



Mise au point d'une méthode d'analyse dérépliative par RMN du carbone 13

Antoine Bruguière

► To cite this version:

Antoine Bruguière. Mise au point d'une méthode d'analyse dérépliative par RMN du carbone 13. Médecine humaine et pathologie. Université d'Angers, 2019. Français. NNT : 2019ANGE0085 . tel-03783739

HAL Id: tel-03783739

<https://theses.hal.science/tel-03783739>

Submitted on 22 Sep 2022

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

THÈSE DE DOCTORAT DE

L'UNIVERSITÉ D'ANGERS
COMUE UNIVERSITÉ BRETAGNE LOIRE

ÉCOLE DOCTORALE N° 600
École doctorale Écologie, Géosciences, Agronomie et Alimentation
Spécialité : Pharmacologie, phytochimie et toxicologie

Par

Antoine BRUGUIÈRE

Mise au point d'une méthode d'analyse déréplicative par RMN du carbone 13

Thèse présentée et soutenue à Angers, le 16/12/2019

Unité de recherche : Substance d'Origine Naturelle et Analogues Structuraux (SONAS), EA 921

Thèse N° : 181324

Rapporteurs avant soutenance :

Jean-Marc NUZILLARD | Directeur de recherches au CNRS,
Université de Reims Champagne-Ardenne
Céline RIVIÈRE | Maître de conférences en pharmacognosie,
Faculté de Pharmacie, Université de Lille

Composition du Jury :

Jean-Marc NUZILLARD | Directeur de recherches au CNRS,
Université de Reims Champagne-Ardenne
Céline RIVIÈRE | Maître de conférences en pharmacognosie,
Faculté de Pharmacie, Université de Lille
Anne-Claire MITAINE-OFFER | Professeur des universités en
pharmacognosie, Faculté des sciences de Santé, Université
de Bourgogne Franche-Comté
Pierre CHAMPY | Professeur des universités en
pharmacognosie, Faculté de Pharmacie, Université Paris-
Sud / Paris Saclay
Pascal RICHOMME | Professeur des universités en
pharmacognosie, Faculté de Pharmacie, Université d'Angers
Directeur de thèse
Séverine DERBRÉ | Maître de conférences en
pharmacognosie, Faculté de Pharmacie, Université d'Angers
Encadrante

L'auteur du présent document vous autorise à le partager, reproduire, distribuer et communiquer selon les conditions suivantes :



- Vous devez le citer en l'attribuant de la manière indiquée par l'auteur (mais pas d'une manière qui suggérerait qu'il approuve votre utilisation de l'œuvre).
- Vous n'avez pas le droit d'utiliser ce document à des fins commerciales.
- Vous n'avez pas le droit de le modifier, de le transformer ou de l'adapter.

Consulter la licence creative commons complète en français :
<http://creativecommons.org/licences/by-nc-nd/2.0/fr/>

Ces conditions d'utilisation (attribution, pas d'utilisation commerciale, pas de modification) sont symbolisées par les icônes positionnées en pied de page.



Ce travail de doctorat implique évidemment la contribution de nombreuses personnes que je me dois de remercier chaleureusement. Même si les remerciements peuvent être succincts, ils ne sont pas pour autant moins sincères.

D'abord un immense merci à Pascal Richomme et Séverine Derbré pour avoir dirigé et encadré ce travail. Bien que leur approche scientifique soit parfois différente, ils m'ont permis d'avancer vers les objectifs fixés et de produire des travaux, je l'espère, de qualité. Ils ont toujours été très disponibles et prêts à me conseiller lorsque j'en avais besoin, et pour cela je les remercie encore. Enfin, ce fût un réel plaisir de partager du temps avec eux, à la fois au niveau professionnel, mais également au niveau personnel. Tout ce que je peux espérer est que j'aurais la chance de pouvoir renouveler cette expérience au travers de futures collaborations. Merci mille fois encore !

Je tiens à remercier Jean-Marc Nuzillard et Céline Rivière d'avoir accepté d'être rapporteurs de ce manuscrit et de faire partie du jury aux côtés d'Anne-Claire Mitaine-Offer et de Pierre Champy que je remercie également.

Merci une seconde fois à Pierre Champy pour avoir, avec Luis Manuel Peña Rodriguez, fait partie de mon comité de suivi de thèse. Leurs remarques et suggestions ont maintenu le projet sur la bonne voie.

Un autre remerciement à Luis Manuel Peña Rodriguez pour m'avoir accueilli deux fois au sein de son laboratoire CICY à Mérida. En plus de faire fructifier la collaboration scientifique entre nos deux laboratoires, cela m'a permis de découvrir le Mexique et d'y faire de super rencontres. Je dois donc remercier encore une fois Pascal Richomme, porteur du projet auprès du programme d'échange ECOS Nord, grâce à qui j'ai pu profiter de cette opportunité.

Je remercie Frédéric Saubion du laboratoire LERIA et les étudiants en master 2 informatique Jules Leguy et Valentine Rahier pour leur participation à la création du programme.

Merci à Béatrice Charreau de m'avoir accueilli au sein du laboratoire INSERM UMR 1064 afin que je puisse effectuer les tests biologiques sur cellules humaines.

Parmi les membres de la SFR MATRIX, merci à Ingrid Freuze pour son investissement dans les analyses LC/MS de nos échantillons (préparation des séries d'analyses, optimisation des conditions MS...) et son expertise sur les résultats. Merci également à Benjamin Siegler pour l'analyse en RMN de nos différents échantillons.

J'ai également eu la chance de pouvoir être moniteur pendant 3 ans à la faculté de Pharmacie. Je tiens donc à remercier le personnel du département Pharmacie de la Faculté de Santé d'Angers. Merci aux enseignants de chimie Séverine Boisard, David Guilet et Jean-Jacques Helesbeux, ainsi qu'aux enseignants de pharmacognosie Pascal Richomme, Anne-Marie Le Ray, Andreas Schinkovitz et Séverine Debré. Merci également à Yannick Abatuci, David Dallerac et Aurore Michaud pour les travaux pratiques en chimie et Patricia Blanchard pour les travaux pratiques en pharmacognosie. Enfin, merci à l'équipe de galénique : Thomas Briot, Florian Fouchet, Vincent Lebreton et Brigitte Pech. Merci à tous pour votre bonne humeur, votre aide précieuse et tout ce que vous m'avez appris pendant les TP.

Last but not least (comme diraient nos collègues anglophones), je tiens à remercier l'ensemble des personnes que j'ai eu la chance de côtoyer au laboratoire SONAS. D'abord mes collègues de bureau les plus récents, Míša Škopíková et Sekhou Cisse, avec qui j'ai passé d'excellents moments (tout en restant très professionnel, évidemment). La chaleur du bureau, celle due aux radiateurs que l'on pousse au max, mais surtout celle due à la bonne ambiance, m'a permis de m'y sentir chez moi quand j'y étais. Merci aussi aux « anciens » du bureau, Paul Engler et Stéphane Dejoie que j'ai un peu moins croisés, mais qui faisaient également des voisins sympathiques. Un grand merci à Anne-Marie Le Ray avec qui j'ai eu le bonheur de travailler en master 2 et avec laquelle il m'est toujours très agréable et intéressant d'avoir des conversations scientifiques... mais pas que ! Merci à Dimitri Bréard de m'avoir laissé l'embêter pendant toutes ces années, à chaque fois que j'en avais besoin. Merci aux différents stagiaires et étudiants avec lesquels j'ai travaillé, qui ont toujours été efficaces et sympathiques. Merci donc à Quentin Pottier, Stanislas Maisonneuve, Léa Gicquel et Yasmine Sakr d'avoir contribué à ce travail. Un merci plus particulier à Chloé Coste et Joel Dietsch, avec lesquels j'ai passé le plus de temps, ce qui était vraiment une joie pour moi. Merci encore pour votre investissement dans ce travail. Merci également à Luis Herbert, doctorant que j'ai eu le plaisir de rencontrer au Mexique, pour sa gentillesse. Enfin, je tiens à remercier tous les autres membres de la team SONAS, pour chacun desquels je pourrais également détailler toutes les fois où ils m'ont aidé. Les portes de leurs bureaux ont toujours été ouvertes (du moins, métaphoriquement parlant), même pour répondre aux plus naïves de mes questions. Merci à Guillaume Viault, Khaled Al Sabil, Maxime Le Bot, Denis Séraphin, Chau-Phi Dinh Alexia Ville, Soprane Suor Cherer, Marie-Christine Aumond et Nadège Blon.

J'espère n'avoir oublié personne, et si c'est le cas, c'est sûrement dû à la fatigue ; le travail de rédaction étant quelque peu énergivore !

Sommaire

PRODUCTIONS SCIENTIFIQUES

1. Publications
2. Communications orales
3. Posters

INTRODUCTION 1

I. ÉTAT DE L'ART 2

1. Article 1 : A highlight on ^{13}C -NMR based dereplication methods 2

II. TRAVAUX PERSONNELS 44

1. Construction des bases de données 44

1.1. L'étape clé des références 44

1.2. Article 2: ^{13}C -NMR dereplication of *Garcinia* extracts: Predicted chemical shifts as reliable databases 47

2. Algorithme de déréplication 60

2.1. Éléments de fonctionnement des algorithmes de déréplication par RMN- ^{13}C existants 60

2.2. Essais de déréplication 63

2.3. Fonctionnement de l'algorithme DerepCrude 69

2.4. Recherche d'améliorations algorithmiques 73

2.5. Description de l'algorithme MixONat 77

2.6. Fonctionnement du *matching* de l'algorithme MixONat 86

2.7. Présentation des résultats de l'algorithme MixONat 88

3. Validation de la méthode 93

3.1. Huile essentielle de menthe poivrée 93

3.2. Article 3 : MixONat, a software for mixtures dereplication based on ^{13}C -NMR experiments. 98

4. Application de la méthode 166

4.1. Article 4 : Polyphenylated polycyclic acylphloroglucinols identification from *Garcinia bancana* bark using ^{13}C -NMR dereplication program MixONat 166

CONCLUSION GÉNÉRALE ET PERSPECTIVES 208

ANNEXES 210

1. Filtre d'intensité des algorithmes de recherche 210

1.1. Le filtre d'intensité de DerepCrude 210

1.2. Méthodes de clustering 214

1.3. Fonctionnement de l'algorithme DBSCAN 216

1.4. Essais de clustering avec DBSCAN 218

2. Valorisation biologique 223

2.1. Type cellulaire 223

2.2. Expression des protéines de surface 224

2.3. Transcrits ARNm 226

2.4. Voie de l'IFN γ 228

2.5. Expression intracellulaire des protéines 228

2.6. Conclusion et perspectives 230

BIBLIOGRAPHIE 231

TABLE DES FIGURES 235

TABLE DES TABLEAUX 237

Productions scientifiques

1. Publications

- **Bruguère, A.**; Derbré, S.; Coste, C.; Le Bot, M.; Siegler, B.; Leong, S. T.; Sulaiman, S. N.; Awang, K.; Richomme, P. ¹³C-NMR Dereplication of *Garcinia* extracts: Predicted chemical shifts as reliable databases. *Fitoterapia* 131, 59-64 (2018)
- Bréard, D.; Viault, G.; Mezier, M.-C.; Pagie, S.; **Bruguère, A.**; Richomme, P.; Charreau, B.; Derbré, S. Additional Insights into *Hypericum perforatum* Content: Isolation, Total Synthesis, and Absolute Configuration of Hyperbiphenyls A and B from Immunomodulatory Root Extracts. *Journal of Natural Products* 81(8), 1850-1859 (2018)

2. Communications orales

- **Bruguère, A. et al.** ¹³C-NMR dereplication of complex mixtures: building a custom search algorithm. 6th AFERP International Conference, Rennes (2018)
- **Bruguère, A. et al.** ¹³C-NMR dereplication analysis of complex mixtures. SFR QUASAV Ph. D. Day, Angers (2018) – *Prix de la meilleure communication orale*

3. Posters

- **Bruguère, A.**; Derbré, S.; Dietsch, J.; Leguy, J.; Rahier, V.; Pottier, Q.; Saubion, F.; Richomme, P. ¹³C-NMR dereplication of medicinal plant extracts using a home-made software. 67th International Congress and Annual Meeting of the Society for Medicinal Plant and Natural Product Research, Innsbruck (2019) – *Prix AFERP du meilleur poster*
- Viault, G.; Bréard, D.; Dinh, C. P.; Blon, N.; **Bruguère, A.**; Le Ray, A. M.; Bataillé-Simoneau, N.; Guillemette, T.; Simoneau, P.; Richomme, P. Xanthone derivatives from NPs library as potential UPR inhibitors for alternative crop protection: molecular modelling and biological activity. 30th International Symposium on the Chemistry of Natural Products, Athens (2018)
- **Bruguère, A.**; Dietsch, J.; Derbré, S.; Bréard, D.; Suor Cherer, S.; Richomme, P. Custom-made algorithm: a powerful tool for ¹³C-NMR dereplication of complex mixtures. Proof of concept on a *G. mangostana* fruit peel extract. 30th International Symposium on the Chemistry of Natural Products, Athens (2018)
- **Bruguère, A.**; Derbré, S.; Dietsch, J.; Le Bot, M.; Siegler, B.; Leong, S. T.; Sulaiman, S. N.; Awang, K.; Richomme, P. ¹³C-NMR dereplications of complex mixtures: predicted vs experimental chemical shifts databases. 5th AFERP International Conference, Angers (2017) – *Prix AFERP du meilleur poster*
- Coste, C.; **Bruguère, A.**; Gérard, N.; Awang, K.; Richomme, P.; Derbré, S.; Charreau, B. Modulation of the expression of MHC molecules by PPAPs from *Garcinia bancana*. 5th AFERP International Conference, Angers (2017)
- Benbelkacen Z.; Le Ray A. M.; **Bruguère, A.**; Suor Cherer, S.; Bréard, D.; Blon, N.; Marchi, M.; Rolland, A.; Simoneau, P.; Bataillé-Simoneau, N.; Guillemette, T.; Richomme, P. Identifying natural products (NPs) as potential UPR inhibitors for crop protection. 5th AFERP International Conference, Angers (2017)

Introduction

Notre laboratoire **S**ubstances d'**O**rigine **N**aturelle et **A**nalogues **S**tructuraux (**SONAS**) travaille depuis plusieurs années sur les métabolites secondaires présentant des propriétés anti-inflammatoires et immunomodulatrices. Les polyphénols, et en particuliers ceux étant prénylés, font partis des composés qui présentent les effets biologiques recherchés [1]. Les Clusiaceae et Calophyllaceae sont deux familles de plantes connues pour contenir une grande variété de polyphénols prénylés [2]. Afin d'étudier les effets de ces molécules sur la réponse inflammatoire et/ou immune, des essais biologiques ont été menés sur différents types de polyphénols prénylés [3]. Ces travaux de doctorat de Caroline Rouger ont démontré que la guttiferone J, un acylphloroglucinol polycyclique polyprénylé (**PPAP** : **P**olycyclic **P**olyprenyated **A**cy**P**hloroglucinol), pouvait diminuer l'expression de plusieurs molécules du **C**omplexe **M**ajeur d'**H**istocompatibilité (**CMH**), notamment les CMH de classe II ainsi que le **H**uman **L**eucocyte **A**ntigen **E** (**HLA-E**) [3]. Ces molécules du CMH contrôlent à la fois l'immunité innée et la réponse immunitaire adaptative de l'organisme, que ce soit en conditions physiologiques ou pathologiques. Moduler l'expression des molécules du CMH pourrait donc présenter des effets thérapeutiques intéressants. En effet, HLA-E permet à certaines cellules cancéreuses d'échapper aux leucocytes; son inhibition constituerait ainsi une cible potentielle pour un agent anti-cancéreux. Les molécules du CMH de classe II, quant à elles, sont impliquées dans plusieurs pathologies telles que les maladies auto-immunes et les phénomènes de rejet de greffe. Ainsi, des drogues capables de moduler la réponse immunitaire présentent des outils utiles à l'immunothérapie et à la compréhension générale des mécanismes cellulaires impliqués dans cette réponse. Notre ambition initiale était donc d'identifier de nouvelles PPAPs actives vis à vis du CMH et/ou du HLA-E, et de tester leurs effets sur le système immunitaire en les comparant à la guttiferone J.

Le fractionnement et la purification de molécules à partir d'extraits végétaux complexes n'est pas tâche aisée car ils nécessitent souvent de nombreuses optimisations des étapes chromatographiques et analytiques avant d'arriver à un composé pur. Ces composés peuvent de plus se révéler d'intérêt limité ou inexistant pour l'objectif que l'on cherche à atteindre, résultant en une perte de temps et de moyens considérables. Des techniques appelées « déréplication » ont fait leur apparition au début des années 90, et continuent de se développer, afin de pallier ces inconvénients. L'objectif de la déréplication est d'être capable, grâce à une comparaison avec des bases de données, de directement identifier les molécules connues à partir d'une analyse de mélange complexe. Une analyse déréplicative permet donc de se focaliser sur des fractions qui contiennent les molécules d'intérêt; dans notre cas, les PPAPs.

Ce travail de doctorat est donc parti de l'objectif de pouvoir utiliser une méthode de déréplication nous permettant de pouvoir rapidement identifier les mélanges riches en PPAPs, afin de pouvoir les purifier pour évaluation biologique. Après avoir fait un état de l'art concernant les méthodes de déréplication reportées dans la littérature (cf. **I. État de l'art**), les travaux personnels seront présentés en seconde partie (cf. **II. Travaux personnels**). Ceux-ci ont d'abord consistés en une réflexion sur la façon la plus adéquate de construire les bases de données utilisées en déréplication (cf. **1. Construction des bases de données**). Ensuite, après avoir testé les différents algorithmes de déréplication disponibles, nous avons développé notre propre programme (cf. **2. Algorithme de déréplication**). Le développement de ce programme a nécessité de travailler sur plusieurs exemples de

déréplication afin de valider la méthode (cf. **3. Validation de la méthode**) avant de pouvoir enfin l'appliquer à la déréplication des PPAPs (cf. **4. Application de la méthode**).

Ce manuscrit de thèse est construit autour de propositions de publications qui seront soumises, ou sont en cours de soumission auprès de journaux scientifiques. Cela permettra de valoriser les différents travaux réalisés au cours du doctorat, mais également la consultation du manuscrit par des lecteurs non-francophones.

I. État de l'art

Avant de choisir une méthode de déréplication adaptée à notre problème, il est important de savoir quels sont les outils nécessaires à la mise en place d'un tel processus ainsi que les divers protocoles existants. Une revue de la littérature a d'abord permis de dégager quelles étaient les manières les plus répandues de conduire une étude dérépliquative et, à partir de ces exemples, le protocole de déréplication a pu être découpé en plusieurs étapes clé. Les différentes façons de procéder pour chacune de ces étapes sont ensuite été analysées et discutées, afin d'en dégager les avantages et inconvénients. L'ensemble de ces éléments est présenté dans la proposition de revue suivante.

1. Article 1 : A highlight on ^{13}C -NMR based dereplication methods

L'analyse des documents de la littérature révèle que la Spectrométrie de Masse (SM) et la spectroscopie par Résonance Magnétique Nucléaire (RMN) sont les méthodes analytiques les plus fréquemment associés aux analyses de déréplication. Bien que l'usage de la SM soit plus répandu, la RMN y trouve quand même sa place. En se basant sur des exemples de déréplication, utilisant notamment les réseaux moléculaires [4], le processus a pu être découpé en plusieurs étapes clés : l'échantillonnage, le choix d'outils analytiques pertinents, la collecte ou l'établissement de références/bases de données, la comparaison entre données à dérépliquer et références et enfin, la confirmation des hypothèses faites par la déréplication. Ces étapes peuvent être abordées de manières très différentes au sein des équipes de recherche, en résultent ainsi des méthodes de déréplication très variées.

Les méthodes utilisant la RMN du ^{13}C ont fait l'objet d'une attention particulière car c'est celles qui semblaient répondre à notre problème de déréplication des acylphloroglucinols polycycliques polyprénylés (PPAPs). En effet, les PPAPs sont une classe de molécules comprenant de nombreux stéréoisomères [5], ce qui les rend difficile à différencier par SM, présentant donc la RMN comme le choix le plus adapté.

A highlight on ^{13}C -NMR based dereplication methods

Antoine Bruguère, Séverine Derbré, Pascal Richomme

SONAS, EA921, UNIV Angers, SFR QUASAV, Faculty of Health Sciences, Dpt Pharmacy,
16 Bd Daviers, 49045 Angers CEDEX 01, France

Abstract: The process allowing to quickly identify known molecules in a mixture, called “dereplication”, is now widely used in the metabolomics and natural products fields. The main analytical methods used to fulfill this purpose are LC-MSⁿ and NMR. Nowadays, where the ways to proceed are very diverse, we took a special look at dereplication methods using ^{13}C -NMR. First, we will give an overview of a few dereplication methods, and then detail each step of a dereplication workflow, before analyzing ^{13}C -NMR based dereplication methods. This will hopefully provide the reader with insights on advantages and drawbacks of the available methods.

INTRODUCTION

The word “dereplication” can refer to a lot of different processes and objectives [1]. In 1990, this term was initially used by Beutler *et al.* in a context of active compounds discovery from natural products (NPs). Whatever the definition, the core of the concept remains a general time-saving method, allowing to make assumptions about compounds present in a complex mixture, based on the comparison between experimental and reference data.

A highlight on ^{13}C -NMR based dereplication methods

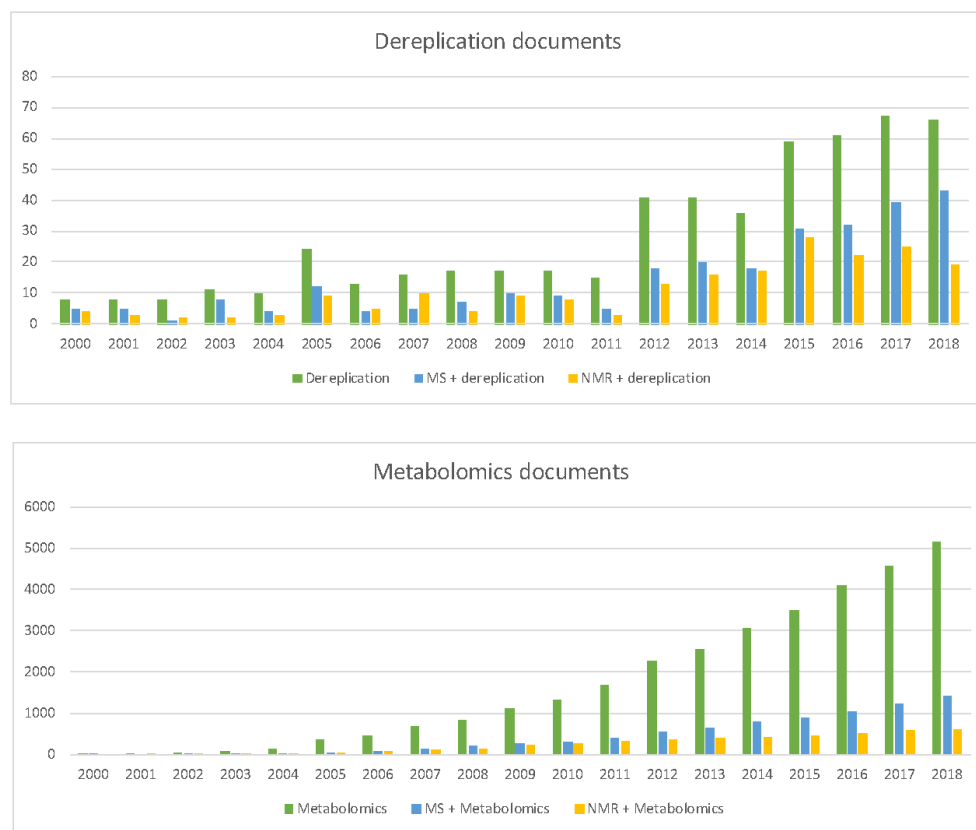
25 The two main domains in which dereplication thrives are NPs valorization (e.g. drugs,[2]
26 dietary supplements, cosmetics,[3, 4] food additives [5] or pesticides [6]) and metabolomics
27 where, integrated within the workflow, it takes part in the objective of metabolite profiling [7-
28 19]. Even if objectives are different, they often both have to deal with metabolites identification,
29 usually present in complex matrices and thus find a common interest in the different
30 dereplication techniques that can save tedious purifications as well as structure elucidation
31 work.

32

33 Mass spectrometry (MS) is one of the most commonly used method, as it represents around
34 25% of documents “metabolomic*” as a keyword and over 50% of documents containing
35 “dereplication” based on Scopus survey [20] (**Figure 1**). In both cases, it is the first associated
36 keyword on an analytical standpoint. In this context, in which MS is the preferred analytical
37 method, but where the improvement of NMR technologies allows a better sensibility, we
38 wonder if ^{13}C -NMR is now able to stand as a method as interesting as MS for dereplication.

39

A highlight on ^{13}C -NMR based dereplication methods



40

41 **Figure 1:** Research results on Scopus for MS and NMR in dereplication and metabolomics
42 related documents.

43

44 In order to do so, documents containing the word “dereplication” were searched on the Scopus
45 and SciFinder [21] databases. This led to around 600 and 1,000 results respectively, starting
46 with a few articles from the 90s and exponentially increasing until now, nowadays reaching
47 respectively 60 and 100 documents published yearly and showing the growing interest for this
48 concept. Searching for “dereplication” AND “NMR” cuts down the number of results to about
49 a third of the initial number but shares the same time period and exponential trend over the
50 years. However, some publications having the same dereplication objectives do not necessarily
51 use the word “dereplication, but “rapid identification”, “fast identification” or even “computer-

A highlight on ^{13}C -NMR based dereplication methods

aided identification”. Those different keywords were also used in combination with “NMR” in order to gather a maximum of information on the subject (**Figure 2**). Out of the 272 collected papers, 71 publications reporting 1D ^{13}C -NMR dereplication works were identified.

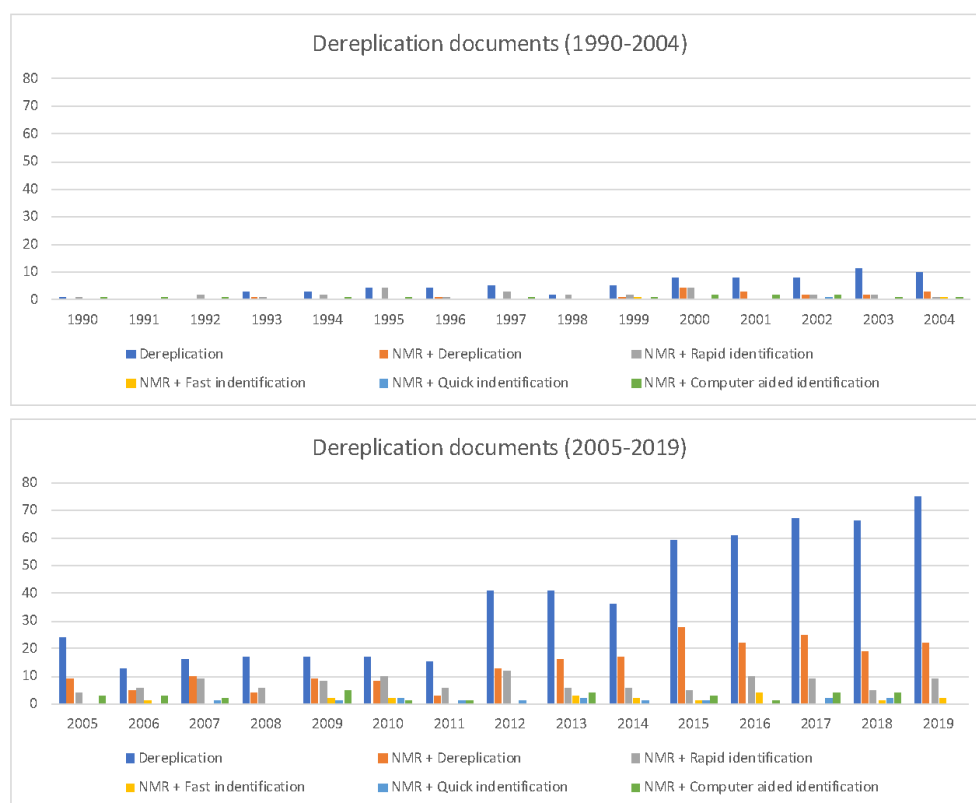


Figure 2: Number of dereplication related documents over the years on Scopus

This publication will first quickly review some of the most used dereplication method as it seems important to discuss benefits and drawbacks of these different strategies. This will also help to split the whole dereplication process in different steps that will be detailed in a second part. Then, this work will focus on the different methods using mainly ^{13}C -NMR as their dereplication tool, describing their process and commenting along the way on the strength and weaknesses of each one.

