary.wiley.com/doi/10.1002/cmt.202300020>
87. Kalchauer H, Robien W. CSEARCH: A Computer Program for Identification of Organic Compounds and Fully Automated Assignment of Carbon-13 Nuclear Magnetic Resonance Spectra. *J Chem Inf Comput Sci* [Internet]. 1984;25(2):103–7. Available from: <https://pubs.acs.org/sharingguidelines>
88. Bremser W. HOSE-A NOVEL SUBSTRUCTURE CODE*. Vol. 103, *Analytica Chimica Acta*. 1978.
89. Steinbeck C, Kuhn S. NMRShiftDB - Compound identification and structure elucidation support through a free community-built web database. *Phytochemistry*. 2004 Oct;65(19):2711–7.
90. Penchev PN, Schulz KP, Munk ME. INFERCNMR: A ¹³C NMR interpretive library search system. *J Chem Inf Model*. 2012 Jun 25;52(6):1513–28.



91. Bruguière A, Derbré S, Dietsch J, Leguy J, Rahier V, Pottier Q, et al. MixONat, a software for the dereplication of mixtures based on ¹³C NMR spectroscopy. *Anal Chem*. 2020;
92. Bruguière A, Derbré S, Coste C, Le Bot M, Siegler B, Leong ST, et al. ¹³C-NMR dereplication of *Garcinia* extracts: Predicted chemical shifts as reliable databases. *Fitoterapia*. 2018 Nov;131:59–64.
93. Robinette SL, Zhang F, Brüscheiler-Li L, Brüscheiler R. Web server based complex mixture analysis by NMR. *Anal Chem*. 2008 May 15;80(10):3606–11.
94. Huber F, Ridder L, Verhoeven S, Spaaks JH, Diblen F, Rogers S, et al. Spec2Vec: Improved mass spectral similarity scoring through learning of structural relationships. *PLoS Comput Biol*. 2021 Feb 16;17(2).
95. Huber F, van der Burg S, van der Hooft JJJ, Ridder L. MS2DeepScore: a novel deep learning similarity measure to compare tandem mass spectra. *J Cheminform*. 2021 Dec 1;13(1).
96. Simon R. Diversité chimique et potentialités antivirales d'Euphorbiacées tropicales. 2019.
97. Ruttkies C, Schymanski EL, Wolf S, Hollender J, Neumann S. MetFrag relaunched: Incorporating strategies beyond in silico fragmentation. *J Cheminform*. 2016;8(1):1–16.
98. Zheng X, Smith RD, Baker ES. Recent advances in lipid separations and structural elucidation using mass spectrometry combined with ion mobility spectrometry, ion-molecule reactions and fragmentation approaches. *Curr Opin Chem Biol* [Internet]. 2018 Feb;42:111–8. Available from: <https://linkinghub.elsevier.com/retrieve/pii/S1367593117301540>
99. Wang F, Liigand J, Tian S, Arndt D, Greiner R, Wishart DS. CFM-ID 4.0: More Accurate ESI-MS/MS Spectral Prediction and Compound Identification. *Anal Chem*. 2021 Aug 31;93(34):11692–700.
100. Nuzillard JM. Taxonomy-Focused Natural Product Databases for Carbon-13 NMR-Based Dereplication. *Analytica*. 2021 Jun 28;2(3):50–6.
101. Allard PM, Péresse T, Bisson J, Gindro K, Marcourt L, Pham VC, et al. Integration of Molecular Networking and In-Silico MS/MS Fragmentation for Natural Products Dereplication. *Anal Chem*. 2016;88(6):3317–23.
102. ISDB: In Silico Spectral Databases of Natural Products | Zenodo [Internet]. [cited 2023 Sep 11]. Available from: <https://zenodo.org/record/8287341>
103. Sorokina M, Steinbeck C. Review on natural products databases: Where to find data in 2020. Vol. 12, *Journal of Cheminformatics*. BioMed Central Ltd.; 2020.
104. Heuer A. Research Data Management. Vol. 62, IT - Information Technology. De Gruyter Oldenbourg; 2020.
105. Thawadee N, Mekruksavanich S. A Web-based Information Management System for Scientific Research. In: 2021 Joint 6th International Conference on Digital Arts, Media and Technology with 4th ECTI Northern Section Conference on Electrical, Electronics, Computer and Telecommunication Engineering, ECTI DAMT and NCON 2021. Institute of Electrical and Electronics Engineers Inc.; 2021. p. 293–6.
106. Conrad LY, Delahunty R, Ding W. The promise and the future of research data sharing. *Learned Publishing* [Internet]. 2022 Jan 1 [cited 2024 Oct 27];35(1):4–6. Available from: <https://onlinelibrary.wiley.com/doi/full/10.1002/leap.1432>