65 FOCUS ON A FEW DEREPLICATION METHODS

66

67 The following paragraphs regroup a few examples of dereplication methods, which are
68 frequently reported in the literature. It seemed important to point out their benefits and
69 drawbacks before focusing on actual and emerging ¹³C-NMR dereplication methods.

70

71 LC-MSⁿ and the molecular networking phenomenon

72

73 As mentioned before, LC-MS is one of the most used analytical method in metabolomics and
74 dereplication workflows. A majority of published papers rely, at least partially, on LC-MSⁿ to
75 carry out a dereplication analysis [22-45]. However, it can be hard and time-consuming to
76 handle the whole datasets when working with a large number of samples, making it difficult to
77 quickly draw conclusions from the experiments. This led to the development of molecular
78 networking introduced in 2013 by Dorrestein *et al.* [46], which is now widely used as a
79 dereplication method.[42, 44, 47-53] It is based on the LC-MS² analysis of different samples
80 considering the fact that similar molecules analyzed in the same conditions will fragment in a
81 comparable way. Mass spectra are aligned two by two to establish a similarity score (*i.e.* cosine
82 score) that relies on the observed differences between the different peaks (*m/z* values and
83 relative intensities). Those similarities and differences are then visualized as a web of nodes,
84 representing specific (de)-protonated molecules, and strings, which thickness depends on the
85 cosine score between the nodes. Potentially similar structures will form clusters of nodes,
86 strongly bounded together. Comparison between the data from the nodes of interest and
87 required references will allow to make assumptions on the NPs and their analogs.

88 The molecular networking process can be done online on the GNPS website [54, 55] and then
89 exported to Cytoscape [56], a software used for visualization and further data processing.

A highlight on ¹³C-NMR based dereplication methods

90 Another goal of the GNPS website is the sharing of MS spectra, thus offering users access to a
91 collective database. It is a powerful tool that can handle a lot of data and allows additional
92 information to be displayed on the network, such as biological activity [53] or type of fraction
93 [51], making a direct link between a specific cluster of molecules and the supplementary data.
94 However, some problems emerge from the MS technique itself like ionization problems for
95 some compounds, reported ions in databases being ionized in a different way than the one
96 desired, or the difficult reproducibility of experimental conditions, pushing researchers to turn
97 to *in-silico* predictions [57] or the acquisition of standards in order to build in-house libraries
98 [50]. The molecular networking itself can result in data “alteration” by merging initially
99 different spectra as a single node (isomers, isotopes and isobars) [58]. The parameters used for
100 the networking, such as the cosine score, are also very important because they will structure the
101 entire web of nodes but can be tricky to set up. It usually requires several tries or some
102 experience with a particular type of sample before creating a molecular network with the proper
103 restriction settings. All of this is possible if the LC conditions were optimized beforehand in
104 order to properly separate the different compounds, adding another step of complexity in the
105 method.

106

107 In NPs valorization, this useful method is more and more used for prioritization of fractions and
108 extracts of interest [59]. But the researchers will often have to turn to NMR analysis for their
109 final identification [24, 31, 33, 34, 36, 37, 41, 43, 45, 49, 60-63], especially when dealing with
110 stereoisomers.

111 **LC-NMR**

112

113 The hyphenation of LC with online NMR has been used in different works aiming at the
114 dereplication of complex mixtures [25-27, 35, 64-78]. Over the years, the addition of a solid
115 phase extraction (SPE) step before NMR analysis, optimized the results given by this method
116 [28, 40, 74, 75, 79-93]. The LC separates the constituents of the mixtures that will be analyzed
117 using ^1H -NMR after SPE treatment to concentrate the samples and get rid of the mobile phase,
118 replacing it with deuterated solvent. Even it is already quite fast to record a proton spectrum
119 (when compared to the current ^{13}C -NMR or 2D NMR acquisition time), a capillary cryoprobe
120 is required to obtain quality data. Indeed, eluates from the liquid chromatography system will
121 be directly analyzed in NMR; but the amount of sample that can be loaded on the column being
122 quite limited, it will influence the strength of the resulting NMR signal. The capillary cryoprobe
123 will allow to increase NMR sensitivity, thus obtaining interpretable data.

124

125 This method is interesting because results can be obtained very quickly, but important and very
126 expensive hardware is required to carry it out. Furthermore, the sample concentration can make
127 it difficult to obtain spectra with a satisfying quality. It also faces the limits of proton NMR,
128 meaning overlapping signals [77] and solvent sensitive chemical shifts.

129

130 The obtained NMR spectra will be manually compared to the references of reported compounds
131 for structure hypothesis, even if additional ^{13}C -NMR will be required for unambiguous structure
132 identification. However, it is still possible to “reconstruct” a ^{13}C spectrum from 2D experiments,
133 under certain conditions. With powerful NMR hardware, a 600 MHz spectrometer equipped
134 with a micro cryoprobe, Williams *et al.* [94] managed to deduce the ^{13}C chemical shifts, from
135 HMBC and HMQC experiments, with only 5 μg of a C_{21} compound.

Search by features

The “search by features” method is not a frequently used for NPs dereplication. Its originality makes it an interesting method to point out. It requires MS and NMR experiments and is based on a search by structural features. A low-resolution mass spectrum allows the obtention of the molecular weight of the compound, and both ^1H -NMR and edited Heteronuclear Single Quantum Correlation (HSQC) spectra help deducing the number of characteristic groups in the NP, such as CH_3 , CH_2 , CH , and additionally quaternary carbons using Heteronuclear Multiple Bond Correlation (HMBC) experiment. After a manual interpretation of the spectra to gather all this information, the users query this information in a home-made algorithm, and a database (DB) built from structures collected from online DBs. One of the databases called DEREPI- NP [95], and made freely available by the authors, can show how many times one of 65 characteristic groups is found in the NPs of the Universal Natural Products Database (UNPD), which gathers 229 358 NPs. In another case [96], the database used by the team contains the structural information of more than 200 000 NPs gathered from the Dictionary of Natural Products (DNP) [97], AntiBase [98] and their own pure compound collection [96].

Searching these structural features often leads to the correct identification of the compounds that were present in the selected databases. The strength of this method consists in the numerous and diverse parameters used for the search: molecular weight, carbon type and specific structural features. However, this method is limited to the analysis of a single compound or, when working with a mixture, needs focusing on the individual signals of one major compounds.

DEREPLICATION TOOLBOXES

From the previous example, we can divide the dereplication methods in several key features for the process (**Figure 3**).

Samples

Dereplication studies have been conducted on very complex mixtures (crude extracts), sometimes fractionated, and even on purified compounds. Needless to say, a dereplication process can perfectly work on a simple sample but tends to fail as the complexity increases. Mixtures can also be composed of very different kind of molecular scaffolds or, the other way around, a lot of NPs from the same structural class. This will obviously impact the dereplication results.

Analytical tools

One or several of them can be used to carry out the dereplication analysis. Each analytical system will provide different type of information and must be chosen accordingly to the sample and the aim of the study. Most of the used analytical tools are hyphenated with a chromatographic system [69, 80] such as high-performance liquid chromatography (HPLC), Ultra Performance Liquid Chromatography (UPLC) [32, 37], supercritical fluid chromatography (SFC) [99], centrifugal partition chromatography (CPC) [78, 100] or gas chromatography (GC) [101]. The resulting separation allows thus an “individual analysis” of the molecules inside the mixture, making it easier to identify them. The analysis itself can be carried out using an ultraviolet (UV) detector, that usually provides information limited to the

chemical class of NP but, as mentioned before, the ones more commonly used are mass spectrometry (MS) and nuclear magnetic resonance (NMR).

While MS is very much appreciated for its high sensibility, a single method cannot be universal and is not especially easy to reproduce from one laboratory to another. This can lead to compounds in the mixture being little to not ionized when analyzed using MS, thus depicting a distorted representation of the actual composition. NMR allows every organic compound to be detected, but it is less sensible than MS [102, 103], making it harder to properly identify the molecules present in a smaller amount. When working with ^1H -NMR, the experiments can be conducted very quickly, which is a huge advantage when working with many samples but considering its small scale of roughly 20 ppm at most, signal overlaps are quite frequent, hence resulting in information loss [52, 88]. One can turn to 2D-NMR (e.g. HSQC, J-RES, TOCSY) to bypass this problem : even if they require a longer acquisition time, proton signals are spread along the second dimension, thus allowing previously overlapping signals to be distinguishable, but losing resolution in the process [102]. ^{13}C -NMR experiment also leads to few signal overlaps, thanks to its larger ppm scale. But its sensibility is much lower (around 1/6400) than proton NMR [102], hence requiring long acquisition times. Further experiments like DEPT analysis can provide additional useful structural information on the molecules (*i.e.* number of CH_3 , CH_2 , CH , C) [104].

When working with a large amount of data, it is often necessary to use statistical methods or other type of data processing in order to “sort” metabolites of interest in the sample, bring out trends in the dataset and thus be able to interpret it. Among those methods, principal component analysis (PCA) [19], hierarchical clustering analysis (HCA) [100] and molecular networking [46] can be cited.

A highlight on ¹³C-NMR based dereplication methods

The analytical tools must be picked depending on the objective of the work: even a robust dereplication method will not work if the tools are not appropriate for the selected sample, or if the quality of the data collection and/or processing is not good enough.

References

In order to quickly identify the compounds, a comparison with published data is a mandatory step. The way the references are selected is also a major stage in the whole method. Once again, very diverse types of datasets are used as references by researchers conducting a dereplication study.

References can be experimental data, gathered from the literature or predicted data, calculated mainly using computational methods, but can also be determined using data trends [105-107]. Even if it seems like the optimal solution, concerning NMR, experimental data is much longer to collect, since a tedious manual work is required. It might sometimes be impossible to be able to collect “relevant” experimental data since it is impossible to find experiments being conducted in the required conditions. For example, when working with NMR, chemical shifts can vary a lot from a deuterated solvent to another, especially when working with ¹H-NMR. The problem is that obviously, if the experimental references used were recorded in different conditions, the comparison with the sample might lead to a “false negative”. It might not be worth building an experimental dataset if one intends to run experiments in a different way that was reported. The alternative is the prediction of the molecules’ properties. Working with MS, a recent paper proposed a dereplication analysis using an extensive *in silico* fragmentation database. This database contained molecules which fragmentation patterns were predicted using a free software [57]. In the NMR field, this can be done using commercially available software

234 programs. Most of them (ACD/Labs [108], Mestrelab [109], ChemDraw [110]) are based on the
235 hierarchically ordered spherical description of environment (HOSE) [111, 112] code of the
236 atoms of the molecules (**Figure 4A**). Centered on one atom, it will look at the atoms
237 immediately surrounding it, first in a sphere of 1 bound, then in a sphere of 2 bounds, 3 bounds,
238 etc... The HOSE may thus be assimilated as a description of the environment, using a specific
239 codex (**Figure 4B**), of each atom in the molecules, at different levels represented by the radius
240 of those spheres. An example of HOSE coding is presented in **Figure 5**. To predict a chemical
241 shift value, the software will search in a dedicated database of experimental values for one
242 molecule with an atom possessing the same HOSE code, with a maximum number of spheres
243 (meaning the maximum similarity). If no results are found, then it will search again but
244 requiring a smaller sphere radius to match. For example, if nothing is found with a sphere radius
245 of 6, the radius will be narrowed down to 5, and so on. When matches are found, all chemical
246 shifts from those atoms possessing the same HOSE code are collected and a statistical
247 calculation, sometimes as simple as a mean, leads to the predicted chemical shift for this atom.
248 The process is repeated for each atom of the molecule, leading to a full set of predicted chemical
249 shifts. The prediction process is much faster than the manual collection of data, and usually just
250 requires the gathering of structures of interest for the dataset, and then some computational
251 time. Concerning commercial NMR prediction software programs, it has been shown that their
252 accuracy is enough to start a dereplication work, but with notable differences between them
253 [113].

254

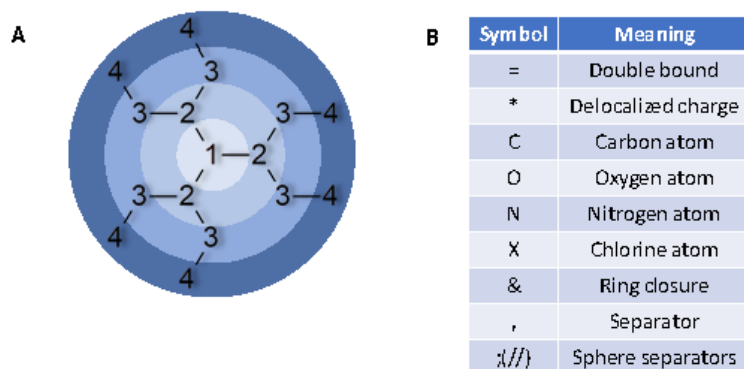


Figure 4: (A) Different spheres of the HOSE code and (B) extract of the HOSE codex.

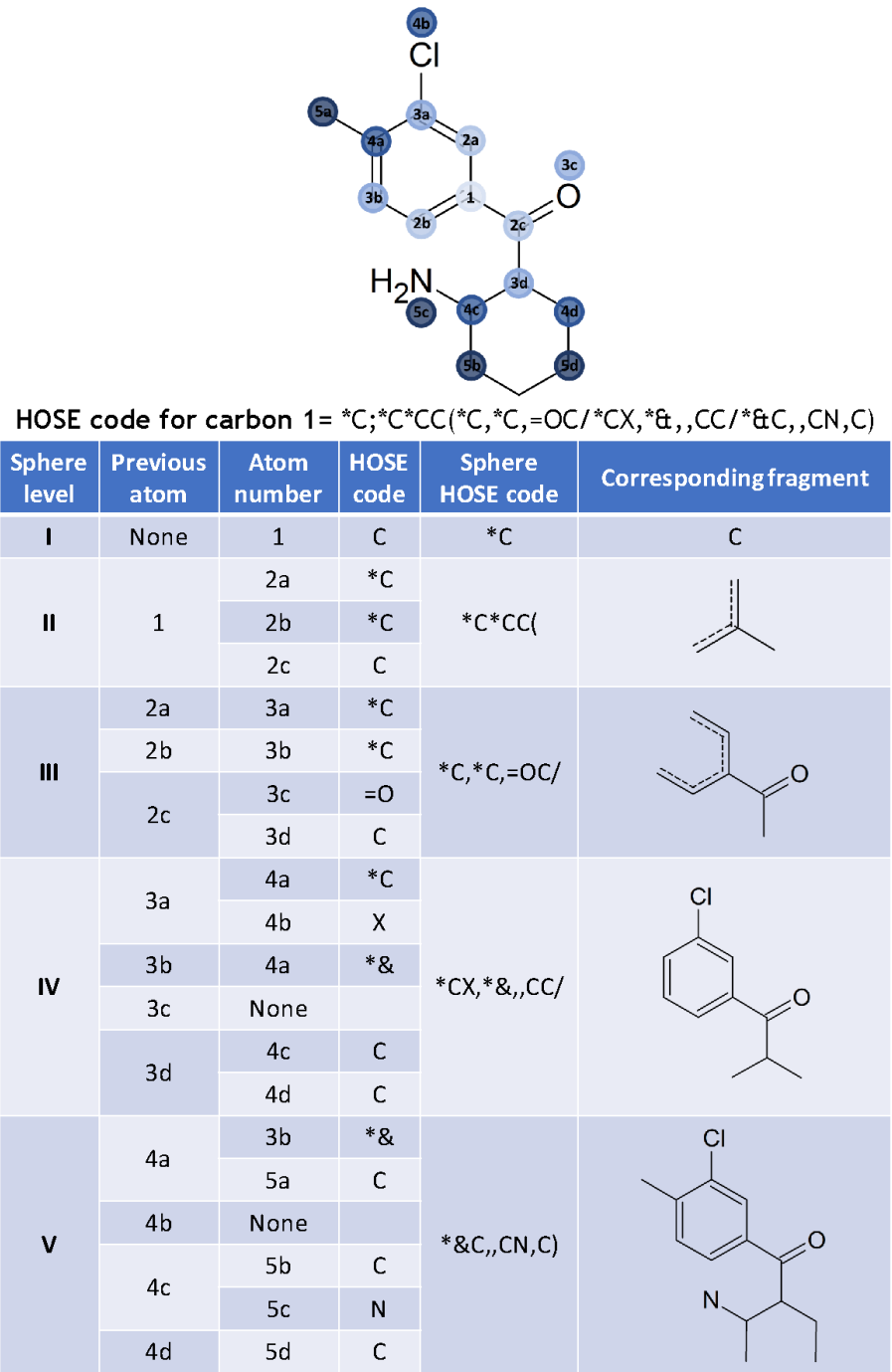


Figure 5: HOSE coding for carbon 1 of the example molecule [112]

A highlight on ^{13}C -NMR based dereplication methods

260 The set of molecules selected for comparison is often chosen upon chemotaxonomic criteria
261 (all molecules reported in a family, genus or species will be gathered) [100, 114-123] but the
262 references can also be a certain type of structure, which presence in the sample was hinted. One
263 can also decide that the goal is to get as many references as possible, whatever their origin or
264 their structural class. Those different ways of proceeding can lead to sets of molecules ranging
265 from around 50 up to several thousands of compounds. The size of the dataset greatly influences
266 the dereplication results: too big of a dataset, and the information can be “drowned” in this huge
267 number of molecules; too small of a dataset, and the molecules of interest might not be present
268 inside.

269

270 Commercial DBs exist and can either help build an appropriate DB for the dereplication or
271 some of them can directly be used for a dereplication study (CH-NMR-NP [124], MarinLit
272 [125], NMRShiftDB2 [126], ...) [7, 127, 128]. None are exhaustive, but they can gather different
273 type of information, ranging from structures only to the complete spectral profile of a molecule
274 (LC, UV, MS, NMR information) [125]. They also cover diverse fields such as NPs whatever
275 their origin (Dictionary of Natural Products [97], CH-NMR-NP [124]), marine products
276 (MarinLit [125]), polyphenolic compounds (MetIDB [129]), metabolites [the human
277 metabolome database (HMDB) [130]] or more generalist databases (NMRShiftDB2 [126],
278 SciFinder [21]) to cite a few. Such data are either experimental [124] or predicted [21]. Those
279 DB or the ones created from them by researchers can sometimes combine both experimental
280 and predicted data [117, 121, 129].

281

282 Once the choice of the type of data (experimental or predicted, type of structures, size of the
283 dataset) is made, all references are usually gathered in a “database” or “library” before the next
284 step: the comparison.

285 **Comparison tools**

286

287 Depending on which data [i.e. experimental (acquisition conditions and parameters) *versus*
288 predicted (chosen prediction method)] is used for comparison with the references, and the way
289 (i.e. manual versus dedicated algorithms) the comparison itself is carried out, the results can be
290 very different. As previously mentioned, the quality of the data acquisition and processing are
291 extremely important since that will define what is considered as a signal and thus submitted for
292 comparison. As an example, one cannot expect to detect the minor compounds in the sample
293 using NMR analysis, the concentration of the analytes inside or the dynamic range problem that
294 can face NMR are not taken into account [131].

295

296 For the comparison between experimental data and references, a manual work is possible but
297 nowadays, most people will use query algorithms to do the job. Those algorithms are also
298 available commercially, usually paired with the aforementioned DB software or websites (e.g.
299 ACD/Labs [108], NMRShiftDB2 [126], MarinLit [125]). Each query algorithm has its own
300 parameters, more or less complex: some of them allow the search for UV, NMR and MS data
301 at the same time, others restrict you to ¹H-NMR only for example [128]. Some parameters can
302 also help, when using a rather large DB, to narrow down the results to a specific species, saving
303 the trouble of filtering them beforehand (restricting the search to a particular molecular weight
304 or a structural class for example). The downside of those programs is that they act like a black
305 box, often making it difficult to know how the matching between experimental data and
306 reference is handled. This led several teams of researchers to create their own query algorithms
307 so that it can work as they intend it to, and process as many information as they require [122,
308 132-135].

309

310 During the comparison step, only “partial information” can be processed, meaning only some
311 of the observed signals are used. This is the case when working with statistical analysis like
312 HCA [100] that will cluster the information: the data will be queried cluster by cluster (*cf.* **The**
313 **CARMEL dereplication process**). Another example is NMR, where the peak picking
314 threshold can be first placed such as to only select major peaks, search for their chemical shifts,
315 and then do the same work for minor peaks. This can allow a “virtual chromatography” of the
316 mixture, separating major compounds from the minor ones, thus simplifying the work of signal
317 comparison. Finally, some teams even look for “minimal information”, but specific one, such
318 as a correlation in 2D NMR or a chemical shifts of an atom, very characteristic of a particular
319 molecule or structural class [25, 105-107]. This last technique is mainly used either when the
320 team already has an idea of the type of compound in the sample, or when different types of
321 analytical tools were used, allowing to compare complementary set of data coming from
322 different sources (MS and NMR for example).

323

324 Last but not least, the main parameter for the comparison, is the accuracy. When are a signal
325 from the sample and a signal from the reference considered one and the same, and can thus be
326 matched? This value, representing the error margin, sometimes called “looseness factor” [108,
327 135] can most of the time be defined by the users, even when using commercial DB. The
328 accuracy will depend on the type of reference used (experimental or predicted), the similarity
329 between the experiments conducted on the samples and the ones on the references (e.g.
330 deuterated solvents sensibility), and the resolution given by the used analytical tool and by the
331 one from the references.

332 **Confirmation**

333

334 All these steps will lead to several hypotheses on the composition of the sample. The researcher
335 usually needs to confirm those hypotheses, either by comparison with additional reference data,
336 by a new analysis using a different analytical method, or by comparison with a standard
337 allowing to identify the molecules structures, depending on the level of metabolite identification
338 required by the team [136].

339 When an unambiguous identification of the structure is needed, ^{13}C -NMR might be the best
340 analytical method to do so [24, 31, 33, 34, 36, 37, 41, 43, 45, 49, 60-63], by comparing the signals
341 of the sample with those of one of the hypothetic molecules, reported in the same deuterated
342 solvent. Indeed, MS tends to struggle to differentiate stereoisomers, and the overlapping signals
343 in ^1H -NMR can make a precise comparison difficult. ^{13}C -NMR is thus the mainly used
344 experiment for structure confirmation.

Sample	Analytical tools				References		Comparison tool	Confirmation
Extract	Chromatographic system				Experimental data		Manual	Manual comparison of ¹³ C-NMR data
	HPLC	UPLC	SFC	CPC	GC	In house data		
Fraction	Detector				Predicted data		Automated	
	UV	MS		NMR	Prediction software	Data trends		
Single compound	Statistical analysis							
	HCA			PCA				

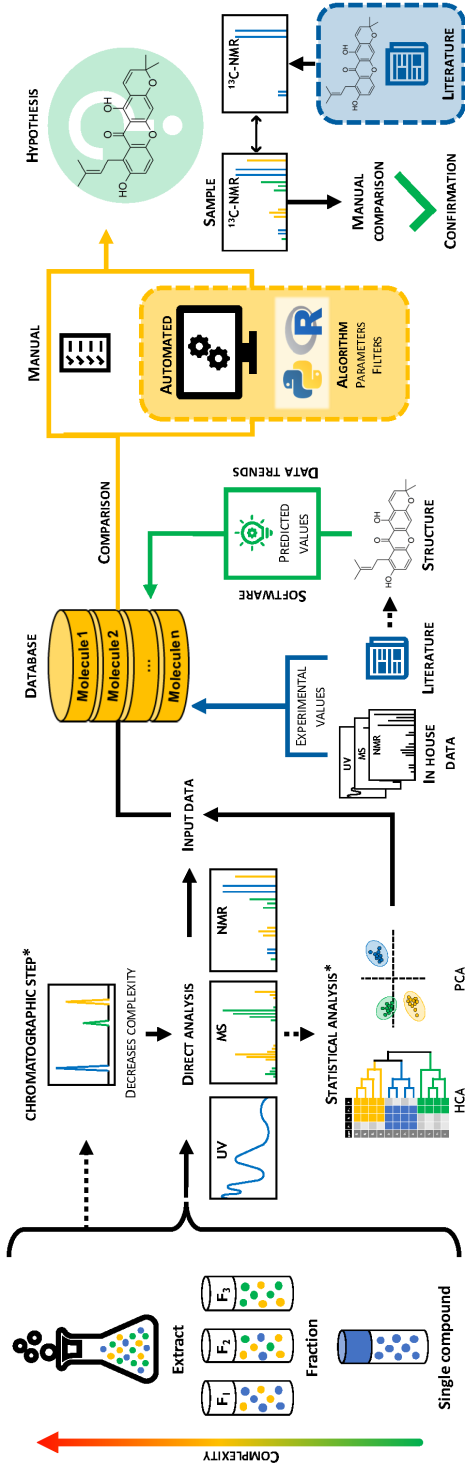


Figure 3: Different ways to conduct a dereplication analysis. From the complexity of the analyzed sample, the analytical tool used, the type of reference gathered for the comparison step, to the final identification of the molecule. Optional steps are signaled by an asterisk.

^{13}C -NMR DEREPLICATION

This next part will present an overview of the type of dereplication methods, mainly based on ^{13}C -NMR, that are used in the literature. The advantages and drawbacks of each method will be discussed.

First automated ^{13}C -NMR dereplication

To our knowledge, the first ^{13}C -NMR based dereplication method was proposed by Tomi, Casanova *et coll.* in the 1990s [137], presenting an automated alternative to the work of Kubezcka *et al.* [138]. This work was mainly focused on the composition of essential oils but also plant extracts and resins containing terpenes. A library of experimental ^{13}C -NMR chemical shifts (δc) was created by gathering the ones of pure mono- and sesquiterpenes whose spectra were directly recorded in the laboratory, in the same deuterated solvent. The DB counted hundreds of NPs.

During the process, a ^{13}C -NMR spectrum of the extract of interest is recorded. A home-made software allows the identification of the terpenes, based on three parameters. First, for each molecule in the database, the number of signals that can be found in the extract is compared to the number of expected carbons. Secondly, for each signal, the difference between the value of the experimental chemical shift and the one of the references is calculated. Thirdly, the number of fortuitously overlapped signals that belong to two different molecules but possessing the same chemical shift value is noted.

A highlight on ^{13}C -NMR based dereplication methods

372 This allows the identification of terpenes in very complex mixtures, even the minor ones (up to
373 0.1%). For most of the NPs, the difference between experimental chemical shifts and the ones
374 of the references are 0.1 ppm or less. Usually, only some of the quaternary carbons from the
375 minor compounds are not visible, but it does not impact negatively the whole dereplication
376 process. An analysis using GC-FID is also usually done to provide further information when a
377 quantitative analysis of the essential oil is needed.

378

379 After validation of the method on artificial mixtures, this technique is now used as a routine
380 dereplication method in this laboratory [[101](#), [139-141](#)].

381

382 This method's strength is that the experiments and the references are conducted using the same
383 conditions (*i.e.* same instrument, same deuterated solvent for each experiment), making the
384 error margin very narrow (0.1 ppm), and thus ensuring that the matched peaks are the correct
385 ones. Moreover, taking in account the fact that δ_{C} of equivalent carbons will be identical. The
386 method avoids a "loss of carbons" signals during the matching process. Indeed, if every signal
387 from the spectra is considered belonging to only one carbon, there will be one carbon "missing"
388 for every equivalent carbon in the database's molecule. The software and the database were not
389 published.

390

391 **SISCONST dereplication program**

392

393 In 2001, in order to identify the NPs inside a mixture of triterpenes, the group of Ferreira started
394 by creating a database containing the ^{13}C -NMR shifts reported in the literature for 1300 mono
395 and 2500 sesquiterpenes [[122](#)].

396

397 In this work, ^{13}C -NMR and DEPT experiments of the samples were recorded. From them,
398 chemical shifts and multiplicities (i.e. CH_3 , CH_2 , CH and Cq) of the carbon atoms were
399 manually listed. That information was queried in a program named SISCONST, developed by
400 the team for the matching process. It also required that a minimal number of carbons to match
401 was chosen by the user. Before the search, the user could also specify the potential class of NP
402 (i.e. monoterpene or sesquiterpene) he was looking for. The program then compared the δ_{C} from
403 the experimental data with those of the terpenes in the DB with an error margin of 1.0 ppm.
404 When no results were found, the error margin is automatically increased by steps of 0.5 ppm to
405 reach 3.0 ppm. A score called “statistical weight” was calculated for each molecule, taking into
406 account the number of associated signals and the error margin for each associated signal. The
407 resulting value allowed the sorting of the different NPs by probability of presence.
408 This method managed to identify 2 monoterpenes and 8 sesquiterpenes in a *Piper cernuum* leaf
409 essential oil [122] as well as 2 monoterpenes and 3 sesquiterpenes in a *P. regnellii* leaf essential
410 oils [122]. Those compounds were confirmed by a careful analysis of the δ_{C} from the initial ^{13}C
411 spectra of the extracts. Those NPs had a concentration of 2% or more in the essential oil. All
412 major compounds were thus identified, and more than 80% of the signals from the original
413 spectra were associated.
414
415 In this process, rather than just being based on the comparison of δ_{C} , the addition carbon
416 multiplicities, information carried by the DEPT data, narrows down the possibilities. Since the
417 references in the DB are not necessarily reported in the same deuterated solvent, it is important
418 for the error margin to be larger than in the previously mentioned method [137]: this time, it
419 ranges from 1.0 to 3.0 ppm. The minimal number of carbons to match is an interesting feature
420 but can also be a double-edge sword if the composition of the mixture is unknown, or if it
421 contains both small and big compounds.

Unfortunately, the program itself was not published.

The CARMEL dereplication process

More recently, Hubert *et al.* has developed a the dereplication method called CARMEL for *CAR*actérisation des *MÉ*Langes in French (*i.e.* Mixture characterization) [100, 121]. As depicted in **Figure 6**, the first step consists in the fractionation of a crude extract using centrifugal partition chromatography (CPC), leading to usually at least 10 fractions. The goal here is double: of course, simplify the analyzed mixtures, but also spreading compounds across several consecutive fractions, which is important for the statistical analysis that will come later in the process. The ^{13}C -NMR spectra of each fractions are recorded, and the table of their chemical shifts and intensities are collected. A bucketing step allows the simplification of the data: the spectrum is divided into buckets of 0.2 ppm and only the most intense signal in every bucket, if there is one, is kept for further processing. A hierarchical clustering analysis (HCA) is then conducted on this bucketed data using PermutMatrix, a software freely available online [142]. The parameters chosen for HCA are the use of Euclidian distance to measure the proximity of the samples and Ward's method for the agglomeration step. This pattern recognition method results in a 2D map showing bins of chemical shifts as the first dimension and the different fractions as the second dimension. Clusters of δ_{C} are formed on this map, depending on the localization of these δ_{C} in the fractions, but also on the intensity of the signals. Indeed, this method relies on the gaussian elution of the different compounds during the chromatographic step. On the 2D map, the intensity of δ_{C} is represented by the brightness of each point. Ideally, one cluster corresponds to one NP: it is then possible to extract its δ_{C} for the comparison step.

A highlight on ^{13}C -NMR based dereplication methods

447 As references, the team usually mixes two types of DB together [117, 120, 121]: one containing
448 diverse NPs, and another one, more specific of the species or genre of the plant they are working
449 on. The δ_{C} in this DB are mostly experimental values [100, 118], gathered from the literature.
450 But they can also be predicted if no experimental data was found. The prediction is made using
451 an ACD/Labs commercial software, namely CNMR Predictor [100, 108, 121].

452

453 Hence, for each cluster, one searches in this DB for bucketed δ_{C} corresponding to one cluster,
454 using the query algorithm proposed by ACD/Labs [108]. The software also requires an error
455 margin, usually selected between 1.0 ppm and 2.0 ppm [117, 120, 121], and a minimum number
456 of matching signals, which is set to 80% [121]. The program then presents the NPs most fitting
457 to this cluster. The initial data was then compared with the chemical shifts reported for the
458 proposed structures to confirm their identity. Published dereplication studies following this
459 method gave good results on the wide range of plant extracts, identifying every time the major
460 constituents of the mixtures within the first ranks [100, 114-121].

461

462 There are a few drawbacks from this method that are mainly linked to the statistical analysis,
463 meaning the bucketing and HCA steps. The data is bucketed in order to simplify the initial
464 spectrum, thus losing some of the carbon signals in the process. Moreover, the δ_{C} indicated on
465 the 2D cluster map are not real chemical shifts, but just the value of the beginning of each bin,
466 forcing the user to go back to the original data for comparison. The bucketing can unfortunately
467 also split a signal in two different buckets. Concerning the HCA, it can only work if the
468 molecules are well spread among several fractions, and not necessarily separated in the best
469 way possible. It can seem a bit counter-productive in an isolation/purification objective. HCA
470 also relies on signal intensity to establish clusters, which can be misleading: For example,
471 quaternary carbons of a molecule usually have a lower intensity than the rest of the molecule

A highlight on ^{13}C -NMR based dereplication methods

and will thus be separated from the main cluster. The intensity is also the one resulting from the bucketing process (meaning the most intense signal that was found in the bucket), explaining that some NPs are split into different clusters. The other way around, several molecules can also be clustered together. Even without any kind of problem, the clusters are not perfectly shaped [113, 115, 117], making it difficult to define their limits. In these cases, it can be harder to properly identify the structures of the molecules. Finally, the team also noted that if NPs are present inside a same fraction, they are more easily discriminated if they belong to different structural classes.

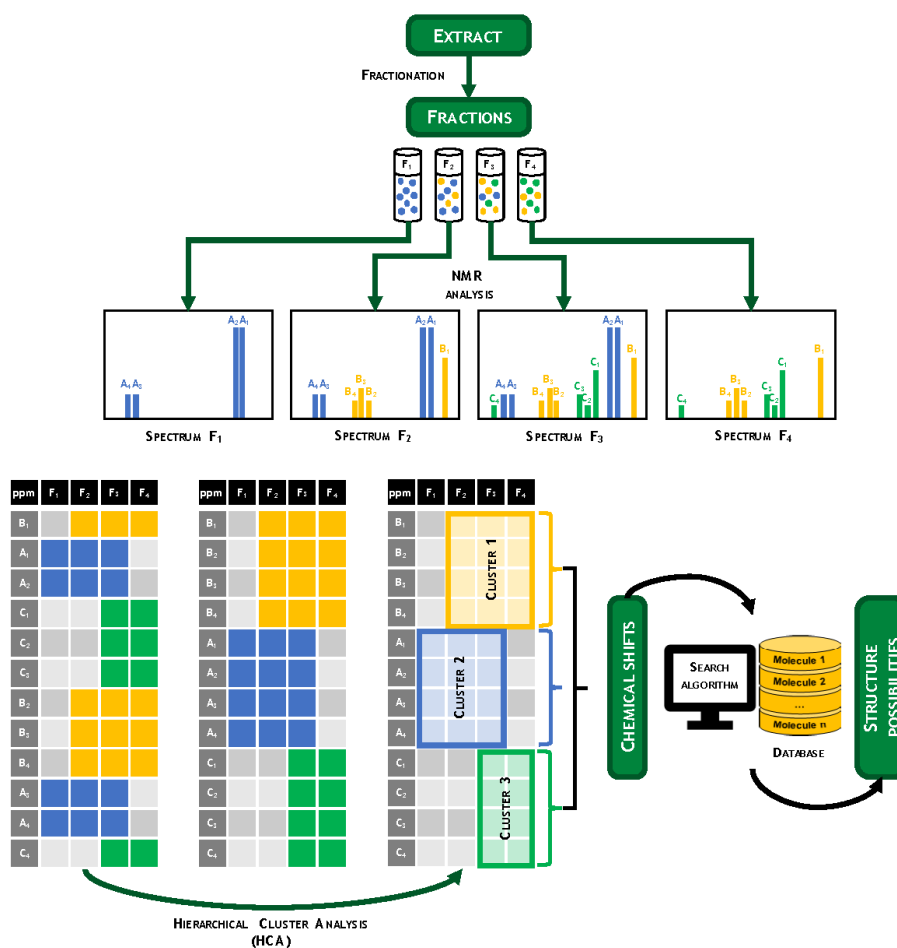


Figure 6: Dereplication process as conducted by the team of J. Hubert

The DerepCrude script for dereplication of crude extracts

In 2017, the method developed by Bakiri *et al.* in the same group presents a dereplication workflow without fractionation [135]. The crude extract chosen as proof of concept was an alkaloid totum from *Peumus boldus*.

After gathering the structure of the 58 NPs reported in the *Peumus* genus, their chemical shifts were predicted using ACD/Labs CNMR Predictor [108] in order to create the DB used for this example. After recording ^{13}C -NMR spectrum, a list of the δ_{C} and their respective intensities were submitted as a text file to a home-made search algorithm called DerepCrude.

The algorithm was written using the programming language Python 2.7. The latter aimed at comparing the experimental δ_{C} with the ones predicted for each NP from the DB. In order for a signal from the spectrum and one from the DB to be associated, their values must be apart from less than the user-defined error margin (± 1 ppm are recommended) and their corresponding intensities must be within the user-defined tolerance (default parameter is twice the standard deviation around the mean intensity of the previously matched signals). Every time a match between experimental data and a compound from the DB was possible, a score was assigned to the molecule. It was calculated as the number of matched carbons for the NP out of the total number of carbons expected for the latter. The score will thus be between 0 and 1. Matched molecules were sorted by decreasing score and saved as a text file, containing the name of the molecule and the matched signals, as well as an image file showing the structure of the matched molecules, along with their name and score (Figure 7).

A highlight on ^{13}C -NMR based dereplication methods

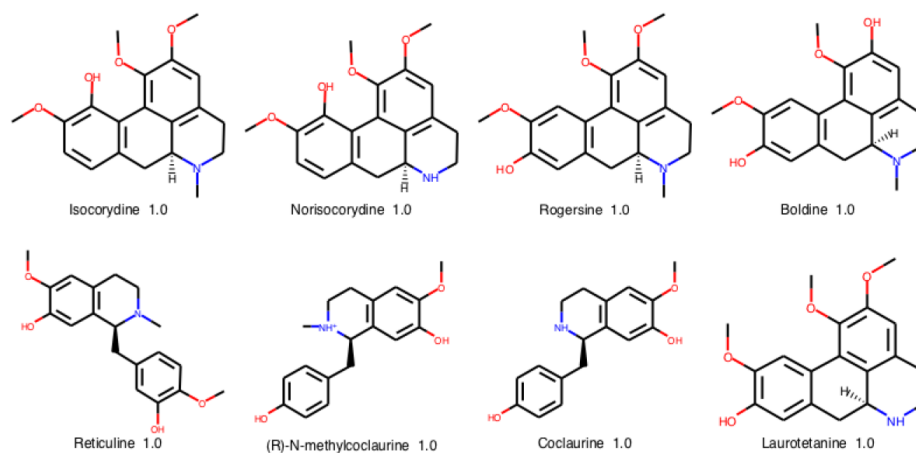


Figure 7: Results image file with the DerepCrude script

In the published example, the program placed 33/58 molecules of the DB with a score higher than 0.9, and 28/58 with a score of 1. Among them, 8 alkaloids were indeed present in the extract, confirmed by GC-MS analysis [135].

This work presented an interesting alternative method to work directly on crude extract without any fractionation. However, the crude extract should be of low complexity regarding the number of molecules and the diversity of the structural classes, and the DB must be adapted to the correct chemical class in order for it to work.

The Python script, as well as the DB and experimental data used in this example were made available in the supporting information.

520 **MixONat software for mixtures dereplication based on ^{13}C NMR and DEPT experiments**

521

522 After a careful analysis of benefits and drawbacks of previous published work on automated
523 dereplication based on ^{13}C -NMR, we also proposed a free available software, mainly based on
524 the improvements of the results after the addition of DEPT data. The matching process was also
525 reworked. It was successfully tested on various examples of increasing complexity, in order to
526 validate the method [143].

527

528 ^{13}C -NMR data is required, and optional DEPT 135 and DEPT 90 spectra can be added for
529 additional information. The information provided by the DEPT 135 greatly improves the
530 dereplication results, as it allows the software to match carbons based on their chemical shift
531 and multiplicity [104]. The DBs are usually based on chemotaxonomy features and thus contain
532 the NPs reported from the family or the genus of the sample; the goal is to reach a reasonable
533 DB size (between 100 and 2000 molecules), to avoid “drowning the information” with too large
534 of a DB or missing out on a compounds with a small DB. The chemical shifts are then predicted
535 using the ACD/Labs CNMR Predictor software [108]. We do not rely on DB of experimental
536 shifts since it is unfortunately difficult to obtain. Our previous work showed that predicted
537 chemical shifts are accurate enough for a dereplication study [113].

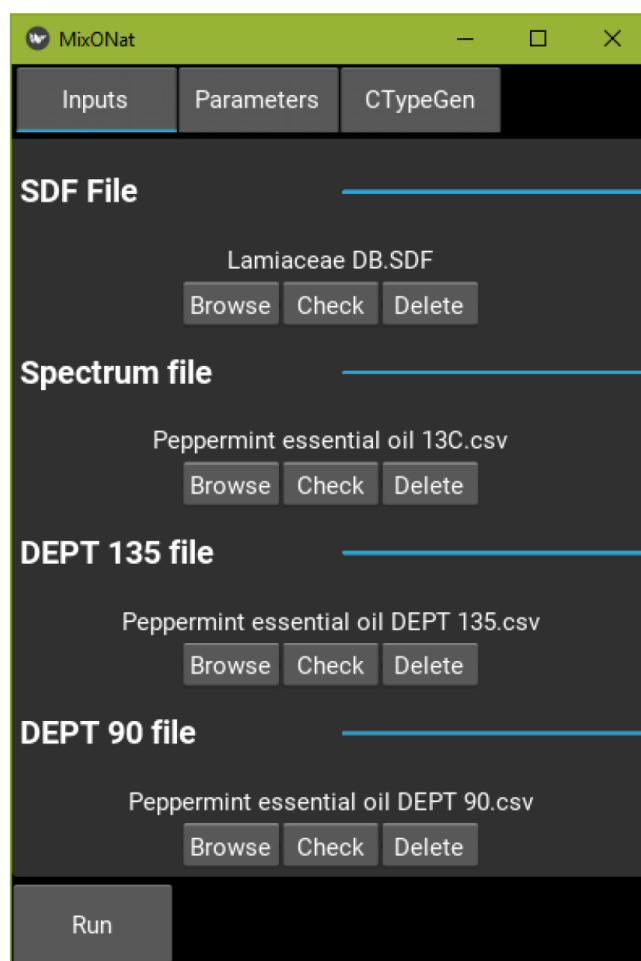
538

539 A locally written program, coded using Python 3.6, has a graphic user interface (**Figure 8**) to
540 be more user-friendly. ^{13}C -NMR, DEPT 135 and DEPT 90 data [122] are imported as a table
541 (.csv file) in the software. The program can also work with just ^{13}C -NMR spectrum or with ^{13}C -
542 NMR and DEPT 135 spectra if not all the experiments were conducted. The error margin used
543 for the matching process is the main parameters for the search: it reflects the accuracy of the
544 DB that is used. We recommend 1.3 ppm for a DB created with ACD/Labs because it is the

A highlight on ^{13}C -NMR based dereplication methods

545 evaluated accuracy of this prediction software [113]. If an experimental DB (in the same
546 solvent) is used instead, the error margin should be reduced accordingly. The other
547 parameters/filters include the accuracy of the alignment between carbon and DEPT data, a
548 molecular weight filter (that can be very useful if additional MS data are available) and a way
549 to allow a single signal to be matched several times (in the case of equivalent carbons) [137].

550



551

552 **Figure 8:** MixONat window for spectra and database inputs

553

A highlight on ^{13}C -NMR based dereplication methods

554 Then the algorithm proceeds to the matching process, matching the carbon signals by carbon
555 type (*e.g.* quaternary carbons with quaternary carbons), searching first for a perfect match (error
556 margin = 0 ppm) and incrementally increasing the error margin, 0.1 by 0.1 ppm, until it reaches
557 the one chosen by the user (1.3 ppm for example). This ensure that the closest chemical shifts
558 are matched together, which is even more important when working with experimental DB. The
559 incrementation is also one of the improvements that allows a better matching. For each
560 molecule in the DB, the score is calculated as the number of matched carbons out of the number
561 of expected carbons for the molecule.

562

563 The results are first presented as an interactive window (**Figure 9**) in the software with the
564 molecules ranked by decreasing score, with their name and structure. The user can click on the
565 NPs to obtain additional information such as the list of matched chemical shifts, their
566 visualization on the structure, as well as a graphical representation of the intensity of the
567 matched peaks. Each chemical shift and/or molecule can be added or deleted from the list, in
568 order to optimize the results. For example, if a signal was not matched because it was not picked
569 in the original spectra, it can be added, or if a molecule is considered a match but it is not a NP
570 (DB problem), it can be removed. Results can then be saved as a text and image file.

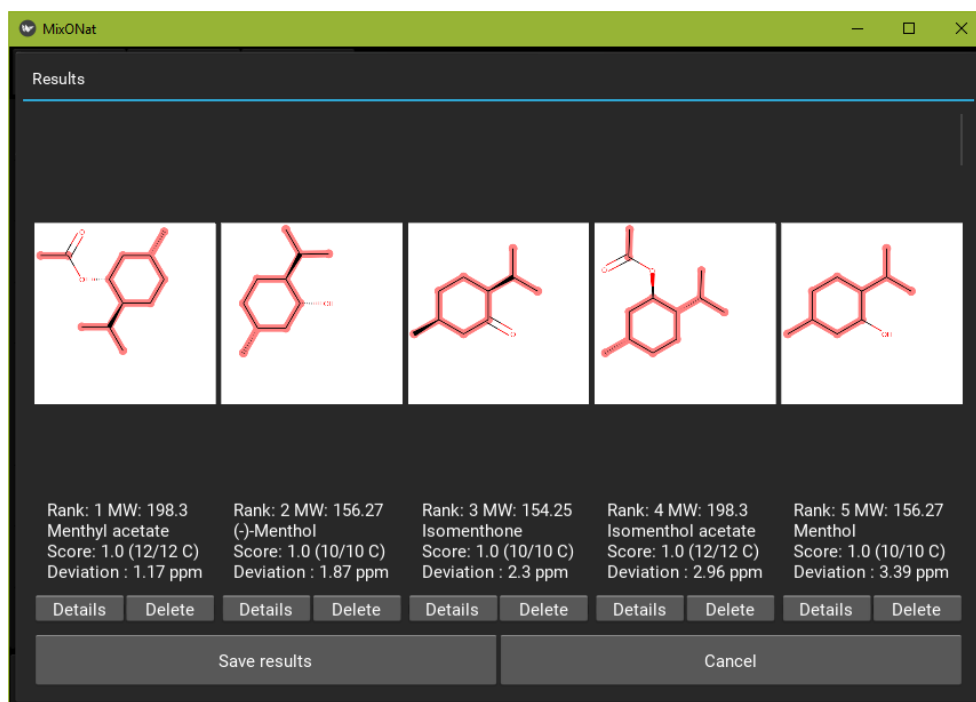


Figure 9: MixONat interactive results window

The method was tested on several examples: a mixture of two alkaloids from *Papaver somniferum*, extracts from *Rosmarinus officinalis*, a crude extract and a fraction from *Garcinia mangostana*. They showed the limits of the method: a dereplication work on crude extract is possible but with relatively simple extracts (number of molecules and/or structure complexity). As for every other dereplication method, it was shown that the quality of the DB is also a really important point, and one must thus be careful when building it.

The program was published along with one of the DB that was used for an example.

ABOUT 2D NMR

As for metabolomics purposes [15], researchers are also starting to turn to 2D NMR exclusive methods in dereplication strategies. The work of Bakiri *et al.* published in 2018 illustrates this emerging concept [134].

HMBC spectra of the samples were recorded and cross peaks were listed. A DB of 2D NMR predicted data was created: the proton and carbon chemical shifts of 2 800 selected structures of NPs were predicted using the ACD/Labs software [108]. A local Python script allowed the team to rework the database by linking ^1H shifts, ^{13}C shifts and multiplicity for each atom of the molecules, thus creating the required 2D DB. Another algorithm builds, from the experimental cross peak list, a clustering network by using the Louvain detection algorithm [144]. HMBC correlations are isolated from each individual cluster and queried in the 2D NMR DB by a matching algorithm. The parameters for the error margin are 1.5 ppm for the ^{13}C chemical shift value and 0.15 ppm for the ^1H chemical shift value. For each cluster, the algorithm ranks the NPs from the DB based on their decreasing score which is the percentage of found signals for each molecule. For the top 20 compounds in each cluster, HMBC data and carbon parity are compared in order to confirm the hits.

The proof of concept was conducted on a controlled mixture of 3 known NPs and was then tested on an extract from *Picea abies*. The method was successful on both examples but encountered a few problems. First, as could have been expected, was the presence of a lot of overlapping signals in the proton dimension, making the clustering a difficult task. The optimal parameters used for the clustering were also hard to find and the researchers had to resort to “trial and error” to obtain the best results. The clustering analysis also separated a single

A highlight on ^{13}C -NMR based dereplication methods

607 molecule in different clusters, *i.e.* the aglycone and its sugar moiety. The last issue was the
608 HSQC analysis that seemed to give a high, if not perfect (100%), score for a lot of NPs, and
609 thus not being as discriminative as expected.

610

611 Unfortunately, neither the programs nor the DB were made available for the moment.

612

613 CONCLUSION

614

615 Dereplication methods based on ^{13}C -NMR provide an interesting alternative method to LC-MS
616 based dereplication. Although less sensible, ^{13}C -NMR is a more universal method, contrary to
617 MS where the detection of the metabolites depends on their ionization. Its analytical conditions
618 are also easier to set up: no LC method needs to be optimized, and no ionization parameters
619 needs to be adjusted. ^{13}C -NMR will also allow the differentiation of stereoisomer, and the
620 identification of metabolites with a high level of confidence [145].

621

622 This explains why researchers turn to ^{13}C -NMR for their dereplication study. Each team
623 published papers where their methods proved to be working on their own different examples.
624 But none of the proposed methods are universal. In general, when choosing a dereplication
625 method, the researcher must have in mind its objectives, and the available tools at his disposal.
626 The quality of the experimental data and the used references must fit these objectives.

627

628 Regarding the complexity of the sample, the ^{13}C -NMR dereplication on crude extract seems
629 possible but really depends on the type of extract. A fractionation step may be required if no
630 interpretable results are obtained. The statistical tools used for big data simplification can be
631 used for NMR dereplication processes: it allows to bring out trends in datasets or to cluster data,

simplifying the comparison step. But their parameters are usually difficult to set. Parameters used for one sample may also not work on another one, especially for ^{13}C -NMR where, for example, the intensity of the signals is something really difficult to work with (mixtures of compounds in different proportions, equivalent carbons).

As far as DB are concerned, predicted DBs represent a faster and interesting alternative to experimental DB: the accuracy of most prediction software using HOSE codes is enough to allow dereplication work. However, the prediction tools themselves are not available for free at the moment. Unfortunately, it is also difficult to improve the existing DB, used by the software for the prediction, with our own experimental data. Since the prediction is based on the software's own experimental DB, we could imagine enriching this DB with our own experimental data, thus improving the quality of future predictions.

Regarding the comparison of spectral data with the ones of a DB, a computer-aided method seems like the best and fastest way to proceed. But it is important as a scientist to have access to the inner working of the algorithm to understand how exactly the matching is made. The user should also be able to tweak as many parameters as possible, and process as much information as he possesses. It is noteworthy that, over the last few years, some algorithms have begun to be freely distributed. Custom-made algorithms, even if generally humbler than the commercial ones, provide this transparency regarding their inner working, as well as the possibility of further modifications, adapted to the need of the user.

Dereplication methods will continue to evolve, trying to find the most discriminating and more universal method but that should remain a fast and approachable process for every team of researchers.

BIBLIOGRAPHY

1. Hubert, J., J.-M. Nuzillard, and J.-H. Renault, *Dereplication strategies in natural product research: How many tools and methodologies behind the same concept?* *Phytochem. Rev.*, 2015. **16**(1): p. 55-95.
2. Newman, D.J. and G.M. Cragg, *Natural products as sources of new drugs from 1981 to 2014*. *J. Nat. Prod.*, 2016. **79**(3): p. 629-661.
3. Kusumawati, I. and G. Indrayanto, *Chapter 15 - Natural antioxidants in cosmetics*, in *Studies in natural products chemistry*, R. Atta ur, Editor. 2013, Elsevier. p. 485-505.
4. Parvez, S., et al., *Naturally occurring tyrosinase inhibitors: mechanism and applications in skin health, cosmetics and agriculture industries*. *Phytother. Res.*, 2007. **21**(9): p. 805-816.
5. Carrocho, M., P. Morales, and I.C.F.R. Ferreira, *Natural food additives: Quo vadis?* *Trends Food Sci. Technol.*, 2015. **45**(2): p. 284-295.
6. Cantrell, C.L., F.E. Dayan, and S.O. Duke, *Natural products as sources for new pesticides*. *J. Nat. Prod.*, 2012. **75**(6): p. 1231-1242.
7. Go, E.P., *Database resources in metabolomics: an overview*. *J. Neuroimmune Pharmacol.*, 2010. **5**(1): p. 18-30.
8. Leiss, K.A., et al., *An overview of NMR-based metabolomics to identify secondary plant compounds involved in host plant resistance*. *Phytochem. Rev.*, 2011. **10**(2): p. 205-216.
9. Tawfike, A.F., C. Viegelmann, and R. Edrada-Ebel, *Metabolomics and dereplication strategies in natural products*. *Methods Mol. Biol.*, 2013. **1055**: p. 227-244.
10. Yuliana, N.D., et al., *Metabolomics for the rapid dereplication of bioactive compounds from natural sources*. *Phytochem. Rev.*, 2013. **12**(2): p. 293-304.
11. Bingol, K., et al., *Unified and isomer-specific NMR metabolomics database for the accurate analysis of $(13)\text{C}$ - $(1)\text{H}$ HSQC spectra*. *ACS Chem. Biol.*, 2015. **10**(2): p. 452-459.
12. Wu, C., et al., *Metabolomics-driven discovery of a prenylated isatin antibiotic produced by *Streptomyces* species MBT28*. *J. Nat. Prod.*, 2015. **78**(10): p. 2355-2363.
13. Palama, T.L., et al., *Identification of bacterial species by untargeted NMR spectroscopy of the exo-metabolome*. *Analyst*, 2016. **141**(15): p. 4558-4561.
14. Ahmad, S.J., et al., *Discovery of antimalarial drugs from *Streptomyces* metabolites using a metabolomic approach*. *J. Trop. Med.*, 2017. **2017**: p. 2189814.
15. Bingol, K. and R. Bruschweiler, *Knowns and unknowns in metabolomics identified by multidimensional NMR and hybrid MS/NMR methods*. *Curr. Opin. Biotechnol.*, 2017. **43**: p. 17-24.
16. Tawfike, A.F., et al., *Metabolomic tools to assess the chemistry and bioactivity of endophytic *Aspergillus* strain*. *Chem. Biodivers.*, 2017. **14**(10).
17. Consonni, R. and L.R. Cagliani, *The potentiality of NMR-based metabolomics in food science and food authentication assessment*. *Magn. Reson. Chem. : MRC*, 2018.
18. Parrot, D., et al., *Mapping the surface microbiome and metabolome of brown seaweed *Fucus vesiculosus* by amplicon sequencing, integrated metabolomics and imaging techniques*. *Sci. Rep.*, 2019. **9**(1): p. 1061.
19. Robinette, S.L., et al., *NMR in metabolomics and natural products research: two sides of the same coin*. *Acc. Chem. Res.*, 2012. **45**(2): p. 288-297.
20. Scopus. [cited 2019; Available from: <https://www.scopus.com/>.]
21. SciFinder. [cited 2019; Available from: <https://scifinder.cas.org/>.]
22. Abbet, C., et al., *Comprehensive analysis of *Phyteuma orbiculare* L., a wild Alpine food plant*. *Food Chem.*, 2013. **136**(2): p. 595-603.