107. Mittal D, Mease R, Kuner T, Flor H, Kuner R, Andoh J. Data management strategy for a collaborative research center. *Gigascience* [Internet]. 2023 [cited 2024 Oct 27];12:giad049. Available from: <https://pmc.ncbi.nlm.nih.gov/articles/PMC10318494/>
108. Anderson B, Braxton S, Imker H, Popp T. 310.4 The Art of Preserving Scientific Data: Building collaboration into the preservation of a legacy database. 2018 [cited 2024 Oct 27]; Available from: <https://osf.io/rh9su/>
109. Gaudry A, Pagni M, Mehl F, Moretti S, Quiros-Guerrero LM, Cappelletti L, et al. A Sample-Centric and Knowledge-Driven Computational Framework for Natural Products Drug Discovery. *ACS Cent Sci*. 2024 Mar 27;10(3):494–510.
110. Magnus P, Cairns PM. Methods for Indole Alkaloid Synthesis. Enantiospecific Synthesis of Pentacyclic Desethylaspidosperma-Type Alkaloids Using an Exceptionally Mild Retro-Diels-Alder Reaction. *J Am Chem Soc* [Internet]. 1986 [cited 2024 Oct 31];108(2):217–21. Available from: <https://pubs.acs.org/doi/abs/10.1021/ja00262a006>
111. Evans JM, Gower JL. Hydrogen transfer in the retro-Diels–Alder fragmentation of trans-4-amino-, isopropyl-amino-, piperidino- and pyrrolidino-2H-1-benzopyran-3-ols. *Organic Mass Spectrometry* [Internet]. 1985 Jul 1 [cited 2024 Oct 31];20(7):440–4. Available from: <https://onlinelibrary.wiley.com/doi/full/10.1002/oms.1210200703>



Université de Lille
UFR3S-Pharmacie
DIPLOME D'ETAT DE DOCTEUR EN PHARMACIE
Année Universitaire 2024 / 2025

Nom : CORDONNIER
Prénom : Julien

Titre de la thèse :

CREATION, GESTION, ET INTERROGATION DE BASES DE DONNEES EN
RESONANCE MAGNETIQUE NUCLEAIRE ET EN SPECTROMETRIE DE MASSE
POUR LA DEREPLICATION DE SUBSTANCES NATURELLES EN MELANGE

Mots-clés : Alcaloïdes, Base de données spectrales, profilage chimique,
déréplication, spectrométrie de masse, résonance magnétique nucléaire

Résumé : Les substances naturelles (SN) sont des molécules issues du métabolisme d'organismes vivants, et jouent un rôle crucial en médecine moderne. Pourtant, moins de 40 000 des quelques 350 000 espèces de plantes à fleurs aujourd'hui identifiées ont été profilées chimiquement. La métabolomique, et en particulier le profilage chimique, permet d'identifier les métabolites spécialisés à partir d'un échantillon, comme les alcaloïdes indolomonoterpéniques de la pervenche de Madagascar (*Catharanthus roseus*), à travers des analyses spectrales (i.e. RMN et LC-HRMS²). Les données spectrales représentent la carte d'identité d'une SN. Leur conservation est essentielle, surtout face aux défis liés à la conservation physique. Conserver ces données est primordial pour la déréplication, un processus clé permettant d'identifier des composés connus *via* la comparaison de leurs spectres expérimentaux avec des spectres de référence. La création de bases de données solides pour stocker ces informations s'avère donc crucial pour accélérer la recherche et améliorer la fiabilité des résultats. De plus, l'application des principes FAIR (Findable, Accessible, Interoperable, Reusable) vise à garantir l'accessibilité et l'exploitation de ces données par la communauté scientifique, favorisant ainsi l'intégration de nouvelles approches, comme l'élucidation structurale assistée par ordinateur ou l'évaluation *in silico* de l'activité biologique. Cette thèse amorce la digitalisation de l'alcaloïdothèque de l'UMR CNRS 7312 – Institut de Chimie Moléculaire de Reims (URCA) en créant la Digital Alkaloid Collection of Reims (DACR), une première base de données pour les spectres MS/MS et RMN 1D ¹³C. L'objectif est également de servir de modèle à de prochaines études afin de faciliter l'interopérabilité des données.

Membres du jury :

Président : PU. RIVIERE Céline | UFR3S, Pharmacie

Directeur, conseiller de thèse : MCU. REMY Simon | URCA, Pharmacie

Assesseur(s) : PU. SAHPAZ, Sevser | UFR3S, Pharmacie