23. Abbet, C., et al., *Comprehensive analysis of Cirsium spinosissimum Scop., a wild alpine food plant*. Food Chem., 2014. **160**: p. 165-170.
24. Petersen, L.M., et al., *Dereplication guided discovery of secondary metabolites of mixed biosynthetic origin from Aspergillus aculeatus*. Molecules, 2014. **19**(8): p. 10898-10921.
25. Brkljaca, R., E.S. Gker, and S. Urban, *Dereplication and chemotaxonomical studies of marine algae of the Ochrophyta and Rhodophyta phyla*. Mar. Drugs, 2015. **13**(5): p. 2714-2731.
26. Brkljaca, R. and S. Urban, *HPLC-NMR and HPLC-MS investigation of antimicrobial constituents in Cystophora monilifera and Cystophora subfarcinata*. Phytochemistry, 2015. **117**: p. 200-208.
27. Brkljaca, R. and S. Urban, *HPLC-NMR and HPLC-MS profiling and bioassay-guided identification of secondary metabolites from the Australian plant Haemodorum spicatum*. J. Nat. Prod., 2015. **78**(7): p. 1486-1494.
28. Cakova, V., et al., *Identification of phenanthrene derivatives in Aerides rosea (Orchidaceae) using the combined systems HPLC-ESI-HRMS/MS and HPLC-DAD-MS-SPE-UV-NMR*. Phytochem. Anal., 2015. **26**(1): p. 34-39.
29. Dang, B.T., et al., *Dereplication of Mammea neurophylla metabolites to isolate original 4-phenylcoumarins*. Phytochem. Lett., 2015. **11**: p. 61-68.
30. Dang, B.T., et al., *Identification of minor benzoylated 4-phenylcoumarins from a Mammea neurophylla bark extract*. Molecules, 2015. **20**(10): p. 17735-17746.
31. Favre-Godal, Q., et al., *Anti-Candida cassane-type diterpenoids from the root bark of Swartzia simplex*. J. Nat. Prod., 2015. **78**(12): p. 2994-3004.
32. Li, Z., et al., *A novel dereplication strategy for the identification of two new trace compounds in the extract of Gastrodia elata using UHPLC/Q-TOF-MS/MS*. J. Chromatogr. B Analyt. Technol. Biomed. Life Sci., 2015. **988**: p. 45-52.
33. Ramli, R., N.H. Ismail, and N. Manshoor, *Identification of oligostilbenes from Dipterocarpus semivestitus through dereplication technique*. Jurnal Teknologi, 2015. **77**(2).
34. Ravu, R.R., et al., *LC-MS- and ¹H NMR spectroscopy-guided identification of antifungal diterpenoids from Sagittaria latifolia*. J. Nat. Prod., 2015. **78**(9): p. 2255-2259.
35. Wolfender, J.-L., et al., *Current approaches and challenges for the metabolite profiling of complex natural extracts*. J. Chromatogr. A, 2015. **1382**: p. 136-164.
36. Ióca, L.P., et al., *A strategy for the rapid identification of fungal metabolites and the discovery of the antiviral activity of pyrenocine a and harzianopyridone*. Química Nova, 2016.
37. Li, P., et al., *UPLC-QTOFMS-guided dereplication of the endangered Chinese species Garcinia paucinervis to identify additional benzophenone derivatives*. J. Nat. Prod., 2016. **79**(6): p. 1619-1627.
38. Chervin, J., et al., *Targeted dereplication of microbial natural products by high-resolution MS and predicted LC retention time*. J. Nat. Prod., 2017. **80**(5): p. 1370-1377.
39. Kang, K.B., et al., *Ceanothane- and lupane-type triterpene esters from the roots of Hovenia dulcis and their antiproliferative activity on HSC-T6 cells*. Phytochemistry, 2017. **142**: p. 60-67.
40. Gomes, N.G.M., et al., *Hybrid MS/NMR methods on the prioritization of natural products: applications in drug discovery*. J. Pharm. Biomed. Anal., 2018. **147**: p. 234-249.
41. Grecco, S.S., et al., *Neolignans isolated from twigs of Nectandra leucantha Ness & Mart (Lauraceae) displayed in vitro antileishmanial activity*. J. Venom Anim. Toxins Incl. Trop. Dis., 2018. **24**: p. 27.
42. Khushi, S., et al., *Cacolides: sesterterpene butenolides from a Southern Australian marine sponge, Cacospongia sp.* Mar. Drugs, 2018. **16**(11).

43. Kimani, N.M., et al., *Sesquiterpene lactones from Vernonia cinerascens Sch. Bip. and their in vitro antitrypanosomal activity*. *Molecules*, 2018. **23**(2).
44. Kristoffersen, V., et al., *Characterization of rhamnolipids produced by an Arctic marine bacterium from the Pseudomonas fluorescens group*. *Mar. Drugs*, 2018. **16**(5).
45. Shady, N.H., et al., *A new antitrypanosomal alkaloid from the Red Sea marine sponge Hyrtios sp.* *J. Antibiot. (Tokyo)*, 2018. **71**(12): p. 1036-1039.
46. Yang, J.Y., et al., *Molecular networking as a dereplication strategy*. *J. Nat. Prod.*, 2013. **76**(9): p. 1686-1699.
47. Allard, P.M., et al., *Integration of molecular networking and in-silico MS/MS fragmentation for natural products dereplication*. *Anal. Chem.*, 2016. **88**(6): p. 3317-3323.
48. Quinn, R.A., et al., *Molecular networking as a drug discovery, drug metabolism, and precision medicine strategy*. *Trends Pharmacol. Sci.*, 2017. **38**(2): p. 143-154.
49. Chervin, J., et al., *Dereplication of natural products from complex extracts by regression analysis and molecular networking: case study of redox-active compounds from Viola alba subsp. dehnhardtii*. *Metabolomics*, 2017. **13**(8).
50. Fox Ramos, A.E., et al., *Revisiting previously investigated plants: a molecular networking-based study of Geissospermum laeve*. *J. Nat. Prod.*, 2017. **80**(4): p. 1007-1014.
51. Remy, S., et al., *Structurally diverse diterpenoids from Sandwithia guyanensis*. *J. Nat. Prod.*, 2018. **81**(4): p. 901-912.
52. Hou, X.M., et al., *Integrating molecular networking and (1)H NMR to target the isolation of chrysogeamides from a library of marine-derived Penicillium fungi*. *J. Org. Chem.*, 2019. **84**(3): p. 1228-1237.
53. Tawfike, A.F., et al., *Isolation of anticancer and anti-trypanosome secondary metabolites from the endophytic fungus Aspergillus flocculus via bioactivity guided isolation and MS based metabolomics*. *J. Chromatogr. B Analyt. Technol. Biomed. Life. Sci.*, 2019. **1106-1107**: p. 71-83.
54. GNPS. Available from: <https://gnps.ucsd.edu/>.
55. Wang, M., et al., *Sharing and community curation of mass spectrometry data with Global Natural Products Social Molecular Networking*. *Nat. Biotechnol.*, 2016. **34**(8): p. 828-837.
56. Cytoscape. Available from: <https://cytoscape.org/>.
57. Allard, P.M., et al., *Integration of molecular networking and in-silico MS/MS fragmentation for natural products dereplication*. *Anal. Chem.*, 2016. **88**(6): p. 3317-3323.
58. Nothias, L.F., et al., *Les réseaux moléculaires, une approche bio-informatique globale pour interpréter les données de spectrométries de masse en tandem*. *Spectra Analyse*, 2015. **307**: p. 73-78.
59. Wolfender, J.L., et al., *Innovative omics-based approaches for prioritisation and targeted isolation of natural products - new strategies for drug discovery*. *Nat. Prod. Rep.*, 2019. **36**(6): p. 855-868.
60. Moss, S., et al., *Efficient structure elucidation of natural products in the pharmaceutical industry*. *Chimia*, 2007. **61**(6): p. 346-349.
61. Klausmeyer, P., et al., *Separation and SAR study of HIF-1alpha inhibitory tubulosines from Alangium cf. longiflorum*. *Planta Med.*, 2008. **74**(3): p. 258-263.
62. Lin, S., et al., *Characterization of chlorinated valepotriates from Valeriana jatamansi*. *Phytochemistry*, 2013. **85**: p. 185-193.
63. Harinantenaina Rakotondraibe, L., et al., *Antiproliferative and antiplasmodial compounds from selected Streptomyces species*. *Bioorg. Med. Chem. Lett.*, 2015. **25**(23): p. 5646-5649.

64. Bobzin, S.C., S. Yang, and T.P. Kasten, *Application of liquid chromatography-nuclear magnetic resonance spectroscopy to the identification of natural products*. J. Chromatogr. B Biomed. Sci. Appl., 2000. **748**(1): p. 259-267.
65. Bobzin, S.C., S. Yang, and T.P. Kasten, *LC-NMR: a new tool to expedite the dereplication and identification of natural products*. J. Ind. Microbiol. Biotechnol., 2000. **25**(6): p. 342-345.
66. Wolfender, J.L., K. Ndjoko, and K. Hostettmann, *The potential of LC-NMR in phytochemical analysis*. Phytochem. Anal., 2001. **12**(1): p. 2-22.
67. Waridel, P., et al., *Identification of the polar constituents of Potamogeton species by HPLC-UV with post-column derivatization, HPLC-MSn and HPLC-NMR, and isolation of a new ent-labdane diglycoside*. Phytochemistry, 2004. **65**(16): p. 2401-2410.
68. Harrigan, G.G. and G.H. Goetz, *Chemical and biological integrity in natural products screening*. Comb. Chem. High Throughput Screen., 2005. **8**(6): p. 529-534.
69. Jaroszewski, J.W., *Hyphenated NMR methods in natural products research, part 1: direct hyphenation*. Planta Med., 2005. **71**(8): p. 691-700.
70. Wolfender, J.L., E.F. Queiroz, and K. Hostettmann, *The importance of hyphenated techniques in the discovery of new lead compounds from nature*. Expert Opin. Drug. Discov., 2006. **1**(3): p. 237-260.
71. Dias, D.A. and S. Urban, *Application of HPLC-NMR for the rapid chemical profiling of a Southern Australian sponge, Dactylospongia sp.* J. Sep. Sci., 2009. **32**(4): p. 542-548.
72. Queiroz, E.F., J.L. Wolfender, and K. Hostettmann, *Modern approaches in the search for new lead antiparasitic compounds from higher plants*. Curr. Drug. Targets, 2009. **10**(3): p. 202-211.
73. Kreiss, W., et al., *Chromatography-bioluminescence coupling reveals surprising bioactivity of inthomycin A*. Anal. Bioanal. Chem., 2010. **398**(5): p. 2081-2088.
74. Brkljača, R. and S. Urban, *Recent advancements in HPLC-NMR and applications for natural product profiling and identification*. J. Liq. Chromatogr. R. T., 2011. **34**(13): p. 1063-1076.
75. Sarker, S.D. and L. Nahar, *An introduction to natural products isolation*. Methods Mol. Biol., 2012. **864**: p. 1-25.
76. Sarker, S.D. and L. Nahar, *Hyphenated techniques and their applications in natural products analysis*. Methods Mol. Biol., 2012. **864**: p. 301-340.
77. Yim, S.-H., et al., *Structure-guided identification of novel phenolic and phenolic amide allosides from the rhizomes of Cimicifuga heracleifolia*. B. Kor. Chem. Soc., 2012. **33**(4): p. 1253-1258.
78. Pawlus, A.D., et al., *Chemical dereplication of wine stilbenoids using high performance liquid chromatography-nuclear magnetic resonance spectroscopy*. J. Chromatogr. A, 2013. **1289**: p. 19-26.
79. Clarkson, C., et al., *Hyphenation of solid-phase extraction with liquid chromatography and nuclear magnetic resonance: application of HPLC-DAD-SPE-NMR to identification of constituents of Kanahia laniflora*. Anal. Chem., 2005. **77**(11): p. 3547-3553.
80. Jaroszewski, J.W., *Hyphenated NMR methods in natural products research, Part 2: HPLC-SPE-NMR and other new trends in NMR hyphenation*. Planta Med., 2005. **71**(9): p. 795-802.
81. Lambert, M., et al., *Rapid extract dereplication using HPLC-SPE-NMR: analysis of isoflavonoids from Smirnowia iranica*. J. Nat. Prod., 2005. **68**(10): p. 1500-1509.
82. Wolfender, J.L., E.F. Queiroz, and K. Hostettmann, *Phytochemistry in the microgram domain - a LC-NMR perspective*. Magn. Reson. Chem. : MRC, 2005. **43**(9): p. 697-709.
83. Tatsis, E.C., et al., *Identification of the major constituents of Hypericum perforatum by LC/SPE/NMR and/or LC/MS*. Phytochemistry, 2007. **68**(3): p. 383-393.

84. Motti, C.A., et al., *FTICR-MS and LC-UVMS-SPE-NMR applications for the rapid dereplication of a crude extract from the sponge Ianthella flabelliformis*. J. Nat. Prod., 2009. **72**(2): p. 290-294.
85. Pedersen, M.M., et al., *Antimalarial sesquiterpene lactones from Distephanus angulifolius*. Phytochemistry, 2009. **70**(5): p. 601-607.
86. Staerk, D., et al., *Accelerated dereplication of crude extracts using HPLC-PDA-MS-SPE-NMR: quinolinone alkaloids of Haplophyllum acutifolium*. Phytochemistry, 2009. **70**(8): p. 1055-1061.
87. Johansen, K.T., et al., *From retrospective assessment to prospective decisions in natural product isolation: HPLC-SPE-NMR analysis of Carthamus oxyacantha*. J. Nat. Prod., 2011. **74**(11): p. 2454-2461.
88. Pieri, V., et al., *¹H NMR-based metabolic profiling and target analysis: a combined approach for the quality control of Thymus vulgaris*. Metabolomics, 2011. **8**(2): p. 335-346.
89. Johansen, K.T., S.G. Wubshet, and N.T. Nyberg, *HPLC-NMR revisited: using time-slice high-performance liquid chromatography-solid-phase extraction-nuclear magnetic resonance with database-assisted dereplication*. Anal. Chem., 2013. **85**(6): p. 3183-3189.
90. van der Hoof, J.J.J., et al., *Structural elucidation of low abundant metabolites in complex sample matrices*. Metabolomics, 2013. **9**(5): p. 1009-1018.
91. Halabalaki, M., et al., *Recent advances and new strategies in the NMR-based identification of natural products*. Curr. Opin. Biotechnol., 2014. **25**: p. 1-7.
92. Corcoran, O., *Hit discovery from natural products in pharmaceutical R&D*. Emagres, 2015. **4**(2): p. 455-461.
93. de Medeiros, L.S., et al., *Dichlorinated and brominated rugulovasines, ergot alkaloids produced by Talaromyces wortmannii*. Molecules, 2015. **20**(9): p. 17627-17644.
94. Williams, R.B., et al., *Isolation and identification of the novel tubulin polymerization inhibitor bifidenone*. J. Nat. Prod., 2017. **80**(3): p. 616-624.
95. Zani, C.L. and A.R. Carroll, *Database for rapid dereplication of known natural products using data from MS and fast NMR experiments*. J. Nat. Prod., 2017. **80**(6): p. 1758-1766.
96. Bitzer, J., et al., *Accelerated dereplication of natural products, supported by reference libraries*. Chimia, 2007. **61**(6): p. 332-338.
97. Dictionary of Natural Products. Available from: dnp.chemnetbase.com/.
98. AntiBase. Available from: <https://application.wiley-vch.de/stmdata/antibase.php>.
99. Koehn, F.E., *High impact technologies for natural products screening*. Progress in Drug Research, 2008. **65**: p. 175, 177-210.
100. Bakiri, A., et al., *Computer-aided dereplication and structure elucidation of natural products at the University of Reims*. Mol. Inform., 2017. **36**(10).
101. Ottavioli, J., et al., *Identification and quantitative determination of resin acids from Corsican Pinus pinaster aitton oleoresin using (¹³) C-NMR spectroscopy*. Chem. Biodivers., 2019. **16**(1): p. e1800482.
102. Emwas, A.H., *The strengths and weaknesses of NMR spectroscopy and mass spectrometry with particular focus on metabolomics research*. Methods Mol. Biol., 2015. **1277**: p. 161-193.
103. Schoenberger, T., Y.B. Monakhova, and D.W. Lachenmeier, *Guide to NMR method development and validation - part I: identification and quantification*. ResearchGate, 2014.
104. Bradshaw, J., et al., *A rapid and facile method for the dereplication of purified natural products*. J. Nat. Prod., 2001. **64**(12): p. 1541-1544.
105. Cheng, H.N. and M.A. Bennett, *Trends in shift rules in carbon-13 nuclear magnetic resonance spectroscopy and computer-aided shift prediction*. Anal. Chim. Acta, 1991. **242**: p. 43-56.

106. Stortz, C.A. and A.S. Cerezo, *The ^{13}C NMR spectroscopy of carrageenans: calculation of chemical shifts and computer-aided structural determination*. Carbohydr. Polym., 1992. **18**(4): p. 237-242.
107. Calcul, L., et al., *NMR strategy for unraveling structures of bioactive sponge-derived oxy-polyhalogenated diphenyl ethers*. J. Nat. Prod., 2009. **72**(3): p. 443-449.
108. ACD/NMR Predictors. Available from: https://www.acdlabs.com/products/adh/nmr/nmr_pred/.
109. Metrelab NMR Predict. Available from: <https://mestrelab.com/software/mnova/nmr-predict/>.
110. ChemDraw Professional. Available from: <https://www.perkinelmer.com/product/chemdraw-professional-chemdrawpro>.
111. Kuhn, S. and S.R. Johnson, *Stereo-aware extension of HOSE codes*. ACS Omega, 2019. **4**(4): p. 7323-7329.
112. Bremser, W., *Hose — a novel substructure code*. Anal. Chim. Acta, 1978. **103**(4): p. 355-365.
113. Bruguiera, A., et al., *(^{13}C)-NMR dereplication of Garcinia extracts: Predicted chemical shifts as reliable databases*. Fitoterapia, 2018. **131**: p. 59-64.
114. Scandolera, A., et al., *GABA and GABA-alanine from the red microalgae Rhodorus marinus exhibit a significant neuro-soothing activity through inhibition of neuro-inflammation mediators and positive regulation of TRPV1-related skin sensitization*. Mar. Drugs, 2018. **16**(3).
115. Lehbili, M., et al., *Two new bis-iridoids isolated from Scabiosa stellata and their antibacterial, antioxidant, anti-tyrosinase and cytotoxic activities*. Fitoterapia, 2018. **125**: p. 41-48.
116. Nivelle, L., et al., *Anti-cancer activity of resveratrol and derivatives produced by grapevine cell suspensions in a 14 L stirred bioreactor*. Molecules, 2017. **22**(3).
117. Hubert, J., et al., *Exploiting the complementarity between dereplication and computer-assisted structure elucidation for the chemical profiling of natural cosmetic ingredients: Tephrosia purpurea as a case study*. J. Nat. Prod., 2015. **78**(7): p. 1609-1617.
118. Abedini, A., et al., *Bioactivity-guided identification of antimicrobial metabolites in Alnus glutinosa bark and optimization of oregonin purification by Centrifugal Partition Chromatography*. J. Chromatogr. B Analyt. Technol. Biomed. Life Sci., 2016. **1029-1030**: p. 121-127.
119. Tisserant, L.P., et al., *(^{13}C) NMR and LC-MS profiling of stilbenes from elicited grapevine hairy root cultures*. J. Nat. Prod., 2016. **79**(11): p. 2846-2855.
120. Sientzoff, P., et al., *Fast identification of radical scavengers from Securigera varia by combining ^{13}C -NMR-based dereplication to bioactivity-guided fractionation*. Molecules, 2015. **20**(8): p. 14970-14984.
121. Hubert, J., et al., *Identification of natural metabolites in mixture: a pattern recognition strategy based on (^{13}C) NMR*. Anal. Chem., 2014. **86**(6): p. 2955-2962.
122. Ferreira, M.J.P., et al., *Computer-aided method for identification of components in essential oils by ^{13}C NMR spectroscopy*. Anal. Chim. Acta, 2001. **447**(1-2): p. 125-134.
123. Pilon, A.C., et al., *NuBBEDB: an updated database to uncover chemical and biological information from Brazilian biodiversity*. Sci. Rep., 2017. **7**(1): p. 7215.
124. CH-NMR-NP. Available from: <https://www.j-resonance.com/en/nmrdb/>.
125. MarinLit. Available from: <http://pubs.rsc.org/marinlit/>.
126. NMRShiftDB2. Available from: <https://nmrshiftdb.nmr.uni-koeln.de/>.
127. v. d. Lieth, C.W., et al., *^{13}C NMR data bank techniques as analytical tools*. Magn. Reson. Chem., 1985. **23**(12): p. 1048-1055.
128. Blunt, J.W. and M.H.G. Munro, *Data, ^1H -NMR databases, data manipulation, ...*. Phytochem. Rev., 2012. **12**(3): p. 435-447.
129. Mihaleva, V.V., et al., *MetIDB: a publicly accessible database of predicted and experimental ^1H NMR spectra of flavonoids*. Anal. Chem., 2013. **85**(18): p. 8700-8707.

130. *The Human Metabolome Database (HMDB)*. Available from: <http://www.hmdb.ca/>.
131. Davies, S., et al., *The dynamic range problem in NMR*. J. Magn. Reson. (1969), 1985. **64**(1): p. 155-159.
132. Pan, H., et al., *Mass defect filtering-oriented classification and precursor ions list-triggered high-resolution mass spectrometry analysis for the discovery of indole alkaloids from Uncaria sinensis*. J. Chromatogr. A, 2017. **1516**: p. 102-113.
133. Dunkel, R. and X. Wu, *Identification of organic molecules from a structure database using proton and carbon NMR analysis results*. J. Magn. Reson., 2007. **188**(1): p. 97-110.
134. Bakiri, A., et al., *Reconstruction of HMBC correlation networks: a novel NMR-based contribution to metabolite mixture analysis*. J. Chem. Inf. Model., 2018.
135. Bakiri, A., et al., *Computer-aided (^{13}C) NMR chemical profiling of crude natural extracts without fractionation*. J. Nat. Prod., 2017. **80**(5): p. 1387-1396.
136. Sumner, L.W., et al., *Proposed minimum reporting standards for chemical analysis Chemical Analysis Working Group (CAWG) Metabolomics Standards Initiative (MSI)*. Metabolomics, 2007. **3**(3): p. 211-221.
137. Bighelli, A., F. Tomi, and J. Casanova, *Computer-aided carbon ^{13}C NMR study of phenols contained in liquids produced by pyrolysis of biomass*. Biomass Bioenerg., 1994. **6**(6): p. 461-466.
138. Kubeczka, K.-H., et al., *The composition of the essential oils of Chaerophyllum hirsutum L.* J. Essent. Oil. Res., 1989. **1**(6): p. 249-259.
139. Esselin, H., et al., *Snyderol derivatives from Laurencia obtusa collected in Corsica*. Biochem. Syst. Ecol., 2019. **82**: p. 24-26.
140. Esselin, H., et al., *New metabolites isolated from a Laurencia obtusa population collected in Corsica*. Molecules, 2018. **23**(4).
141. Duquesnoy, E., et al., *Identification of taxanes in extracts from leaves of Taxus baccata L. using (^{13}C)-NMR spectroscopy*. Phytochem. Anal., 2009. **20**(3): p. 246-252.
142. Caraux, G. and S. Pinloche, *PermutMatrix: a graphical environment to arrange gene expression profiles in optimal linear order*. Bioinformatics, 2005. **21**(7): p. 1280-1281.
143. Bruguière, A., et al., *MixONat software for mixture dereplication base on ^{13}C NMR and DEPT experiments*. Anal. Chem. (Submitted).
144. Blondel, V.D., et al., *Fast unfolding of communities in large networks*. J. Stat. Mech.-Theory E., 2008. **2008**(10).
145. Schrimpe-Rutledge, A.C., et al., *Untargeted metabolomics strategies-challenges and emerging directions*. J. Am. Soc. Mass. Spectr., 2016. **27**(12): p. 1897-1905.

II. Travaux personnels

1. Construction des bases de données

1.1. L'étape clé des références

Comme la précédente revue nous a permis de le constater, une des étapes-clé du processus de déréplication est la construction d'un set de molécules de références, souvent compilées en bases de données, utilisé pour l'étape de comparaison entre les signaux de l'échantillon et ceux des références. C'est donc une des premières problématiques à laquelle nous avons été confrontés pour la construction de bases adaptées à notre objectif de déréplication par RMN ^{13}C . En effet, on retrouve dans la littérature des façons très différentes de procéder à cette étape, tant au niveau qualitatif (types de molécules) que quantitatif (nombre d'entrées) des bases (cf. **1. Article 1 : A highlight on ^{13}C -NMR based dereplication methods**).

Les molécules servant de référence peuvent être sélectionnées en se basant sur un ou plusieurs critères chimiotaxonomiques (famille, genre ou espèce d'où est extrait l'échantillon), structuraux (type de composés supposés présents dans l'échantillon) ou bien chercher à regrouper le maximum de molécules possible. En fonction de la façon dont une base est construite, cela peut conduire à des bases contenant une cinquantaine de molécules [6], tout comme des dizaines de milliers [7], ce qui impacte bien évidemment les résultats de la déréplication. Une base de données trop large peut en effet « noyer » l'information, en proposant énormément de molécules « parasites » à comparer. En revanche, une base de données trop restreinte risque de ne pas contenir les molécules d'intérêt, ou leurs analogues.

L'autre différence importante concernant les bases de données utilisées dans la littérature est le choix d'utiliser des données expérimentales [8], des données prédites [6], ou un mélange des deux [9]. D'un premier abord, il semble plus intéressant de construire des bases de données à partir de données expérimentales, permettant ainsi de jouir d'une précision plus importante lors de la comparaison des data. Cependant, deux inconvénients majeurs sont à pointer. D'abord, la construction de ce genre de base représente une quantité de travail conséquente, impliquant d'associer manuellement tous les δ_c aux structures moléculaires, dans le cas de bases à visée déréplicative par RMN- ^{13}C . L'autre problème rencontré est que ce long travail peut ne pas porter ses fruits dans le cas où les données récoltées dans la littérature ont été enregistrées dans un solvant deutéré différent de celui utilisé pour l'échantillon d'intérêt. Pour la plupart des molécules, et notamment les PPAPs qui nous intéressent, il serait impossible de construire une base expérimentale contenant uniquement des déplacements chimiques des molécules enregistrées dans le même solvant deutéré. En effet les données sont indifféremment reportées dans le DMSO, le méthanol (\pm acide trifluoroacétique) et dans la pyridine deutérés [10, 11].

L'utilisation de logiciels permettant, à partir des structures des molécules d'intérêt, la prédiction des déplacements chimiques de chaque atome est également possible (ACD/Labs® [12], MestReNova® [13] et ChemDraw® [14]). Il ne s'agit en fait pas d'une prédiction *de novo* mais plus d'une estimation par similarité(s) observée(s). Les logiciels, pour la plupart commerciaux, se basent en effet sur les codes HOSE (*Hierachically Ordered Spherical*

description of Environment) des atomes. Le code HOSE d'un atome est défini par les autres atomes présents dans son environnement, dans une sphère d'un diamètre plus ou moins large (**Figure 1**) [15]. L'environnement immédiat de l'atome, situé à une liaison autour de celui-ci, représente le niveau 1 du code HOSE. Une sphère de 2 liaisons de diamètre représentant le niveau 2, ainsi de suite. Un exemple de HOSE *coding* est présenté en **Figure 2**. Les logiciels intègrent leur propre base de données de données expérimentales (*i.e.* δ_c). Pour une molécule donnée, l'algorithme recherche dans cette base de données, atome par atome, un code HOSE identique de plus grand niveau possible (c'est-à-dire possédant le niveau sphérique le plus large possible). Si aucun résultat n'est trouvé avec un niveau élevé la recherche recommencera avec un diamètre de sphère réduit. Si, par exemple, aucun atome ne partage le même code HOSE de niveau 6, le programme cherchera un atome possédant un code HOSE de niveau 5. Lorsque des résultats sont trouvés, le déplacement chimique de l'atome est prédit par un calcul, qui peut être une simple moyenne des déplacements expérimentaux trouvés, mais qui n'est pas explicité sur la plupart des logiciels (effet « boîte noire ») [12-14]. La qualité des prédictions repose donc également sur la « richesse » (nombre de structures différentes et nombre d'environnement similaires) de la base de données.

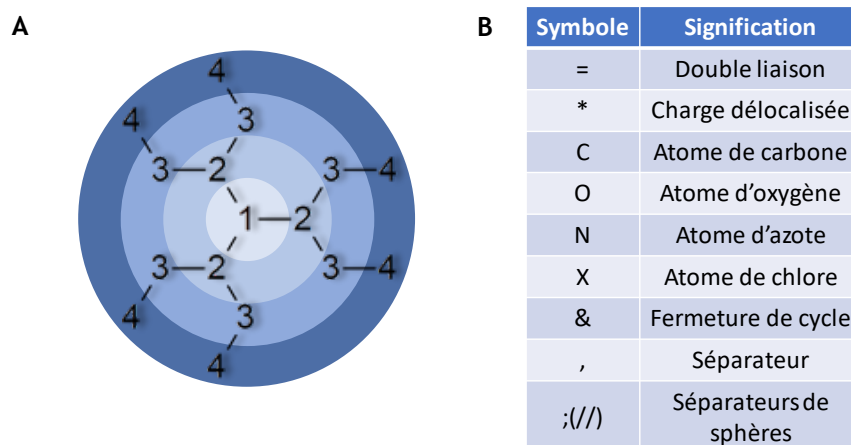
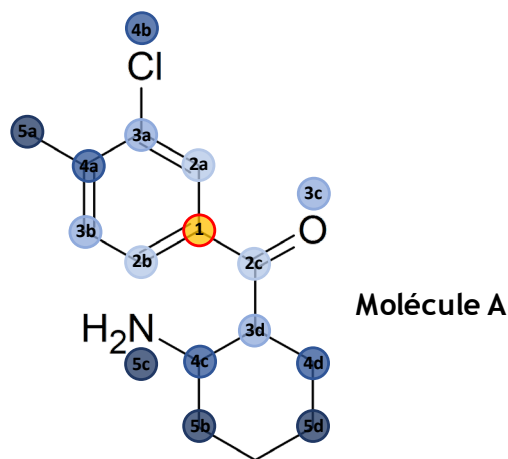


Figure 1: (A) Principe général du code HOSE et (B) extrait du codex HOSE.



Niveau de sphère	Atome précédent	Numéro d'atome	Code HOSE	Code HOSE de la sphère	Fragment correspondant
I	Aucun	1	C	*C	C
II	1	2a	*C	*C*CC(
		2b	*C		
		2c	C		
III	2a	3a	*C	*C,*C,=OC/	
	2b	3b	*C		
	2c	3c	=O		
		3d	C		
IV	3a	4a	*C	*CX,*&,,CC/	
		4b	X		
	3b	4a	*&		
	3c	Aucun			
	3d	4c	C		
		4d	C		
V	4a	3b	*&	*&C,,CN,C)	
		5a	C		
	4b	Aucun			
	4c	5b	C		
		5c	N		
	4d	5d	C		

Code HOSE pour carbone 1 = *C;*C*CC(*C,*C,=OC/*CX,*&,,CC/*&C,,CN,C)

Figure 2: HOSE code détaillé du C-1 de la molécule A

Cette méthode représenterait ainsi une alternative intéressante à la construction de bases de données expérimentales car, une fois les structures des molécules d'intérêt collectées, elle permet d'obtenir des bases de données de grande taille de façon très rapide, notamment grâce à la possibilité d'automatiser certains processus grâce à des macros. Cependant, il est important de savoir si la précision de ces prédictions est suffisante pour

atteindre l'objectif de déréplication que l'on s'est fixé. Notre première étude avait donc pour but d'analyser la qualité des données obtenues *via* 3 logiciels prédictifs, afin de déterminer si cette méthode pouvait être utilisée, sans altérer la qualité des résultats finaux, dans notre processus d'analyse.

1.2. Article 2: ^{13}C -NMR dereplication of Garcinia extracts: Predicted chemical shifts as reliable databases

1.2.1. Résumé de l'article 2

Cet article [16] reprend les principaux résultats issus de cette étude, *i.e.* dans un premier temps, l'évaluation de la précision de 3 logiciels de prédictions des δ_c (ACD/Labs® [12], MestReNova® [13] et ChemDraw® [14]), et dans un second temps, la comparaison de l'utilisation de banques de données expérimentales et prédictives sur un exemple d'analyse déréplicative d'un extrait de *Garcinia bancana* malaisien (Clusiaceae).

Pour ce faire, 80 structures de produits naturels ont été recueillies de façon à obtenir des molécules de classes chimiques différentes, représentatives de la famille des Clusiaceae : benzophénones, biphényles, chromanones, coumarines, depsidones, PPAPs, tocotriénols, triterpènes et xanthones. D'une part, les déplacements expérimentaux de ces molécules ont été recueillis à partir de diverses publications; d'autre part, les déplacements chimiques de ces mêmes molécules ont été prédits par les 3 différents logiciels à notre disposition. Pour chaque logiciel, la différence en ppm entre valeur prédite et expérimentale a été calculée et répertoriée dans un tableur. Grâce à la fonction de tableau croisé dynamique proposé par Microsoft Excel®, il a en effet été possible de visualiser l'importance de cette différence de précision en fonction de la classe chimique des molécules ou du type de carbone considéré (carbone quaternaire, tertiaire, secondaire primaire, carbone lié à des fonctions oxydée, etc...).

L'exemple précis de déréplication a ensuite consisté à appliquer la méthode décrite par de Hubert *et al.* [9] à un extrait dichlorométhanique de *G. bancana*. Après fractionnement, analyse ^{13}C -NMR, *binning*¹ des déplacements chimiques et HCA, les déplacements chimiques présents au sein de chaque cluster ont été soumis à une recherche dans deux banques de données. La première est une banque de données prédictives, construite à partir des structures de 718 produits naturels reportés dans le genre *Garcinia* (disponible en *supporting information* de 1.2.2. Article 2). La seconde, est la base de données CH-NMR-NP® de la société JEOL [17], librement disponible en ligne (<https://www.j-resonance.com/en/nmrdb/>), regroupant les données des déplacements chimiques expérimentaux de 30 500 produits naturels, reportés dans la littérature entre 2000 et 2014. Les hypothèses proposées par ces deux types de banques de données ont ensuite été comparées à la réelle composition de l'extrait initial, celle-ci ayant parallèlement fait l'objet d'une étude phytochimique menée par HPLC semi-préparative et analyse RMN des molécules purifiées.

¹ Le *binning* permet de découper l'axe des ordonnées du spectre RMN (ppm) en plusieurs intervalles égaux appelés *bins* (généralement 0,2 ppm). Dans chacun des *bins*, seuls le signal possédant la plus forte intensité sera conservé. Cela permet de simplifier le jeu de données.

1.2.2. Article 2



¹³C-NMR dereplication of *Garcinia* extracts: Predicted chemical shifts as reliable databases



Antoine Bruguère^a, Séverine Derbré^{a,*}, Chloé Coste^a, Maxime Le Bot^a, Benjamin Siegler^b, Sow Tein Leong^c, Syazreen Nadia Sulaiman^c, Khalijah Awang^c, Pascal Richomme^a

^a SONAS SFR QUASAV, University of Angers, France

^b SFR MATRIX, University of Angers, France

^c Department of Chemistry, Faculty of sciences, University of Malaya, Malaysia

ARTICLE INFO

Keywords:

Clusiaceae

Database

¹³C-NMR Dereplication

Garcinia

PPAPs

Prediction software

ABSTRACT

Usually isolated from *Garcinia* (Clusiaceae) or *Hypericum* (Hypericaceae) species, some Polycyclic Polyprenylated Acylphloroglucinols (PPAPs) have been recently reported as potential research tools for immunotherapy. Aiming at exploring the chemodiversity of PPAPs amongst *Garcinia* genus, a dereplication process suitable for such natural compounds has been developed. Although less sensitive than mass spectrometry, NMR spectroscopy is perfectly reproducible and allows stereoisomers distinction, justifying the development of ¹³C-NMR strategies. Dereplication requires the use of databases (DBs). To define if predicted DBs were accurate enough as dereplication tools, experimental and predicted δ_c of natural products usually isolated from Clusiaceae were compared. The ACD/Labs commercial software allowed to predict 73% of δ_c in a 1.25 ppm range around the experimental values. Consequently, with these parameters, the major PPAPs from a *Garcinia bancana* extract were successfully identified using a predicted DB.

1. Introduction

Manipulating the immune system is a therapeutic approach to either activate immunity against tumors and infected cells or to prevent its activation in autoimmune diseases or transplantation [1,2]. The Major Histocompatibility Complex (MHC) is a group of genes encoding cell surface proteins that control immune responses in both normal and pathological conditions. In this context, natural products (NPs) able to modulate MHC protein expression may provide new therapeutic strategies for immunotherapy. The authors have recently shown that Polycyclic Polyprenylated Acylphloroglucinols (PPAPs) such as guttiferone J (1) (Fig. 1) represent efficient modulators of some MHC proteins [3]. PPAPs share one acylphloroglucinol scaffold, polysubstituted by prenyl or geranyl groups with various degrees of oxidation involved in different types of secondary cyclizations [4]. Guttiferone J (1), like many PPAPs, was previously isolated from *Garcinia* species (i.e. *G. virgata* and *G. yunnanensis*) [4–6] which also biosynthesize a variety of different NPs including xanthenes, triterpenes, biphenyls, chromanones, coumarins, depsidones and tocotrienols [7]. Exploring the chemodiversity of PPAPs may reveal further valuable leads. However, very few of these PPAPs are commercially available, whereas a wide variety of such derivatives may be isolated from Clusiaceae, Calophyllaceae

and Hypericaceae species with 618 entities reported so far [4,8].

In order to limit time consuming fractionation and purification procedures and rapidly focus on crude extracts containing desired molecular scaffolds, the authors have developed a targeted dereplication protocol specifically adapted for the detection of PPAPs. High-performance liquid chromatography (HPLC) combined with mass spectrometry (MS) is commonly used as method of first choice as it allows to compare retention times, molecular weights and fragmentation patterns with references data [9]. Moreover, these data may be used for building up molecular networks [10]. However, as far as PPAPs are concerned, asymmetric centers and prenylated/geranylated groups result in fragmentation patterns that are sometimes hardly distinguishable by MSⁿ [11]. Alternatively, dereplication may also be achieved through nuclear magnetic resonance (NMR): ¹H-NMR [12], ¹³C-NMR [13] or 2D-NMR [14]. As dereplication tools, both MS and NMR exhibit advantages and drawbacks. While MS provides a higher sensitivity, a standard ionization protocol may not be suitable for a wide range of different molecules often present in plant materials. Though direct inspection of polyphenol extracts has proved to be efficient in some specific cases [15,16], hyphenation with chromatography is generally requested prior to MS analysis. A contrario, NMR can be applied to complex mixtures [17]. Moreover, it allows differentiation of

* Corresponding author.

E-mail address: severine.derbre@univ-angers.fr (S. Derbré).

<https://doi.org/10.1016/j.fitote.2018.10.003>

Received 22 June 2018; Received in revised form 25 September 2018; Accepted 1 October 2018

Available online 12 October 2018

0367-326X/ © 2018 Elsevier B.V. All rights reserved.

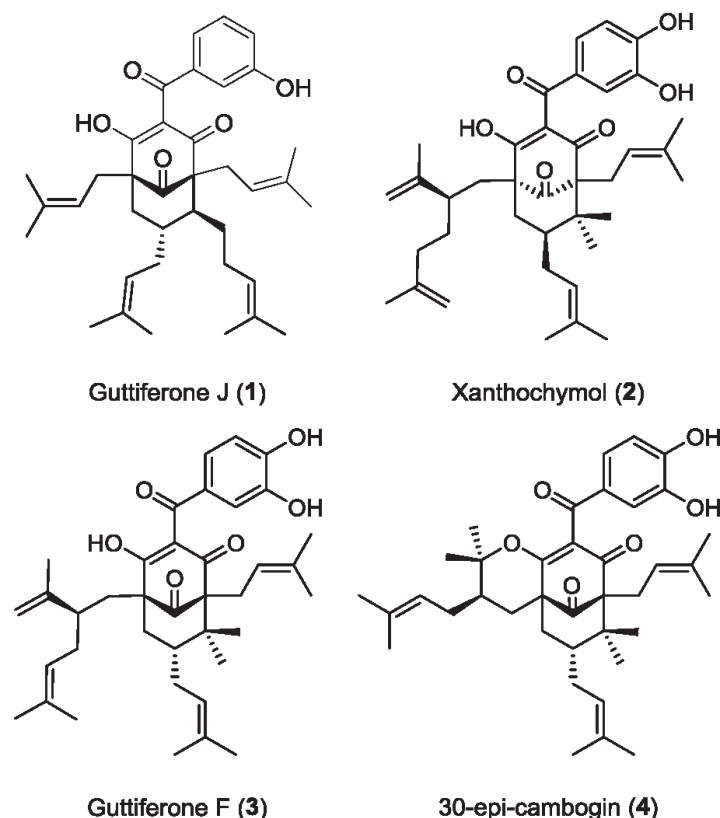


Fig. 1. Structures of guttiferone J (1) and of the compounds isolated from *Garcinia bancana* bark: xanthochymol (2), guttiferone F (3) and 30-epi-cambogin (4).

stereoisomers, which may be very challenging or even impossible with MS analysis [18]. On the one hand, ^1H -NMR spectra are quickly recorded, but δ_{H} are strongly solvent-dependent whereas complex mixtures also generate regions of intense δ_{H} overlapping which impair spectra interpretation. On the other hand, ^{13}C -NMR requires longer acquisition time, but δ_{C} are less solvent-sensitive whilst signals overlapping are seldom observed. 2D-NMR may provide additional key information on the spatial structure of the molecule, but for practical reasons these experiments must be recorded with much lower resolution than for 1D-NMR [19]. Therefore, benefitting from the assets of ^{13}C -NMR a dereplication strategy was thus applied in order to quickly identify PPAPs in a *Garcinia bancana* (Clusiaceae) bark extract [20]. In this context, a δ_{C} database (DB) was required.

The literature suggests that several types of DBs can be used. First, the use of an internal DB gathering the experimental chemical shifts of secondary metabolites isolated and analyzed by a given laboratory is possible. In such a case, it is possible to identify compounds with an accuracy of 0.1 ppm if all analyses are conducted in the same deuterated solvent and with the same parameters [17]. Then, working on more complex mixtures, it is usually impossible to use the same NMR conditions, or to have an extensive internal DB at our disposal. Therefore, ^{13}C chemical shifts can be gathered from the literature to build a bigger “experimental DB” [21]. However, this type of DB is quite long to build and obtaining an exhaustive one is a difficult task. For the present work, this is even more difficult with PPAPs for which NMR data are usually reported in pyridine- d_6 , deuterated chloroform and methanol- d_4 with or without 0.1% TFA, generating strongly solvent

dependant DBs. The third and last option is the construction of a theoretical DB using a ^{13}C -NMR prediction software. To our knowledge, reported ^{13}C -NMR dereplication processes indifferently rely on experimental or predicted DBs but, without real justification [22]. Hence, this work will try to figure out if predictive DBs are accurate enough to be used as a general dereplication tool suitable for *Garcinia* extracts.

2. Experimental

2.1. Predictive and experimental databases

As a general validation trial, 80 NPs representative of the Clusiaceae family, chosen for their structural diversity and including benzophenones, biphenyls, chromanones, coumarins, depsidones, PPAPs, tocotrienols, triterpenes and xanthenes (Table S1), were included in both an experimental DB1 and a predictive DB2. For the experimental one, structure-data files (SDF) of these metabolites were built using ChemSketch application (ACD/Labs Release 12.00) whilst literature chemical shifts were entered into a Microsoft Excel 2016 spreadsheet. For the predictive DB, SDF were analyzed with three different software programs: *Spectrum Processor* from ACD/Labs (2014 release), the module NMRP from MestReNova (v12.0.1) and ChemDraw Professional (v17.0.0.206). The absolute difference between experimental and predicted ^{13}C -NMR chemical shifts was calculated for each NP. The PivotTable feature allowed the evaluation of the distance between predicted and experimental values, ranked by accuracy, depending on the type of structure or carbon type, making the interpretation easier.

Table 1
Percentage of carbon signals predicted in a given interval using ACD/Labs prediction (blue), MNova (Orange) or ChemDraw (green) software. For each prediction software and each structural class, the percentage of carbon signals that have their predicted chemical shifts in a given interval (in ppm) around the experimental chemical shift value is shown. For example, concerning the ACD/Labs prediction software, in average, 73% of the predicted carbon signals are less than 1.25 ppm away from the value described in the literature.

		ACD/Labs										MNova										ChemDraw									
		Structural class										Structural class										Structural class									
		Benzophenone		Biphenyl		Chromanone		Coumatin		Depsidone		Benzophenone		Biphenyl		Chromanone		Coumatin		Depsidone		Benzophenone		Biphenyl		Chromanone		Coumatin		Depsidone	
		Absolute difference	Average	Absolute difference	Average	Absolute difference	Average	Absolute difference	Average	Absolute difference	Average	Absolute difference	Average	Absolute difference	Average	Absolute difference	Average	Absolute difference	Average	Absolute difference	Average	Absolute difference	Average	Absolute difference	Average	Absolute difference	Average	Absolute difference	Average	Absolute difference	Average
		ppm < 1.00	41%	60%	66%	65%	68%	69%	73%	68%	68%	41%	60%	66%	65%	68%	69%	73%	68%	68%	68%	38%	48%	37%	48%	39%	46%	38%	48%	38%	48%
		ppm < 1.25	53%	67%	72%	74%	74%	75%	78%	72%	72%	53%	67%	72%	74%	74%	75%	78%	72%	72%	72%	46%	56%	50%	56%	46%	53%	46%	56%	46%	56%
		ppm < 1.50	63%	77%	80%	80%	80%	80%	85%	78%	78%	63%	77%	80%	80%	80%	85%	80%	80%	80%	80%	56%	66%	57%	66%	55%	63%	55%	66%	55%	66%
		ppm < 1.75	69%	80%	85%	85%	85%	85%	90%	82%	82%	69%	80%	85%	85%	85%	90%	85%	85%	85%	85%	60%	70%	61%	70%	59%	68%	60%	70%	60%	70%
		ppm < 2.00	79%	88%	92%	92%	92%	92%	95%	88%	88%	79%	88%	92%	92%	92%	95%	92%	92%	92%	92%	69%	79%	69%	79%	69%	78%	69%	79%	69%	79%
		ppm < 3.00	86%	92%	95%	95%	95%	95%	98%	92%	92%	86%	92%	95%	95%	95%	98%	95%	95%	95%	95%	77%	87%	77%	87%	77%	86%	77%	87%	77%	87%
		ppm < 4.00	88%	96%	98%	98%	98%	98%	100%	95%	95%	88%	96%	98%	98%	98%	100%	98%	98%	98%	98%	88%	98%	88%	98%	88%	97%	88%	98%	88%	98%
		ppm < 5.00	98%	100%	100%	100%	100%	100%	100%	100%	100%	98%	100%	100%	100%	100%	100%	100%	100%	100%	100%	98%	100%	98%	100%	98%	100%	98%	100%	98%	100%
		ppm < 10.00	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
		ppm < 20.00	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
		20.00 ≤ ppm	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%

2.2. Method validation on a *Garcinia bancana* extract

2.2.1. Plant material

Garcinia bancana bark was collected in October 2000 around Mersing, Johor. The plant was identified by the botanist Mr. Teo Leong Eng and the voucher specimen (KL4967) was deposited at the Herbarium of the Department of Chemistry, University of Malaya, Kuala Lumpur, Malaysia.

2.2.2. Fractionation

10 g of bark powder were successively extracted by sonication (3 h) with dichloromethane (740 mg) and methanol (1.35 g). 650 mg of the dichloromethane extract were then fractionated using normal phase flash chromatography on silica gel (Chromabond® flash RS 40 SiOH) from 100% cyclohexane to 70% cyclohexane / 30% ethyl acetate (flow: 20 mL/min). 10 successive sub-fractions likely to contain PPAPs (LC-UV, ¹H NMR) were selected for ¹³C NMR dereplication: F1 (15.2 mg), F2 (12.2 mg), F3 (13.2 mg), F4 (18.8 mg), F5 (30.1 mg), F6 (6.5 mg), F7 (53.7 mg), F8 (27.6 mg), F9 (22.9 mg) and F10 (25.9 mg).

2.2.3. NMR experiments

¹³C-NMR experiments of the fractions and pure products were done in deuterated methanol and 0.1% trifluoroacetic acid using the JEOL 400 MHz YH spectrometer (Jeol Europe). Parameters were 16,000 scans for 15 mg of fraction.

2.2.4. Bucketing and cluster analysis

A bucketing step was automated using a macro on a Microsoft Excel spreadsheet. The hierarchical cluster analysis (HCA), allowing the formation of chemical shifts clusters, was processed using Permutmatrix 1.9.3 software [23]. HCA can recognize and sort specific set of ¹³C-NMR chemical shifts, each set is supposed to belong to a different molecule.

2.2.5. Compounds purification and identification

30 mg of fraction 7 were separated using semi-preparative HPLC (Agilent HP 1100 Series, Agilent Technologies, Les Ulis, France) on a reversed phase column (Phenomenex Luna C18, 100 Å, 250 × 10 mm, 5 µm), using a 50 mg/mL concentration for the injection (100 µL), with a 97% methanol + 0.1% formic acid / 3% water + 0.1% formic acid system (flow: 2.8 mL/min). Fractions were collected using the Agilent Technologies 1260 Infinity G1364C fraction collector and the ChemStation for LC 3D software for automatic UV peak detection (diode array detector G13115A). This led to 8.7 mg of xanthochymol (2) [24] and guttiferone F (3) [25] as a mixture.

Another semi-preparative HPLC on a different column (Hypersil Gold PFP, 150 mm × 10 mm), using a 50 mg/mL concentration for the injection (100 µL), with a 75% methanol / 25% water + 0.1% formic acid system (flow: 4.7 mL/min) yielded 2.5 mg of xanthochymol (2) and 1.5 mg of guttiferone F (3) from fraction 6. The same method led to the purification of 2.9 mg of xanthochymol (2), 3.6 mg of guttiferone F (3) and 4.3 mg of 30-epi-cambogin (4) [25] from fraction 7 (Fig. 1).

2.2.6. Database choice for *G. bancana*

A predictive DB3 in *C + H NMR Predictor* (ACD/Labs) was built using a SDF gathering 718 NPs described in the *Garcinia* genus on the Dictionary of Natural Products website [26]. On the other hand, the CH-NMR-NP JEOL database [27] was used as the experimental DB4. This website gathers the ¹³C-NMR chemical shifts of around 30,500 NPs published between 2000 and 2014, including compounds from *Garcinia*.

3. Results and discussion

3.1. Accuracy of the chemical shift prediction (DB1 vs DB2)

In average, as far as 80 NPs isolated from Clusiaceae and

Calophyllaceae species were concerned, the ACD/Labs software could predict 73% of carbon chemical shifts with a ± 1.25 ppm accuracy versus 2.00 and 3.00 ppm for MNova and ChemDraw software programs respectively (Table 1). It also appeared that the accuracy of these prediction software strongly depended on NPs structural class but not on the type of carbon hybridization.

3.2. Dereplication analysis on *Garcinia bancana* extract

Garcinia genus is known for producing various PPAPs of biological interest but also different kind of secondary metabolites [4]. To focus on PPAPs, a preliminary study was conducted on 124 DCM and MeOH extracts from the bark, leaf, and sometimes fruit of 17 Malaysian *Garcinia* species (30 batches): LC-UV-MS² analyses suggested PPAPs as major compounds in 19 DCM extracts including in a *G. bancana* bark extract. However, their structures could not be firmly characterized using MS² data due to the similar fragmentation pattern of the stereoisomers. Little work has been reported on this endemic Malaysian plant. Only two PPAPs (i.e. garcinol and isogarcinol), two isocoumarin derivatives, two triterpenes, a monoterpene glycoside and a biphenyl were isolated from the twigs when two flavones glycosides and tannins were reported in the leaves [28–30]. Thus, focusing on PPAPs, an NMR dereplication study was conducted on this DCM bark extract, requiring a database of NPs and their chemical shifts.

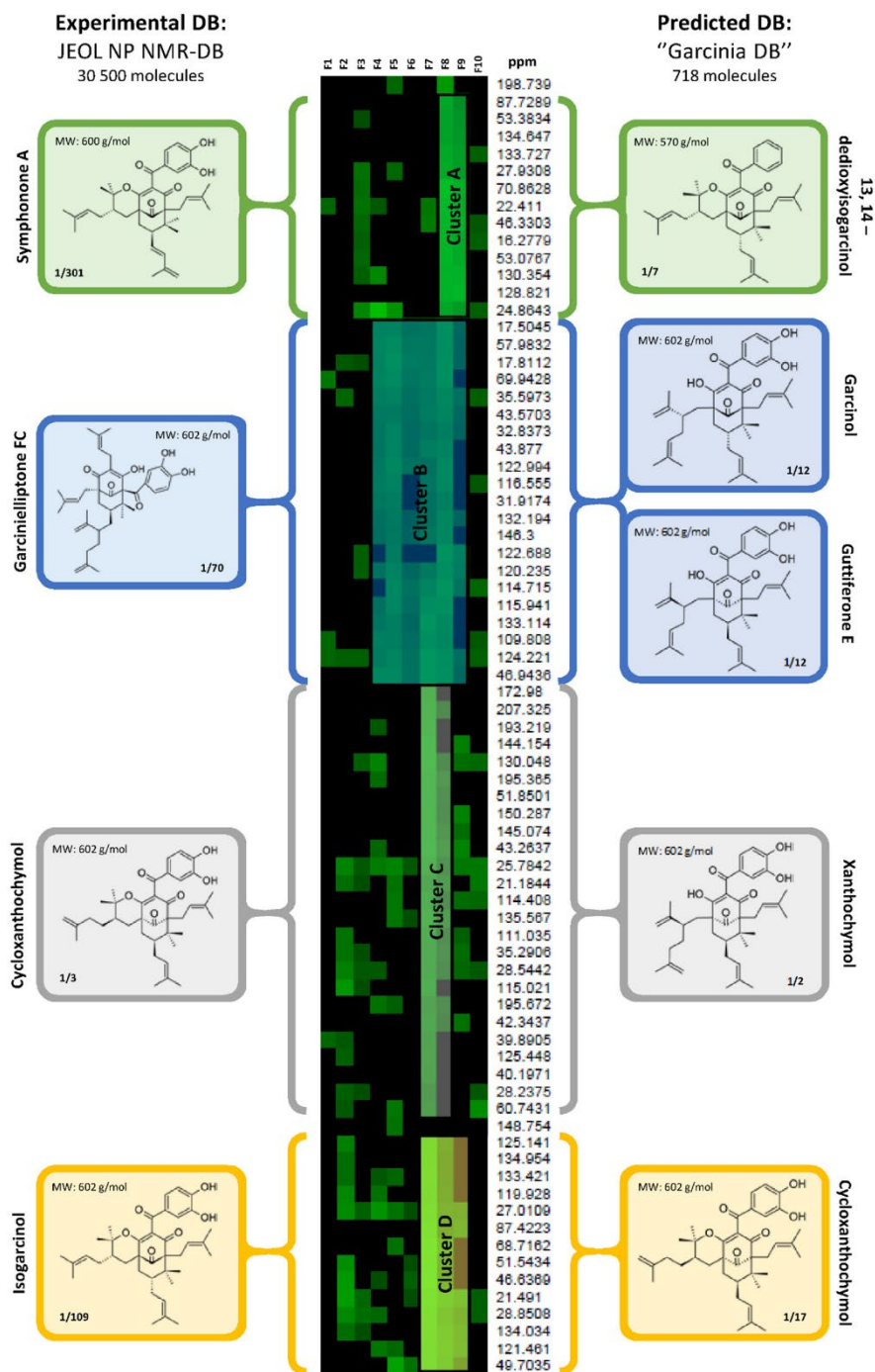
Exploring the potential of predicted DBs, we turned to a ¹³C-NMR dereplication using a fractionation step, as nicely described by Hubert et al. [20]. The hierarchical cluster analysis (HCA) step established 4 clusters corresponding to the major NPs across the several fractions. For each cluster, the set of chemical shifts were searched both in the predicted DB3 (*Garcinia*'s DNP) and DB4 (JEOL). The parameters defined in the previous experiment were used as query parameters (i.e. looseness factor 1.25 ppm; 75% of carbon signals matching). As shown in Fig. 2, both DBs managed to identify similar structural type of PPAPs but DB3 and DB4 often predicted different isomers. Thus, major NPs, namely xanthochymol (2), guttiferone F (3) and 30-epi-cambogin (4) (MSI level 1), were purified to determinate if the different DBs properly identified them. Predicted DB3 correctly suggested xanthochymol, stereoisomers of guttiferone F (garcinol or guttiferone E) and one isomer of 30-epi-cambogin (cyloxanthochymol). Concerning the experimental DB4, it proposed one isomer of guttiferone F (garcinielliptone FC) and one stereoisomer of 30-epi-cambogin (isogarcinol). Actually, as the experimental JEOL DB4 did not include the NPs of the crude extract, a perfect match was not possible.

4. Conclusion

To explore efficiently the chemodiversity of PPAPs and their ability to modulate MHC molecules expression, we have been developing a dereplication process suitable for *Garcinia* species. It should be noticed that LC-MS² analyses could not distinguish the different PPAPs stereoisomers. Therefore, a ¹³C-NMR dereplication study was chosen requiring an appropriate DB. When available, working with an experimental DB is ideal, increasing the chance for better matches if the NMR experiments are conducted in the same solvent. However, as no experimental DBs are exhaustive and require a long time to build, this work demonstrates for the first time that predictive DBs are an attractive alternative, at least for first dereplication steps. Indeed, the accuracy (± 1.25 ppm for 75% δ_C) of one predicted DB3 was sufficient to identify the major PPAPs successively isolated from a *Garcinia bancana* crude extract. As far as Clusiaceae and Calophyllaceae NPs are concerned, this work also shows great differences in prediction accuracies depending on software programs.

Conflict of interest

K. Awang, A Bruguère, C. Coste, S. Derbré, M. Le Bot, S. T. Leong, P.



(caption on next page)

Fig. 2. Molecules predicted for each cluster by the experimental DB (on the left) and the predicted DB (on the right). After hierarchical cluster analysis (HCA) of the fractions, this map was produced, showing the presence (green square) or absence (black square) of a given chemical shifts in one of the fractions. Clusters of green squares are recognized as chemical shifts belonging to the same molecule. Those chemical shifts were thus entered as query in both DBs. For each cluster and for each DB, the structure of the first proposed molecule is displayed. The predicted DB also shows its ranking. In cluster A for example, the software proposed 7 molecules as a potential match (out of the 718 in this DB), 13, 14-dedioxysogarcinol being the one with the higher score. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Richomme, B. Siegler and S. N. Sulaiman declare no conflict of interest.

Author contributions

This article is a part of the results obtained by AB during his doctoral thesis. STL and SNS, supervised by KA, extracted and fractionated *Garcinia bancana* bark extract. CC was involved in PPAPs purification. MLB and BS automated the bucketing step using a macro on a Microsoft Excel spreadsheet. This work was supervised by SD and PR. AB prepared the figures and tables. AB, SD and PR wrote the manuscript together. All authors discussed the results from the experiments and commented on the manuscript.

Acknowledgements

The authors thank Le ministère de l'Europe et des Affaires Étrangères (MEAE) for its financial support to travel grants to STL, SNS and KA (PANASIA project). This work was also supported by grants from the Ministère de l'enseignement supérieur et de la recherche (MENRT to AB).

Appendix A. Supplementary data

CAS number, name and structural class of all molecules used to compare the difference between their experimental chemical shifts and the chemical shifts that were predicted by a software for these same structures (Table S1); CAS numbers of PPAPs dereplicated and then isolated from *Garcinia bancana* bark (Table S2).

The DB3 containing the predicted chemical shifts [C + H NMR Predictor (ACD/Labs)] of 718 compounds described in *Garcinia* species (Dictionary of Natural Products) is available as a SDF: *Garcinia* DB ¹³C NMR Chemical shifts.sdf. Supplementary data to this article can be found online at doi: <https://doi.org/10.1016/j.fitote.2018.10.003>.

References

- [1] P. Sharma, J.P. Allison, The future of immune checkpoint therapy, *Science* 348 (6230) (2015) 56–61.
- [2] S.L. Topalian, J.M. Taube, R.A. Anders, D.M. Pardoll, Mechanism-driven biomarkers to guide immune checkpoint blockade in cancer therapy, *Nat. Rev. Cancer* 16 (5) (2016) 275–287.
- [3] C. Rouger, S. Pagie, S. Derbré, A.M. Le Ray, P. Richomme, B. Charreau, Prenylated polyphenols from Clusiaceae and Calophyllaceae with immunomodulatory activity on endothelial cells, *PLoS One* 11 (12) (2016) e0167361.
- [4] X.-W. Yang, R.B. Grossman, G. Xu, Research progress of polycyclic polyprenylated acylphloroglucinols, *Chem. Rev.* 118 (7) (2018) 3508–3558.
- [5] J. Merza, S. Mallet, M. Litaudon, V. Dumontet, D. Séraphin, P. Richomme, New cytotoxic guttiferone analogues from *Garcinia virgata* from New Caledonia, *Planta Med.* 72 (01) (2006) 87–89.
- [6] G. Xu, C. Feng, Y. Zhou, Q.-B. Han, C.-F. Qiao, S.-X. Huang, D.C. Chang, Q.-S. Zhao, K.Q. Luo, H.-X. Xu, Bioassay and ultraperformance liquid chromatography/mass spectrometry guided isolation of apoptosis-inducing benzophenones and xanthone from the pericarp of *Garcinia yunnanensis* Hu, *J. Agric. Food Chem.* 56 (23) (2008) 11144–11150.
- [7] M. Hemshekar, K. Sunitha, M.S. Santhosh, S. Devaraja, K. Kemparaju, B.S. Vishwanath, S.R. Niranjana, K.S. Girish, An overview on genus *Garcinia*: phytochemical and therapeutical aspects, *Phytochem. Rev.* 10 (3) (2011) 325–351.
- [8] R.B. Grossman, Table of Naturally Occurring PPAPs, <http://www.chem.uky.edu/research/grossman/PPAPs/allPPAPs.html#list>, (2018), Accessed date: June 2018.
- [9] J. Chervin, M. Stierhof, M.H. Tong, D. Peace, K.O. Hansen, D.S. Urgast, J.H. Andersen, Y. Yu, R. Ebel, K. Kyeremeh, V. Paget, G. Cimpan, A.V. Wyk, H. Deng, M. Jaspars, J.N. Tabudravu, Targeted dereplication of microbial natural products by high-resolution MS and predicted LC retention time, *J. Nat. Prod.* 80 (5) (2017) 1370–1377.
- [10] A.E. Fox Ramos, C. Alcover, L. Evanno, A. Maciuk, M. Litaudon, C. Duplais, G. Bernadat, J.F. Gallard, J.C. Jullian, E. Mouray, P. Grellier, P.M. Loiseau, S. Pomel, E. Poupon, P. Champy, M.A. Benidrir, Revisiting previously investigated plants: a molecular networking-based study of *Geissospermum laeve*, *J. Nat. Prod.* 80 (4) (2017) 1007–1014.
- [11] G. Marti, V. Eparvier, C. Moretti, S. Prado, P. Grellier, N. Hue, O. Thoisson, B. Delpech, F. Gueritte, M. Litaudon, Antiplasmodial benzophenone derivatives from the root barks of *Symphonia globulifera* (Clusiaceae), *Phytochemistry* 71 (8–9) (2010) 964–974.
- [12] J. Hubert, J.-M. Nuzillard, J.-H. Renault, Dereplication strategies in natural product research: how many tools and methodologies behind the same concept? *Phytochem. Rev.* 16 (1) (2015) 55–95.
- [13] A. Bakiri, J. Hubert, R. Reynaud, S. Lanthony, D. Harakat, J.H. Renault, J.M. Nuzillard, Computer-aided ¹³C NMR chemical profiling of crude natural extracts without fractionation, *J. Nat. Prod.* 80 (5) (2017) 1387–1396.
- [14] A.L. Guennec, P. Giraudeau, S. Caldarelli, Evaluation of fast 2D NMR for metabolomics, *Anal. Chem.* 86 (12) (2014) 5946–5954.
- [15] P. Le Pogam, A. Schinkovitz, B. Legouin, A.-C. Le Lamer, J. Boustie, P. Richomme, Matrix-free UV-laser desorption ionization mass spectrometry as a versatile approach for accelerating dereplication studies on lichens, *Anal. Chem.* 87 (20) (2015) 10421–10428.
- [16] A. Schinkovitz, S. Boisard, I. Freuze, J. Osuga, N. Mehlmer, T. Brück, P. Richomme, Matrix-free laser desorption ionization mass spectrometry as a functional tool for the analysis and differentiation of complex phenolic mixtures in propolis: a new approach to quality control, *Anal. Bioanal. Chem.* 410 (24) (2018) 6187–6195.
- [17] F. Toml, P. Bradesi, A. Bighelli, J. Casanova, Computer aided identification of individual components of essential oils using carbon 13 NMR spectroscopy, *J. Magn. Reson.* 1 (1995) 25–34.
- [18] G. Marti, V. Eparvier, C. Moretti, S. Susplugas, S. Prado, P. Grellier, P. Retailleau, F. Gueritte, M. Litaudon, Antiplasmodial benzophenones from the trunk latex of *Moronobea coccinea* (Clusiaceae), *Phytochemistry* 70 (1) (2009) 75–85.
- [19] Z. Miao, M. Jin, X. Liu, W. Guo, X. Jin, H. Liu, Y. Wang, The application of HPLC and microprobe NMR spectroscopy in the identification of metabolites in complex biological matrices, *Anal. Bioanal. Chem.* 407 (12) (2015) 3405–3416.
- [20] J. Hubert, J.M. Nuzillard, S. Purson, M. Hamzaoui, N. Borie, R. Reynaud, J.H. Renault, Identification of natural metabolites in mixture: a pattern recognition strategy based on ¹³C NMR, *Anal. Chem.* 86 (6) (2014) 2955–2962.
- [21] M.J.P. Ferreira, M.B. Costantin, P. Sartorelli, G.V. Rodrigues, R. Limberger, A.T. Henriques, M.J. Kato, V.P. Emerenciano, Computer-aided method for identification of components in essential oils by ¹³C NMR spectroscopy, *Anal. Chim. Acta* 447 (1–2) (2001) 125–134.
- [22] A. Bakiri, B. Plainchont, V. de Paulo Emerenciano, R. Reynaud, J. Hubert, J.H. Renault, J.M. Nuzillard, Computer-aided dereplication and structure elucidation of natural products at the University of Reims, *Mol. Inform.* 36 (10) (2017) 1700027.
- [23] G. Caraux, S. Pinloche, PermutMatrix: a graphical environment to arrange gene expression profiles in optimal linear order, *Bioinformatics* 21 (7) (2005) 1280–1281.
- [24] M. Iinuma, H. Tosa, T. Tanaka, S. Kanamaru, F. Asai, Y. Kobayashi, K. Miyauchi, R. Shimano, Antibacterial activity of some *Garcinia* benzophenone derivatives against methicillin-resistant *Staphylococcus aureus*, *Biol. Pharm. Bull.* 19 (2) (1996) 311–314.
- [25] R.W. Fuller, J.W. Blunt, J.L. Boswell, J.H. Cardellina, M.R. Boyd 2nd, Guttiferone F, the first prenylated benzophenone from *Allanblackia stuhlmannii*, *J. Nat. Prod.* 62 (1) (1999) 130–132.
- [26] ChemNetBase, Dictionary of Natural Products 26.2, <http://dnp.chemnetbase.com>, (2015), Accessed date: November 2015.
- [27] JEOL, K. Hayamizu, Natural Product NMR-DB "CH-NMR-NP", <https://www.j-resonance.com/en/nmrdb/>, (2000–2014), Accessed date: April 2017.
- [28] V. Rukachaisirikul, W. Naklue, Y. Sukpondma, S. Phongpaichit, An antibacterial biphenyl derivative from *Garcinia bancana* Miq, *Chem. Pharm. Bull.* 53 (3) (2005) 342–343.
- [29] I. Jantan, S.H. Goh, Flavonoids, xanthones and triterpenes of the leaves and heartwood of *Garcinia bancana* Miq, *Sains Malaysiana* 24 (1995) 23–30.
- [30] B. Elya, K. Basah, A. Mun'im, W. Yuliatuti, A. Bangun, E.K. Septiana, Screening of α-glucosidase inhibitory activity from some plants of Apocynaceae, Clusiaceae, Euphorbiaceae, and Rubiaceae, *J. Biomed. Biotechnol.* 2012 (2012) 6.

Supplementary data

**¹³C-NMR DEREPLICATION OF GARCINIA EXTRACTS: PREDICTED
CHEMICAL SHIFTS AS RELIABLE DATABASES**

Antoine Bruguère^a, Séverine Derbré^{a*}, Chloé Coste^a, Maxime Le Bot^a, Benjamin Siegler^b,
Sow Tein Leong^c, Syazreen Nadia Sulaiman^c, Khalijah Awang^c, Pascal Richomme^a

^a *SONAS SFR QUASAV, University of Angers, France.*

^b *SFR MATRIX, University of Angers, France.*

^c *Department of Chemistry, Faculty of sciences, University of Malaya, Malaysia*

* To whom correspondence should be addressed. Tel: 33 249 180 440. E-mail:
severine.derbre@univ-angers.fr (S.D.).

S1

Supplementary data

Table S1 CAS number, name and structural class of all molecules used to compare the difference between their experimental chemical shifts and the chemical shifts that were predicted by a software for these same structures.

CAS number	Molecule	Structural class
597564-82-4	4,6,4'-Trihydroxy-2,3'-dimethoxy-3-prenylbenzophenone	Benzophenone
597564-83-5	4,6,3',4'-Tetrahydroxy-2-methoxybenzophenone	Benzophenone
686315-69-5	Garciosaphenone A	Benzophenone
81827-55-6	Kolanone	Benzophenone
*	Multibiphenyl A	Biphenyl
1850284-78-4	Multibiphenyl B	Biphenyl
1850284-79-5	Multibiphenyl C	Biphenyl
856451-92-8	Garcibiphenyl A	Biphenyl
856451-93-9	Garcibiphenyl B	Biphenyl
916206-07-0	Garcibiphenyl D	Biphenyl
916206-09-2	Garcibiphenyl E	Biphenyl
356053-82-2	Atrovirione	Biphenyl
18196-05-9	Apetalic acid	Chromanone
17243-98-0	Blancoic acid	Chromanone
29077-60-9	Brasilensic acid	Chromanone
686318-86-5	Caledonic acid	Chromanone
686318-85-4	Calolongic acid	Chromanone
34366-34-2	Isoapetalic acid	Chromanone
40737-92-6	Isocalolongic acid	Chromanone
142382-37-4	Pseudobrasilensic acid	Chromanone
*	Demethylisorecedensone	Chromanone
686318-87-6	Calolongic acid lactone	Chromanone
382146-54-5	Dispariol B	Coumarin
221379-33-5	Disparinol A	Coumarin
382146-52-3	Disparinol D	Coumarin
1805814-45-2	Lepidotin A	Coumarin
1805814-40-7	Lepidotol A	Coumarin
1805814-44-1	Lepidotol E	Coumarin
18483-64-2	Mammea A/AA	Coumarin
2289-11-4	Mammea A/AA cyclo D	Coumarin
30563-62-3	Mammea A/AA cyclo F	Coumarin
382146-55-6	Mammea A/AB cyclo E	Coumarin
1805814-47-4	Mammea A/OB	Coumarin
30390-12-6	Mammea B/AB	Coumarin
5022-20-8	Mammea B/BB	Coumarin
906794-54-5	Parvifolidone A	Depsidone
906794-55-6	Parvifolidone B	Depsidone
299188-32-2	Garcidepsidone A	Depsidone
299188-56-0	Garcidepsidone B	Depsidone
299188-81-1	Garcidepsidone C	Depsidone

S2

Supplementary data

299188-86-6	Garcidepsidone D	Depsidone
356053-83-3	Atrovirisidone	Depsidone
1141871-31-9	Coccinone A	PPAP
1141870-93-0	Coccinone B	PPAP
1141870-95-2	Coccinone D	PPAP
1141870-97-4	Coccinone F	PPAP
1141870-99-6	Coccinone G	PPAP
1141871-01-3	Coccinone H	PPAP
219793-20-1	Furohyperforin	PPAP
78824-30-3	Garcinol	PPAP
1073496-59-9	Garciyunnanin A	PPAP
1073496-60-2	Garciyunnanin B	PPAP
11079-53-1	Hyperforin	PPAP
71117-97-0	Isogarcinol	PPAP
1639050-36-4	γ -(E)- deoxyamplexichromanal	Tocotrienol
2086743-22-6	γ -(E)-amplexichromanoic acid	Tocotrienol
1492001-47-4	γ -(Z)- deoxyamplexichromanol	Tocotrienol
1492001-45-2	γ -amplexichromanol	Tocotrienol
1639050-39-7	δ -(E)-amplexichromanal	Tocotrienol
1639050-37-5	δ -(E)- deoxyamplexichromanol	Tocotrienol
1492001-43-0	δ -amplexichromanol	Tocotrienol
14440-41-6	Canophyllol	Triterpene
83-48-7	Stigmasterol	Triterpene
83-46-5	β -sitosterol	Triterpene
508-02-1	Oleanolic acid	Triterpene
545-47-1	Lupeol	Triterpene
559-74-0	Friedelin	Triterpene
33018-28-9	3-hydroxy-2- methoxyxanthone	Xanthone
105037-94-3	6-deoxy- γ -mangostin	Xanthone
686318-83-2	Caledonixanthone M	Xanthone
105037-93-2	Calothwaitesixanthone	Xanthone
155566-36-2	Caloxanthone C	Xanthone
686318-84-3	Caloxanthone L	Xanthone
78859-48-0	Calozeyloxanthone	Xanthone
106897-03-4	Demethylcalabaxanthone	Xanthone
201859-39-4	Dombakinaxanthone	Xanthone
5848-14-6	Macluraxanthone	Xanthone
84002-57-3	Padiaxanthone	Xanthone
55785-61-0	Thwaitesixanthone	Xanthone

*Not found

Supplementary data

Table S2 CAS numbers of PPAPs dereplicated and then isolated from *Garcinia bancana* bark (MSI level 1).

CAS number	Molecule
52617-32-0	Xanthochymol
219538-86-0	Guttiferone F
219524-66-0	30-epi-cambogin

Ces expériences ont permis de souligner qu'il existe des différences notables de qualité des estimations selon les différents logiciels de prédictions. Pour une structure donnée, ACD/Labs® [12] prédit en moyenne 75% des signaux dans un intervalle de 1,25 ppm autour des valeurs expérimentales. Les logiciels MestReNova® [13] et ChemDraw® [14], quant à eux, nécessitent un intervalle de respectivement 2,00 et 3,00 ppm pour recouvrir 75% des signaux, $\frac{3}{4}$ des signaux étant empiriquement une valeur reflétant une bonne identification de molécule. On observe que les déplacements chimiques de certaines classes structurales sont prédits avec plus de précision que d'autres, ce qui peut notamment être lié à l'abondance respective de leur description dans la littérature (cf. **1.1. L'étape clé des références**). Cependant, aucun lien n'a pu être établi entre qualité de la précision et un type de carbone donné.

ACD/Labs® [12], qui proposait le logiciel de prédiction le plus précis, a donc été choisi pour la construction des bases de données prédictives. L'algorithme de recherche d'ACD/Labs® a aussi été utilisé pour l'étape de comparaison, en tenant compte des paramètres expérimentalement définis (75% des δ_c trouvés pour une marge d'erreur de $\pm 1,25$ ppm). Concernant la base de données expérimentale, celle-ci est indissociable de l'algorithme de recherche proposé par le site CH-NMR-NP [17], c'est donc ce dernier qui a été utilisé pour l'étape de comparaison.

Les recherches avec les deux différents types de bases ont donné des résultats très similaires. La présence de stéréoisomères des garcinol, xanthochymol et cycloxanthochymol était suggérée par les algorithmes. Les hypothèses ont été confirmées concernant le stéréoisomère du garcinol (guttiferone F) et du xanthochymol lors de la purification des produits présents dans les fractions (cf. **1.2.2. Article 2**).

Après la publication de ces travaux, des purifications additionnelles de composés ont permis de confirmer la présence de (-)-cycloxanthochymol et de 30-epi-cambogin au sein de ce même extrait. Cela permet d'appuyer le fait que les hypothèses proposées par les deux types de banques de données sont correctes ou très proches de la réalité.

Dans notre étude, l'utilisation du logiciel ACD/Labs® [12] pour prédire les bases de données et son utilisation lors d'un processus de déréplication a donc donné des résultats de qualité comparable à celle d'une base expérimentale avec l'avantage, bien sûr, de pouvoir construire en mode semi-automatisé (utilisation de macros informatiques) - et donc très rapidement - de larges bases de δ_c prédits.

2. Algorithme de déréplication

La nature des bases de données à utiliser étant maintenant connue, la partie de comparaison automatisée des *data*, c'est-à-dire les différents algorithmes utilisés pour la déréplication par RMN-¹³C, reste à explorer. Certains logiciels commerciaux permettent déjà la comparaison informatique des jeux de données, mais l'effet « boîte noire » masquant leur fonctionnement précis, ne permet parfois pas de pouvoir expliquer et discuter les hypothèses qu'ils proposent. Ils ne permettent pas non plus de pouvoir apporter de nouvelles données aux bases expérimentales sur lesquels leur prédiction est basée, et donc de potentiellement améliorer la qualité de la prédiction. De ce fait, certaines équipes de recherches s'orientent vers la création de leurs propres algorithmes de recherche. A notre connaissance, pour la première fois en avril 2017, Bakiri *et al.* ont mis librement à disposition l'algorithme codé par leurs soins, appelé **DerepCrude**, permettant la déréplication d'extrait brut par RMN-¹³C, sans étape de fractionnement [6]. L'étude a été conduite sur un extrait d'alcaloïdes totaux obtenu à partir des feuilles de *Peumus boldus* (Monimiaceae), plante contenant diverses isoquinoléines de type aporphinique.

Deux points ont attiré notre attention. D'abord, cette méthode propose de travailler directement sur des extraits bruts, ce qui permet de ne pas passer par des étapes de fractionnement pour formuler des hypothèses sur la composition des échantillons. Ensuite le code est disponible, ce qui nous autorise à tester ce nouvel algorithme et à le comparer aux outils déjà à notre disposition, c'est-à-dire l'algorithme commercial de recherche d'ACD/Labs® [12], intégré au logiciel de bases de données, et l'algorithme de recherche de la société JEOL, lié à la base CH-NMR-NP disponible en ligne [17].

Après avoir présenté les éléments essentiels nécessaires au fonctionnement des algorithmes existants, les résultats de tests menés avec ces derniers seront présentés. Puis, grâce à la lecture du code de l'algorithme DerepCrude proposé par Bakiri *et al.*, une analyse plus poussée de son fonctionnement sera détaillée afin de chercher des terrains d'optimisation qui seront exposés dans la dernière partie.

2.1. Éléments de fonctionnement des algorithmes de déréplication par RMN-¹³C existants

Il convient de succinctement décrire chacun des algorithmes ainsi que les éléments nécessaires à la recherche car, même s'ils sont similaires, ils comportent chacun des spécificités. Tous comportent 4 types d'éléments : (A) les données initiales, (B) les paramètres de recherche, (C) les systèmes de filtrage des résultats ou des banques de données, (D) le calcul et la présentation des résultats.

Concernant l'algorithme proposé par ACD/Labs® [12], l'utilisateur rentre (A) les déplacements chimiques qu'il souhaite soumettre à la recherche (Figure 3). La base de données utilisée pour la comparaison est un fichier SDF (Structure Data File), ensemble de fichier .mol (fichiers de structure de molécules), dans lequel la valeur (expérimentale ou prédite) des déplacements chimiques carbone (δ_c) est incluse. (B) Un « *looseness factor* », qui représente la marge autorisée pour un déplacement chimique donné (\pm ppm), et donc la précision attendue, doit être renseigné. (C) Un *minimal number of shifts to match* doit également être renseigné, ce qui empêche les

molécules trop « petites » (faible nombre de carbone) d'apparaître dans les résultats. Si l'utilisateur connaît la classe chimique recherchée, ce qui est loin d'être toujours le cas, nous avons observé expérimentalement qu'une valeur de *minimal number of shifts to match* située entre 50% et 75% du nombre moyen de carbones associé à cette classe chimique donnait d'excellents résultats. Choisir ce nombre de *shifts* minimal devient plus problématique quand des échantillons contenant des molécules de tailles inconnues et/ou très différentes sont analysés. **(B)** Le dernier paramètre est la possibilité de cocher une case « *Do not match one chemical shift for several shift queries* », ce qui permet d'autoriser ou non la réutilisation de déplacements chimiques (en cas de carbones équivalents par exemple). **(D)** Les résultats seront automatiquement classés en fonction d'un score calculé qui prend en compte le nombre de déplacement chimiques qui ont pu être associés entre la banque de données et les données de l'utilisateur. **(C)** Une fois les résultats obtenus, il est possible de les filtrer en utilisant n'importe quelle information stockée dans la base de données utilisée (poids moléculaire, formule brute, etc...).

Figure 3: Fenêtre de recherche dans ACD Labs®.

L'algorithme disponible sur le site **CH-NMR-NP** [17] requiert également **(B)** une « *allowance* » en ppm qui sert de marge autorisée pour **(A)** chacun des déplacements chimiques fournis par l'utilisateur (**Figure 4**). Ce paramètre est fixé par défaut à 2 ppm mais, la base JEOL étant une base de données de déplacements chimiques expérimentaux, cette marge peut être largement réduite (jusqu'à 0,1 ppm). **(C)** Un pourcentage de similarité minimum, s'apparentant à un score (nombre de signaux associés / nombre de signaux totaux de la molécule) peut également être utilisé comme filtre. Il est aussi possible d'ajouter des filtres additionnels, comme des poids moléculaires, renseigner les régions de spectre RMN-¹³C dépourvues de signaux, ou bien la formule brute de la molécule. **(D)** Les molécules sont ensuite présentées par pourcentage de similarité décroissant.

NMR Database Search

Basic Information

☒ Name
Example: Pseudoanchnazine / Pseudo*zine*

☒ Atoms C H N O
Example: C21-23 H18 N4 O5

☒ Molecular Formula
Example: C15H18BrS2

☒ Molecular Weight
Example: 545 / 545 - 558

☒ ¹³C Chemical Shift ±
± Allowance / ppm Example: 40, 41, 71 ± 2 Similarity ≥ %

☒ ¹³C No Signal Region to ppm
Example: 40 to 41 ppm

☐ Structure Search To search structure, [Java Runtime](#) needs to be installed. Please see page 6 in [the instruction manual](#) if the structure search doesn't work well.

☒ NP No.
Example: 15 / 30 - 100

☒ CAS Registry No.
Example: 59392-53-9 / 5932-*

Sort Molecular Formula (C) ▼
☐ Ascending
☒ Descending

☐ Enable Detailed Search

Figure 4: Fenêtre de recherche sur CH-NMR-NP.

Le script Python **DerepCrude** développé par Bakari *et al.* [6] nécessite (A) la liste des déplacements chimiques mais également leurs intensités respectives qui sont fournies par l'utilisateur sous la forme d'un fichier .txt. La base de données à utiliser est soumise sous la forme d'un SDF, ensemble de fichiers .mol (fichiers de structure de molécules), dans lequel la valeur des déplacements chimiques est incluse et formatée de façon à être correctement reconnue par l'algorithme. (B) Le paramètre de marge autorisée, « *looseness factor* », doit être renseigné par l'utilisateur, les créateurs de l'algorithme recommandant ± 1 ppm. (C) Un score minimum doit également être indiqué, ce qui permettra d'uniquement afficher des résultats ayant dépassé un certain score (le score étant le rapport entre le nombre de δ_C associés et le nombre total de carbones au sein d'une molécule). Enfin, un nombre e d'écart-type σ doit être choisi. Ce dernier permettra de prendre en compte l'intensité des pics associés, et de faire en sorte que seuls des déplacements chimiques dont les intensités sont comprises entre la moyenne \pm le nombre d'écart-type choisi ($e \times \sigma$) soient pris en compte pour le *matching* (cf. 2.3. **Fonctionnement de l'algorithme DerepCrude**). La valeur par défaut est de 2 écart-types. (D) Les différents composés sont classés par scores décroissants dans le fichier de résultats.

Quant au fonctionnement interne de ces algorithmes, il est difficile de le décrire de façon certaine pour les commerciaux (ACD/Labs® [12], CH-NMR-NP [17]) qui agissent comme des boîtes noires. Pour DerepCrude [6], les informations communiquées par publication et l'étude du code Python ont permis de décrypter son

fonctionnement. Celui-ci sera détaillé et discuté dans le paragraphe **2.3. Fonctionnement de l'algorithme DerepCrude**.

2.2. Essais de déréplication

2.2.1. Partie expérimentale

Afin d'évaluer la qualité des résultats des différents algorithmes à notre disposition, un second exemple a été choisi pour une étude déréplicative. Il s'agit cette fois d'une fraction enrichie d'*Allanblackia floribunda* (Clusiaceae) qui nous a été fournie pour analyse dans le cadre d'une collaboration avec l'équipe de Félix Tomi (Université de Corse Pascal Paoli). Un profil HPLC-UV de la fraction a été obtenu en phase inverse avec une colonne Lichrospher (150 mm x 4,6 mm ; 5 µm), avec un gradient démarrant à 5% de méthanol / 95% d'eau + 0,1 % d'acide formique pour atteindre 100% de méthanol / 0% d'eau + 0,1% d'acide formique en 50 min. Le débit est de 1 mL / min et la détection UV se fait à 254 nm. Le chromatogramme est disponible en **Figure 5**, accompagné de la structure des molécules qui ont été ensuite identifiées au sein de la fraction.

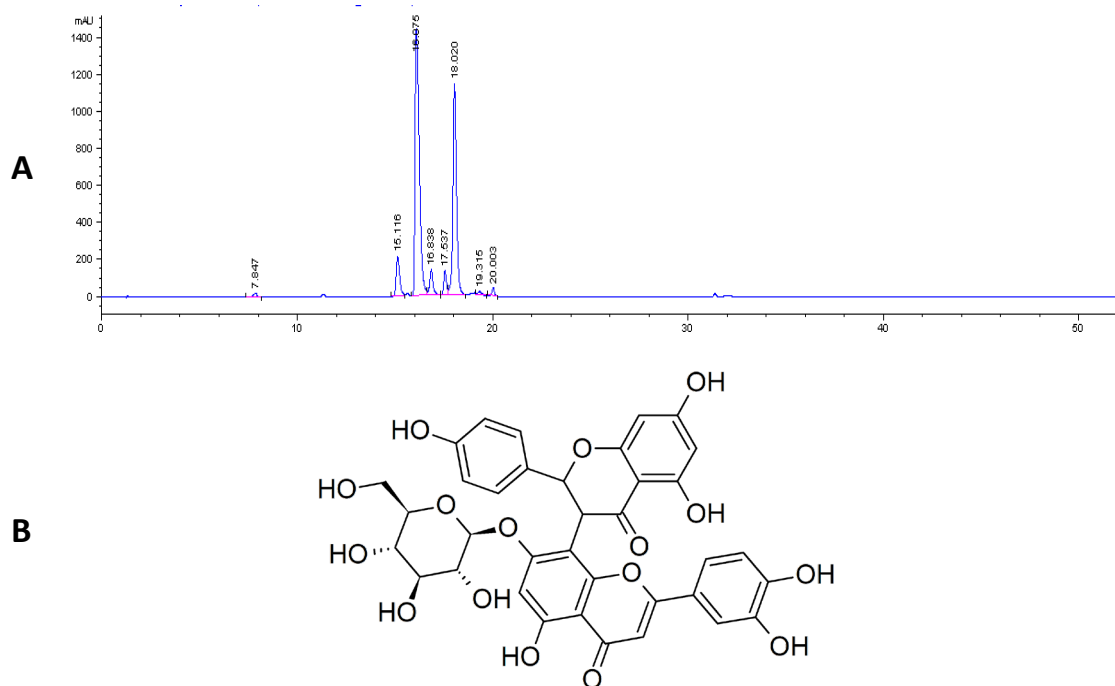


Figure 5: (A) Profil de l'extract d'*Allanblackia floribunda* à 254 nm. (B) Type de structure identifiée dans la fraction : fukugiside.

Les bases de données utilisées pour ces essais ont été (1) la base de donnée de déplacements chimiques réels CH-NMR-NP [17] (puisque indissociable de l'algorithme de recherche du site), (2) une base de données de déplacement chimiques prédits avec ACD/Labs® [12] «*Garcinia* », construite à partir des structures des molécules reportées dans le genre *Garcinia* d'après le Dictionary of Natural Products (DNP) [18], (3) une base de données chimiques prédits avec ACD/Labs® [12] «*Allanblackia* », construite à partir des structures des molécules reportées dans le genre *Allanblackia* d'après une recherche sur SciFinder [19]. Il convient de noter que ces bases de données sont de tailles très différentes puisqu'elles comportent respectivement (1) 30 500, (2) 718 et (3) 39

molécules. Les bases de données *Garcinia* et *Allanblackia* ont été utilisées avec les algorithmes d'ACD/Labs® [12] et DerepCrude [6].

2.2.2. Résultats

Pour l'algorithme CH-NMR-NP et sa base de données associée [17], en suivant les paramètres de recherche recommandés par le site et décrits dans la partie 2.1. **Éléments de fonctionnement des algorithmes de déréplication par RMN-¹³C existants**, le premier résultat fût le fukugiside, un biflavonoïde glycosylé, dont la présence a été ensuite confirmée par comparaison manuelle des signaux de la fraction (Figure 5) avec ceux publiés [20].

Pour l'algorithme d'ACD/Labs® [12], la recherche dans la base *Allanblackia* avec un *looseness factor* de 0,9 ppm et au moins 16 carbones requis pour un *match* (70 % des 24 carbones d'une xanthone), propose 2 structures biflavonoïdiques : la GB 2a et la (+)-volkensiflavone (Figure 6).

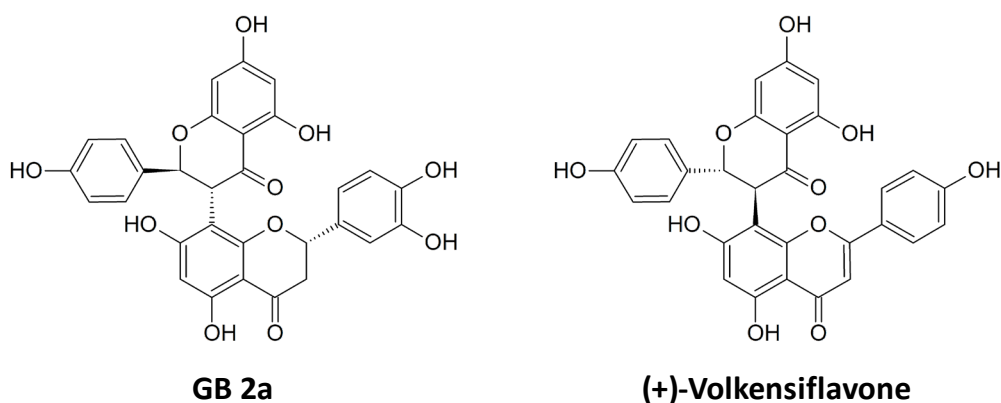


Figure 6: Résultats proposés par l'algorithme ACD/Labs® avec la base de données *Allanblackia* concernant la composition de la fraction enrichie d'*Allanblackia floribunda*.

Une recherche dans la base de données *Garcinia* avec les mêmes paramètres (0,9 ppm de *looseness factor* et 16 carbones minimum pour un *match*), aboutit à 26 molécules possédant également des structures biflavonoïdiques, la plupart d'entre elles étant des hétérosides. Afin de réduire le nombre de résultats, la recherche a été renouvelée en réduisant le *looseness factor* à 0,5 ppm. 4 molécules sont suggérées : le spicataside, le fukugiside, la rhusflavanone et le GB 1a glucoside (Figure 7).

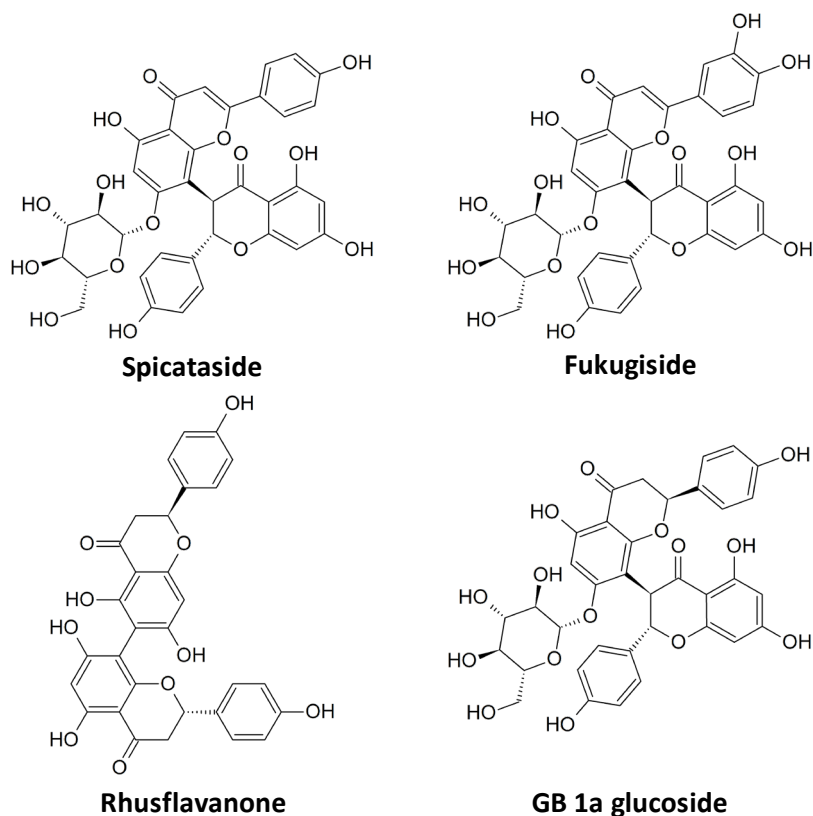


Figure 7: Résultats proposés par l'algorithme ACD/Labs® avec la base de données *Garcinia* concernant la composition de la fraction enrichie d'*Allanblackia floribunda*.

L'algorithme DerepCrude de Bakiri *et al.* [6] a également été testé sur ce même exemple, d'abord avec la base *Allanblackia*. Les paramètres utilisés sont ceux recommandés par les créateurs du script, à savoir un *looseness factor* de 1 ppm pour la recherche et 2 écart-types pour le filtre d'intensité. Le score minimal a été placé à 0,70 afin de limiter le nombre de résultats. Les molécules proposées sont la (+)-volkensiflavone, la fukugetine, le fukugiside, la GB 2a, la 4-phénylcoumarine et la catéchine (**Figure 8**).

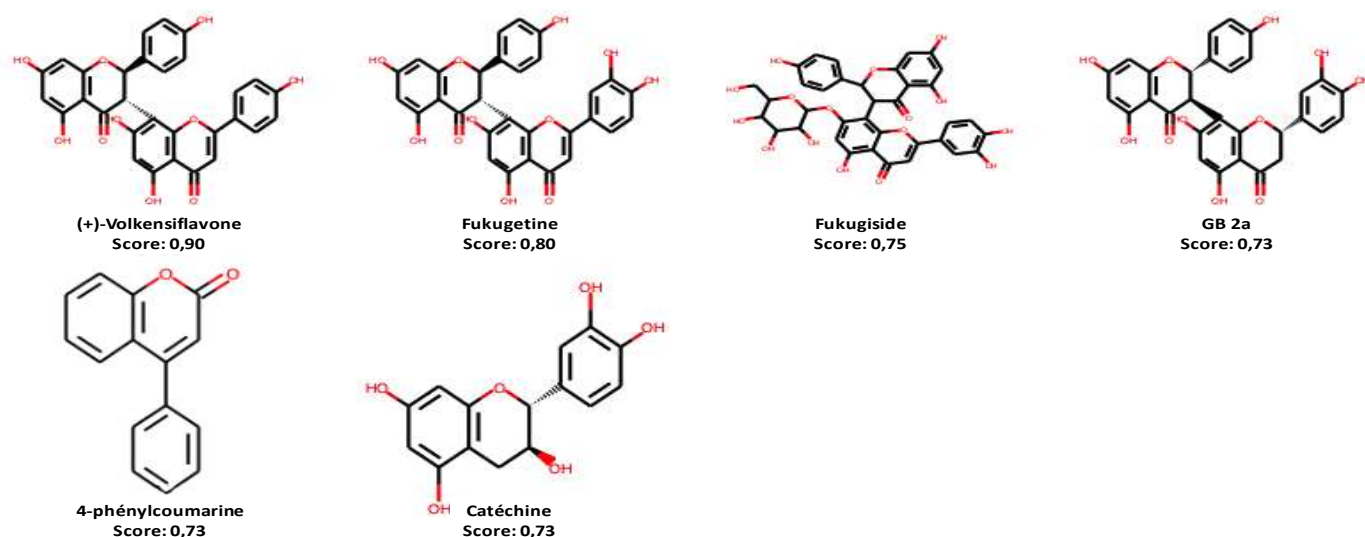
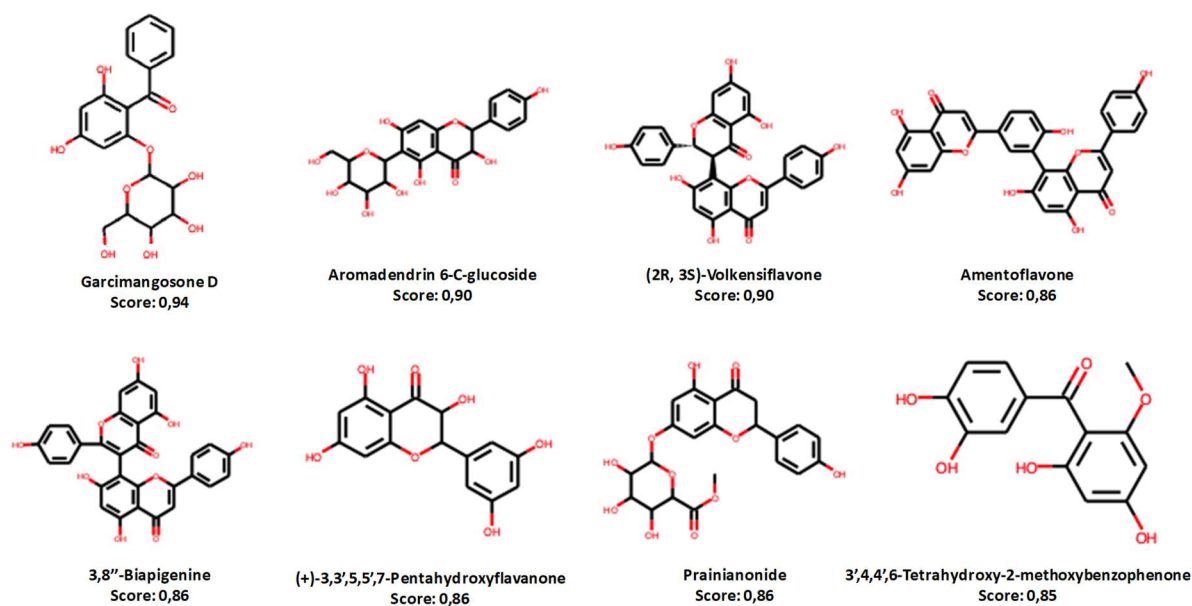


Figure 8: Résultats proposés par l'algorithme DerepCrude avec la base de données *Allanblackia* concernant la composition de la fraction enrichie d'*Allanblackia floribunda*.

Une recherche avec les mêmes paramètres, mais cette fois dans la base *Garcinia*, donne 39 propositions de molécules, dont les 24 premières sont présentées en **Figure 9**.



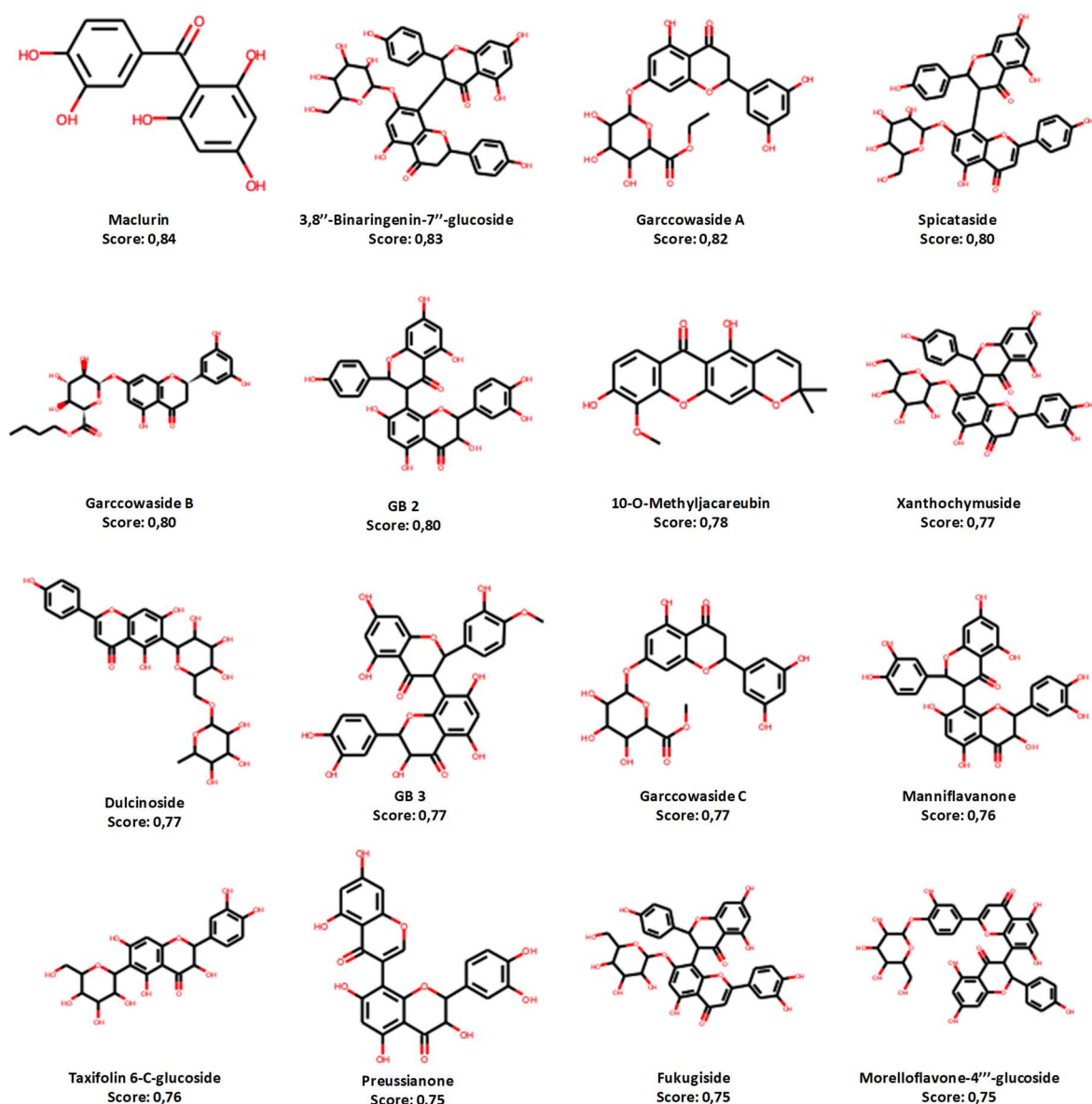


Figure 9 : Résultats proposés par l'algorithme DerepCrude avec la base de données *Garcinia* concernant la composition de la fraction enrichie d'*Allanblackia floribunda*.

2.2.3. Discussion

La présence du fukugiside dans la fraction enrichie ayant été confirmée (**Tableau 1**), il est possible d'évaluer la pertinence des réponses données par les différents algorithmes. Notons également que, même si la seconde molécule majoritaire n'a pas été formellement identifiée, on peut supposer, d'après les déplacements chimiques résiduels (*i. e.* non attribués), qu'il s'agit d'un hétéroside comportant deux sucres et une génine de structure flavonoïdique (**Tableau 1**).

Tableau 1: Déplacements chimiques carbone 13 en ppm correspondant au fukugiside et aux signaux restants.

Extrait (MeOD)	Fukugiside (MeOD)	Extrait (MeOD)	Fukugiside (MeOD)	Signaux restants	
197,5	197,6	115,4	115,5	198,5	102,1
184,0	183,9	114,2	114,4	184,1	101,0
168,3	168,3	106,5	106,5	168,8	96,6
166,2	166,1	104,1	104,1	162,9	96,4
165,8	165,7	103,6	103,6	162,7	84,1
164,9	164,8	103,3	103,3	159,1	78,2
162,8	162,7	101,5	101,5	132,3	78,0
161,7	161,6	99,6	99,5	129,9	74,7
158,5	158,5	97,6	97,7	129,8	74,6
156,7	156,7	96,5	96,5	129,5	62,8
151,3	151,1	82,8	82,8	129,5	62,2
146,9	146,7	78,6	78,5	129,2	49,9
130,3	130,4	78,3	78,3	122,7	30,8
129,4	129,4	75,2	75,2	117,0	19,3
123,1	123,1	71,0	71,1	116,4	15,0
120,8	120,8	62,4	62,5	116,0	14,4
116,9	116,9	51,0	51,0	104,5	9,2

La recherche sur CH-NMR-NP [17] illustre parfaitement la puissance des bases de données expérimentales lorsque les échantillons sont analysés dans un solvant deutéré identique : on retrouve la molécule correcte avec un score maximal.

La recherche à partir de la base *Allanblackia*, avec l'algorithme ACD/Labs® [12], donne cependant moins d'informations car elle présente uniquement des génines analogues des molécules présentes dans la fraction, alors que la base contient bien des hétérosides. La base prédite *Garcinia* donne tout de même un résultat permettant directement de s'orienter vers un type de structure particulier (biflavonoïdes glycosylés), et comporte également le fukugiside parmi ces propositions.

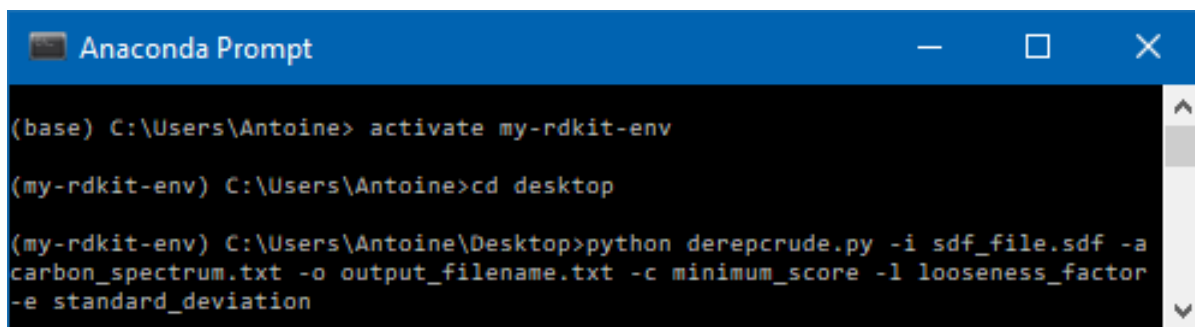
Avec l'algorithme DerepCrude [6], une recherche dans la base *Allanblackia* permet d'identifier le fukugiside assez rapidement, même si les autres hypothèses orienteraient plus vers des biflavonoïdes non glycosylés. Le passage sur la base *Garcinia*, plus large en termes de nombre d'entrées, rend le processus de déréplication plus difficile pour cet algorithme. En effet, même s'il présente en score correct, le fukugiside n'arrive qu'en 23^{ème} position. On remarque aussi que de nombreuses classes structurales sont suggérées, avec un score important, avant d'arriver au premier biflavonoïde glycosylé (10^{ème} position) : on retrouve ainsi des benzophénones (glycosylées ou non), des flavonoïdes (glycosylés ou non) et des biflavonoïdes non glycosylés.

On remarque que ces 3 algorithmes produisent des résultats relativement différents, et que le choix d'une banque de données influence considérablement le processus de recherche. Une base de données plus large, dans notre cas, la base *Garcinia*, peut soit permettre de trouver les molécules correctes (ACD/Labs® [12]), soit « noyer » les bons résultats parmi le nombre trop important de molécules (DerepCrude [6]). Bien qu'il ne soit pas possible de savoir comment l'algorithme proposé par ACD/Labs® fonctionne pour expliquer ces résultats, l'algorithme

DerepCrude présente en revanche cet avantage : il est possible de savoir comment le procédé de *matching* (c'est-à-dire d'association des δ_c) est réalisé par examen du code associé. On peut donc ainsi chercher à comprendre pourquoi certaines molécules sont associées d'une façon ou d'une autre.

2.3. Fonctionnement de l'algorithme DerepCrude

Le script DerepCrude proposé par Bakiri *et al.* [6] a été codé avec le langage de programmation informatique Python (version 2.7) [21]. A partir du spectre RMN- ^{13}C considéré sont récupérés la liste des déplacements chimiques, et leurs intensités respectives, sous la forme d'un fichier .txt. Ce fichier sera traité par le programme et comparé à une base de données fournie par l'utilisateur sous forme d'un fichier SDF (ensemble de fichiers .mol, fichiers de structure de molécules ; sera illustré dans le chapitre 2.5. **Description de l'algorithme MixONat** en Figure 16 et Figure 17). Après avoir rentré les paramètres précédemment décrits, à savoir le *looseness factor* (« l »), le score minimal (« c ») et le nombre d'écart-type (« e ») (Figure 10), le processus de *matching* (association des déplacements chimiques) peut commencer pour chaque molécule comprise dans le SDF.



```
Anaconda Prompt

(base) C:\Users\Antoine> activate my-rdkit-env

(my-rdkit-env) C:\Users\Antoine> cd desktop

(my-rdkit-env) C:\Users\Antoine\Desktop> python derepcrude.py -i sdf_file.sdf -a
carbon_spectrum.txt -o output_filename.txt -c minimum_score -l looseness_factor
-e standard_deviation
```

Figure 10: Lancement de l'algorithme DerepCrude.

Comme illustré par la Figure 11, les listes de δ_c du spectre et de chaque molécule de la base de données sont d'abord triées par ordre croissant. La différence absolue entre la valeur du premier δ_c de chaque liste est calculée, c'est-à-dire $\delta_{C1-SDF-M}$ (déplacement chimique du 1^{er} carbone de la liste pour la molécule du SDF) et δ_{C1-13C} (déplacement chimique du 1^{er} carbone du spectre ^{13}C).

- Si cette différence est égale ou inférieure à la valeur du *looseness factor* « l » définie par l'utilisateur, les valeurs $\delta_{C1-SDF-M}$ et δ_{C1-13C} et l'intensité correspondante à δ_{C1-13C} , I_{C1-13C} , sont stockées. Le programme passe à la valeur suivante de la molécule $\delta_{C2-SDF-M}$ et calcule la différence absolue avec la valeur suivante du spectre δ_{C2-13C} , recommençant ainsi un nouveau cycle.

- Si cette différence est supérieure à la valeur du *looseness factor* « l », le programme cherche à comparer la valeur suivante du spectre δ_{C2-13C} avec la même valeur de la molécule que précédemment $\delta_{C1-SDF-M}$, et continue la boucle. Si de plus, $\delta_{C1-SDF-M} - \delta_{C1-13C} < l$, l'algorithme passe directement au $\delta_{C2-SDF-M}$ suivant. En effet, les déplacements chimiques étant classés par ordre décroissant, si δ_{C1-13C} dépasse $\delta_{C1-SDF-M}$ de plus d'une fois l, alors il n'y aura pas de solution qui satisferont les conditions imposées par la marge l plus loin dans la liste de δ_{C1-13C} . Par exemple : l = 1,0 ppm, $\delta_{C1-SDF-M}$ = 10,0 ppm et δ_{C1-13C} = 8,0 ppm. On a bien $|\delta_{C1-SDF-M} - \delta_{C1-13C}| > l$, soit |10,0 -

$8,0| > 1,0$. Les signaux ne peuvent pas être associés. $\delta_{C1-SDF-M} - \delta_{C1-13C} = 2,0$, ce qui est bien supérieur au *looseness factor*. L'algorithme passe donc au δ_{C-13C} suivant. Maintenant, $I = 1,0$ ppm, $\delta_{C1-SDF-M} = 10,0$ ppm et $\delta_{C2-13C} = 12,0$ ppm. On a toujours $|\delta_{C1-SDF-M} - \delta_{C2-13C}| > I$, soit $|10,0-12,0| > 1,0$. Mais cette fois, $\delta_{C1-SDF-M} - \delta_{C2-13C} = -2,0$, ce qui est inférieur à I . Il n'est donc pas nécessaire de continuer d'explorer la liste des δ_{C-13C} puisque toutes les valeurs suivantes seront supérieures à 12,0 ppm, et donc impossible à associer avec 10,0 ppm. L'algorithme passe donc au $\delta_{C2-SDF-M}$.

Dès que deux signaux sont associés, le programme examine si deux déplacements consécutifs de la molécule sont équivalents ($\delta_{Cx-SDF-M} = \delta_{Cx+1-SDF-M}$). Si c'est le cas, le déplacement chimique « unique » du spectre sera alors associé 2 fois : au lieu d'avancer vers le carbone suivant, le programme réutilise le même déplacement que précédemment ($\delta_{Cy-13C} = \delta_{Cy-1-13C}$).

A partir du 3^{ème} déplacement chimique associé, le programme applique le filtre d'intensité : les couples de signaux qui seront dorénavant associés ne seront considérés valables que si l'intensité des déplacements chimiques « en cours de *matching* » est comprise dans un intervalle de σ écart-type autour de la moyenne des intensités des déplacements chimiques déjà associés. Dans le cas contraire, ils seront rejetés.

Lorsque toutes les associations possibles ont été réalisés entre les deux listes de déplacements chimiques, celle de la molécule du SDF et celle du spectre, un score est attribué à la molécule. Ce dernier correspond au nombre de δ_c qui ont été associés sur le nombre de δ_c total de la molécule. L'intégralité du processus de *matching* est répétée pour chaque molécule du SDF jusqu'à ce que le programme atteigne la fin de liste.

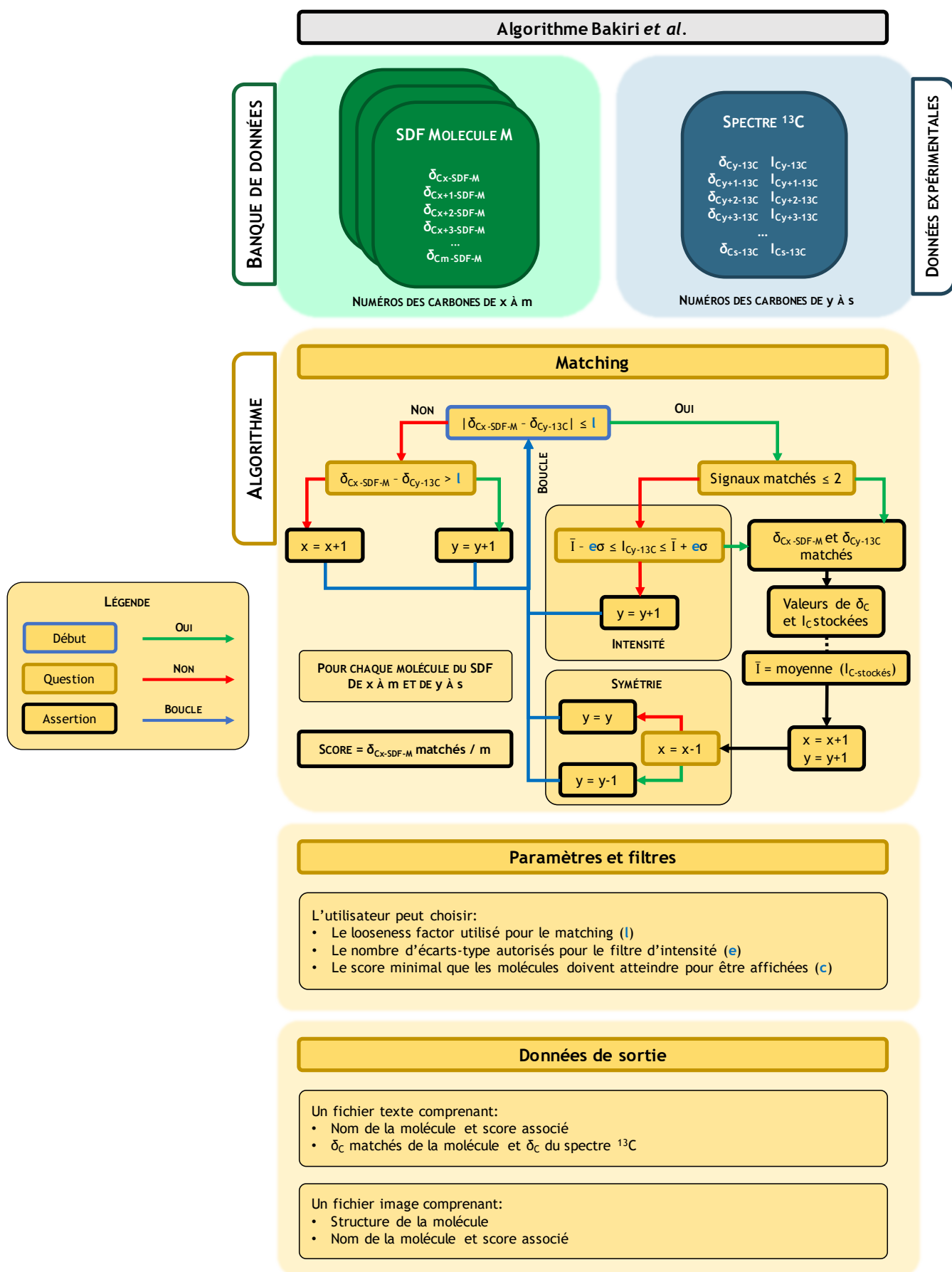
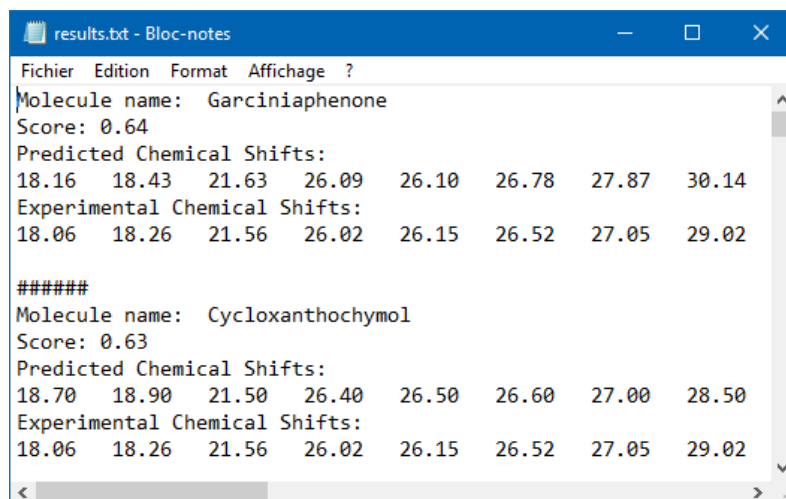


Figure 11: Fonctionnement de l'algorithme DerepCrude.

Les résultats sont présentés sous la forme de 2 fichiers, qui font apparaître les molécules par scores décroissants, si ces derniers sont supérieurs à la valeur minimale **c** définie par l'utilisateur. Tout d'abord un fichier texte reprend, pour chaque molécule, son nom, son score et la liste des déplacements chimiques qui ont été associés (**Figure 12**). Ensuite, un module graphique permet une illustration des structures moléculaires, de leurs noms et de leurs scores (**Figure 13**).



```

Fichier Edition Format Affichage ?
Molecule name: Garciniaphenone
Score: 0.64
Predicted Chemical Shifts:
18.16 18.43 21.63 26.09 26.10 26.78 27.87 30.14
Experimental Chemical Shifts:
18.06 18.26 21.56 26.02 26.15 26.52 27.05 29.02

#####
Molecule name: Cycloxanthochymol
Score: 0.63
Predicted Chemical Shifts:
18.70 18.90 21.50 26.40 26.50 26.60 27.00 28.50
Experimental Chemical Shifts:
18.06 18.26 21.56 26.02 26.15 26.52 27.05 29.02
  
```

Figure 12: Fichier texte (.txt) de résultats générés par l'algorithme DerepCrude.

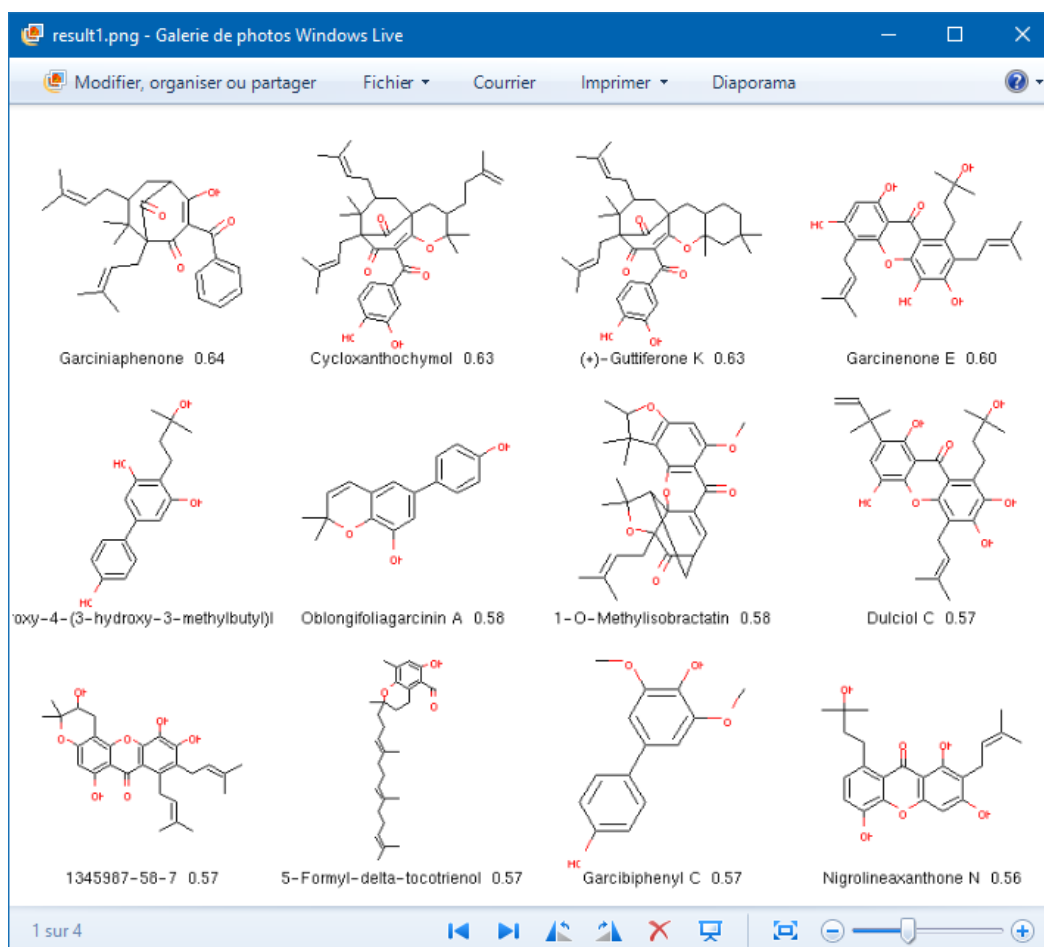


Figure 13: Fichier image (.png) de résultats de l'algorithme DerepCrude.

2.4. Recherche d'améliorations algorithmiques

A partir des observations effectuées sur les algorithmes existants, dans une optique d'optimisation des résultats, nous avons cherché à comprendre quels paramètres pouvaient être implémentés à un nouvel algorithme afin d'améliorer son potentiel discriminant. L'idée principale est d'introduire une sélection des carbones selon leur nombre de liaisons : carbones primaires, secondaires, tertiaires ou quaternaires. Il serait en effet possible d'obtenir cette information sur les données expérimentales à partir de 2 spectres DEPT (135 et 90), et la base de données pourra être reclassée en fonction du nombre de liaison de chaque carbone. Ce travail sera détaillé dans les paragraphes suivants.

2.4.1. Exploitation des données issues des expériences DEPT

Le premier, et principal paramètre, a été l'utilisation des données obtenues lors de l'utilisation de séquences **DEPT** (*Distortionless Enhancement by Polarization Transfer*) en complément des valeurs des δ_c . En effet, les expériences DEPT permettent de déduire la multiplicité des carbones en se basant sur le phasage des signaux (**Figure 14**). En DEPT 135, les signaux des CH et CH₃ apparaîtront « positifs » (phase), ceux des CH₂ « négatifs » (antiphase) tandis que ceux des C quaternaires, ne bénéficiant pas du transfert de polarisation $^1\text{H} \rightarrow ^{13}\text{C}$ seront d'intensité si faible qu'ils seront généralement confondu avec le bruit de fond et donc « absents » du spectre. En DEPT 90, seuls les signaux des CH apparaîtront, phasés positivement, sur le spectre. La combinaison de ces informations permet ainsi de créer des couples déplacement chimique/multiplicité du carbone.

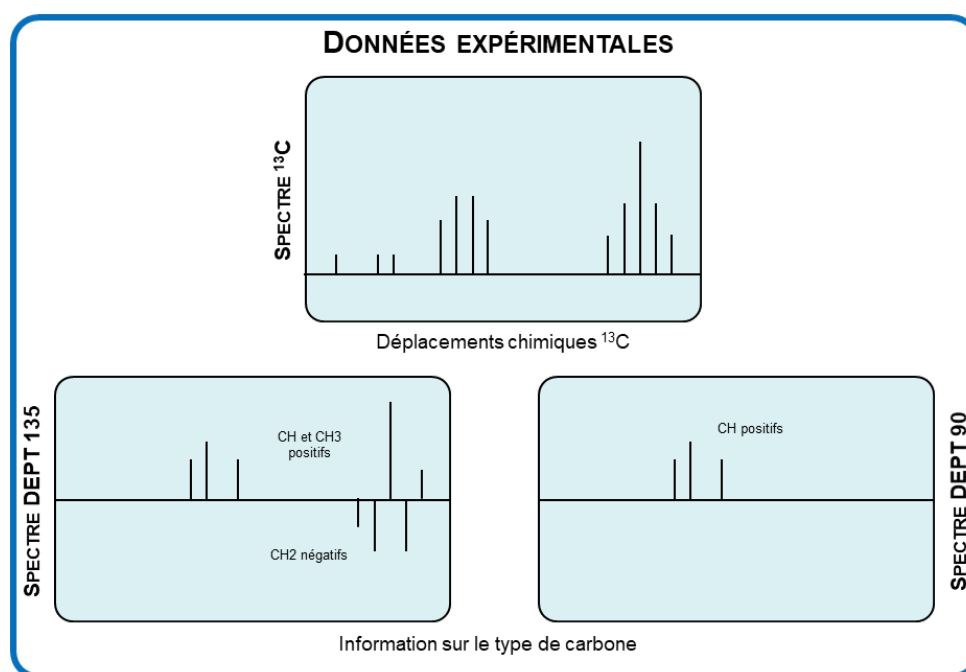


Figure 14: Information sur la multiplicité des atomes de carbone déduite des expériences de RMN-¹³C et DEPT 135 & 90.

Cette double information permet en effet une analyse beaucoup plus discriminante qu'une simple analyse des δ_c . Afin de d'évaluer cet effet discriminant, la banque de données *Garcinia*, comprenant les 718 produits naturels reportées dans le genre *Garcinia* d'après le DNP [18], a été analysée pour visualiser les « combinaisons » de type

de carbones spécifiques à chaque molécule. C'est-à-dire que pour chaque molécule de la base de données, le nombre de C quaternaire, CH₂ et CH+CH₃ a été répertorié en tableau (**Tableau 2**). Dans ce tableau, les combinaisons les plus fréquentes ont ensuite été triées par ordre décroissant. Par exemple, on observe que, pour les produits naturels isolés des *Garcinia*, la combinaison la plus fréquente est « 13 carbones quaternaires, 2 carbones secondaires et 9 carbones tertiaires ou primaires » et qu'elle n'est partagée que par 21 molécules ne représentant qu'environ 3% de la base de données.

Tableau 2: Nombre de produits naturels de la base de données *Garcinia* partageant la même combinaison de [C, CH₂, CH₃+CH].

Cq	CH ₂	CH + CH ₃	Nb de molécules
13	2	9	21
15	3	10	16
11	1	7	15
15	6	17	15
12	2	9	13
13	2	8	12
16	6	16	12
13	1	10	10
13	1	9	10
15	2	11	10
12	1	10	9
15	5	18	9
8	0	5	9
11	1	9	8
12	3	8	8

Cq	CH ₂	CH + CH ₃	Nb de molécules
12	3	9	8
13	2	10	8
14	4	10	8
16	4	13	8
17	6	15	8
11	0	7	7
11	1	6	7
11	4	15	7
13	3	7	7
14	4	11	7
15	1	12	7
15	8	15	7
16	8	19	7
10	0	8	6
...

L'analyse plus approfondie de ces données a permis de trouver à quelle fréquence un groupe de taille particulière était retrouvé. Le **Tableau 3** montre ainsi que, dans 164 cas (c'est-à-dire dans plus de 50% des cas), 1 combinaison particulière mène à 1 seule possibilité de molécule dans la base de données, et que dans 49 cas, une combinaison mène à 2 possibilités de structures. Donc, dans plus de 70% des cas, une combinaison particulière de C, CH₂ et CH₃+CH n'oriente que vers 1 ou 2 molécules en particulier.

Tableau 3: Nombre de fois (colonne de gauche) où une combinaison de [C, CH₂, CH₃+CH] amène à un groupe dont la taille est indiquée de la colonne de droite.

Nb d'occurrence	Taille du groupe
164	1
49	2
24	3
18	4
10	5
8	7
7	6
7	8
3	9
3	10
2	12
2	15
1	13
1	16
1	21

Cela semble bien confirmer que la réalisation additionnelle d'une expérience de type DEPT 135 (CH₃ et CH étant ici indifférenciés), permet de limiter drastiquement les possibilités de structures et donc d'arriver rapidement aux molécules d'intérêt. La différenciation supplémentaire des carbones tertiaires et primaires, permise par l'expérience DEPT 90 (**Tableau 4**), augmente également le taux de discrimination, comme illustré par le **Tableau 5**. Dans 218 cas, soit environ 60% du total des possibilités, 1 combinaison correspond à 1 molécule particulière dans la base de données. Dans 66 cas, 1 combinaison peut correspondre à 2 molécules de cette même base. Donc, dans plus de 75% des cas, 1 combinaison de C_q, CH, CH₂, CH₃ oriente vers 1 ou 2 structures.

Tableau 4: Nombre de produits naturels de la base de données *Garcinia* partageant la même combinaison de [C_q, CH, CH₂, CH₃].

C _q	CH	CH ₂	CH ₃	Nb de molécules
13	4	2	5	19
15	8	6	9	15
11	4	1	3	15
15	4	3	6	14
13	4	2	4	12
12	5	2	4	10
13	5	1	5	10
12	6	1	4	9
15	5	2	6	9
16	7	6	9	9
8	5	0	0	9
13	5	1	4	9
14	5	4	5	8
13	4	2	6	8
15	7	8	8	7

C _q	CH	CH ₂	CH ₃	Nb de molécules
16	9	8	10	7
16	6	4	7	7
14	5	4	6	7
11	4	1	2	7
11	5	4	10	7
11	4	1	4	6
10	5	1	2	6
12	5	3	4	6
15	6	1	6	6
9	4	0	1	6
15	3	4	6	5
11	6	3	3	5
10	5	1	3	5
11	5	0	3	5
...

Tableau 5: Nombre de fois (colonne de gauche) où une combinaison de [Cq, CH, CH₂, CH₃] amène à un groupe dont la taille est indiquée de la colonne de droite.

Nb d'occurrence	Taille du groupe
218	1
66	2
27	3
13	4
12	5
6	7
5	6
5	9
2	8
2	10
2	15
1	12
1	14
1	19

2.4.2. Matching des déplacements chimiques des ¹³C

La seconde optimisation proposée se fait au niveau de l'association des déplacements chimiques. En effet, tous les algorithmes proposent une marge autorisée lors de la comparaison des signaux mais, la plupart d'entre eux, considèrent que le premier déplacement chimique compris dans cet intervalle est le déplacement correct qui sera donc associé. Il semblerait plus intéressant de faire en sorte que les déplacements chimiques soit associés avec les valeurs les plus proches possible, c'est-à-dire commencer par rechercher des *matches* « parfaits » (aucune différence entre les valeurs des δ_c) puis graduellement augmenter la marge jusqu'à atteindre la valeur maximale définie par l'utilisateur. Cette façon de procéder serait évidemment capitale lors de l'utilisation de bases de données expérimentales, dans lesquelles les différences observées entre valeurs sont généralement très faibles : de l'ordre de 0,1 à 0,2 ppm (lorsque les données sont enregistrées dans le même solvant deutéré). Cet avantage s'appliquera également aux bases associant données prédites et expérimentales. Cependant l'intérêt de cette méthode de calcul peut paraître amoindri pour une utilisation limitée aux données prédites mais l'expérience a montré la fiabilité des outils de prédictions, et qu'associer les déplacements les plus proches améliore les résultats (cf. **3.1. Huile essentielle de menthe poivrée**).

2.4.3. Utilisation de filtres

La possibilité de filtrer les résultats, comme le proposent déjà certains algorithmes, est un outil puissant permettant d'apporter autant d'informations que l'on en possède, et donc de réduire le champ des possibles. Proposer l'utilisation d'un filtre de masse moléculaire semblait adapté car des analyses SM sont assez fréquentes en phytochimie, et un grand nombre de chercheurs peuvent avoir accès à cette information. Cela permet aussi de sélectionner des produits naturels ayant une masse moléculaire comprise dans un intervalle donné si l'utilisateur a « une idée » de la composition de son mélange. Cela rejoint un peu le principe de « *minimal number of shifts to match* » proposé par l'algorithme d'ACD/Labs® [12], dans le sens où ces deux paramètres constituent un critère de taille de molécule, permettant ainsi de discriminer d'avantage les différents candidats entre eux.

2.4.4. Interactivité

La dernière amélioration possible réside dans « l'interactivité » de l'utilisateur avec les résultats. Les outils utilisés pour la déréplication ne présentent en effet en aucun cas des résultats absolus ou définitifs : il s'agit d'hypothèses, qui doivent être discutées et triées grâce aux connaissances phytochimiques et analytiques de l'utilisateur. C'est pour cela qu'il semble important que ce dernier puisse, à partir des résultats, avoir accès à un maximum d'informations sur la façon dont les signaux ont été associés, mais également faire rapidement le lien avec les données initiales, *i.e.* le spectre RMN et la base de données SDF. Non seulement l'accès à ces informations mais également la possible optimisation manuelle de ces dernières, d'une façon rapide et intuitive, a constitué un des objectifs de ce programme.

2.4.5. Intensité des signaux

Enfin, une discussion sur l'utilisation d'un filtre d'intensité nous est apparue essentielle. La mise en place d'un tel système devrait permettre de faire en sorte, en se basant sur l'intensité des signaux observés, que seuls les signaux appartenant à une même molécule, donc avec une intensité d'un même ordre de grandeur, soient associés ensemble. Cela permettrait théoriquement de réduire les possibilités de *matching* dans lesquelles des signaux issus de composés différents sont attribués à une même molécule. C'est ce que Bakiri *et al.* [6] ont proposé dans leur script, en calculant une moyenne d'intensité des signaux associés. Puis, considérant que toutes les intensités sortant de l'intervalle de ± 2 écart-types appartenaient à des molécules distinctes, leur algorithme excluait ces composés. Cependant, après utilisation de cet algorithme, il nous est apparu que supprimer ce filtre d'intensité aboutissait à de meilleurs résultats (cf. **Annexe 1.1. Le filtre d'intensité de DerepCrude**). Après quelques essais infructueux de mise en place d'un filtre d'intensité fonctionnel (cf. **Annexes 1.4. Essais de clustering avec DBSCAN**), nous avons décidé de faire en sorte que l'évaluation de l'homogénéité des intensités des signaux associés soit laissée aux soins de l'utilisateur, en lui procurant, au niveau de l'interface graphique, une visualisation directe des signaux associés et de leurs intensités respectives.

2.5. Description de l'algorithme MixONat

Compte-tenu des observations précédentes, nous avons développé un nouveau programme. Puis avec l'aide de Frédéric Saubion (Laboratoire LERIA, EA2645, UNIV Angers, SFR MathSTIC, Faculté des Sciences) et d'étudiants en master 2 informatique 1) le code que nous avons créé a été optimisé afin de pouvoir y associer une interface graphique, 2) des fonctions permettant l'amélioration des résultats post-*matching* ont été ajoutées et 3) des scripts permettant une installation et un lancement du programme facilité ont été créés. Le programme a été baptisé **MixONat**, pour **Mixture Of Natural** products dont la version 1.0 est présentée dans les paragraphes suivants.

L'interface graphique a été implémentée afin de faciliter l'utilisation de MixONat, la manipulation des différents paramètres, et la visualisation des résultats lors de l'utilisation routinière du logiciel. Lors de l'installation, un script multi-plateformes permet la création d'un environnement virtuel sous Conda [22] ainsi que le téléchargement et la mise en place de toutes les dépendances du programme, à savoir RDKit [23] gérant les lectures de fichier SDF et les tâches associées aux dessins de structures de molécules, Matplotlib [24] permettant le tracé de graphiques, et Kivy [25] responsable de l'affichage de l'interface graphique créée.

La fenêtre qui s'affiche au lancement du programme présente aujourd'hui 3 onglets (**Figure 15**) : *Inputs* (Données), *Parameters* (Paramètres) et CTypeGen (Générateur de base de données intégrant les δ_c triés par type de carbone).

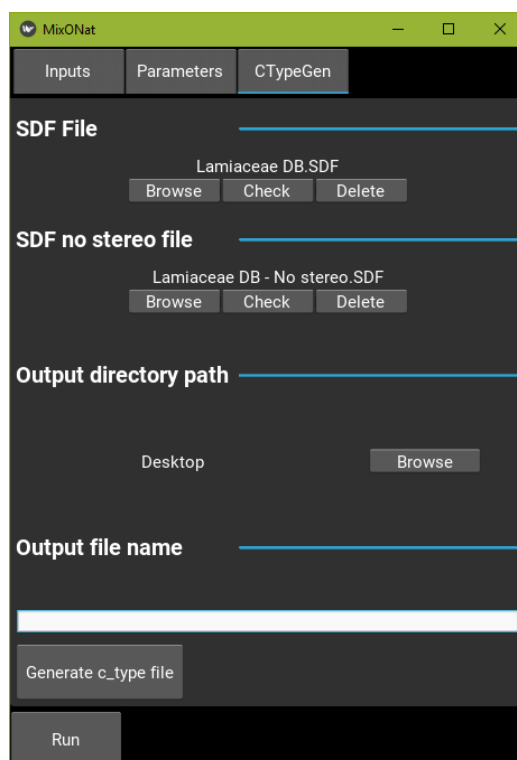


Figure 15: Onglet CTypeGen permettant de trier les déplacements chimiques de chaque molécule de la base de données SDF en fonction du type de carbone.

2.5.1. Génération des bases de données : onglet CTypeGen

L'onglet CTypeGen (**Figure 15**) permet la modification des bases de données originales fournies au format SDF afin que les déplacements chimiques ^{13}C qu'elles contiennent soient triés selon qu'il s'agit de carbones quaternaires, de méthines, de méthylènes ou de méthyles. Les données doivent être correctement triées pour être utilisables par le programme. Comme décrit précédemment, les bases de données sont créées à partir d'ensembles de fichier de structures, les fichiers .mol, qui peuvent également contenir un certain nombre d'informations (nom, formule brute, etc...). Ces ensembles de fichiers .mol, appelés SDF, doivent également contenir l'information concernant les valeurs des déplacements chimiques des carbones des molécules, prédites par un logiciel ou entrées manuellement, avant d'entamer l'étape de transformation.

Afin de mieux comprendre comment s'effectue cette transformation des bases de données, l'anatomie d'un SDF - dans laquelle le code de fichier .mol correspondant au benzaldéhyde est explicité - est présentée en **Figure 16**. Le fichier commence par indiquer le nom du logiciel utilisé pour sa création : ici ACD/Labs® [12] a été utilisé. Le premier ensemble de ligne qui va suivre s'appelle le *Mol block*, qui est en fait la description minimale nécessaire au bon fonctionnement d'un fichier .mol. Ce bloc est composé d'une ligne d'introduction, et de deux blocs (*atom block* et *bound block*). La première ligne d'introduction comporte 2 numéros. Le premier indique le nombre de lignes dans le bloc des atomes (ou *atom block*), situé juste dessous, ce qui correspond généralement au nombre

d'atomes de la molécule qui ne sont pas des hydrogènes (à l'exception de certains hydrogènes explicites, généralement associés à une information stéréochimique, qui peuvent apparaître). Le second numéro indique le nombre de ligne dans le bloc des liaisons (ou *bound block*), situé sous le bloc précédent, correspondant ici au nombre de liaisons entre les différents atomes non-hydrogènes (avec la même exception que précédemment citée).

Le bloc des atomes décrit les atomes ligne par ligne, la première ligne correspondant à l'atome numéro 1 et ainsi de suite². Les coordonnées spatiales X, Y et Z sont indiquées en première, seconde et troisième colonne ; la quatrième colonne présentant le symbole de l'atome. Ici, la structure étant en 2D, la troisième dimension n'est pas renseignée. Pour prendre un exemple, on remarque que la ligne 2 est l'origine (0,0,0) et correspond à un atome d'oxygène. Les atomes 5 et 7 possèdent la même valeur que 2 pour leur coordonnée selon X. C'est bien ce qui est représenté sur la structure du benzaldéhyde dans l'encadré : l'atome d'oxygène portant le numéro 2 est bien aligné selon X (verticalement) avec les deux atomes de carbone 5 et 7. On peut faire cette analyse pour d'autres couples d'atome comme 5 et 4, qui partagent la même coordonnée selon Y (horizontalement).

Le bloc des liaisons, situé sous le bloc des atomes, décrit ligne par ligne les liaisons inter atomiques (les liaisons pouvant être simples, doubles ou triples). La première colonne présente le numéro du premier atome de la liaison, la seconde colonne présente son partenaire, et la troisième colonne indique le nombre de liaison entre ces deux atomes. La première ligne, on lit 1 2 2, ce qui peut se traduire par l'atome 1 et l'atome 2 sont reliés par 2 liaisons. Cela est facilement vérifié sur la structure de la molécule puisqu'il s'agit ici de la liaison aldéhydique. La seconde ligne indique que l'atome 1 est également lié à l'atome 3 par 1 liaison. On peut ainsi en déduire le nombre total de liaisons de chaque atome, et donc leur multiplicité. C'est donc grâce à l'analyse automatisée de ce bloc que le tri par type de carbone effectué par CTypeGen sera possible. Enfin, la quatrième colonne dans le bloc des liaisons n'indique ici que des 0. Cela est signe d'une structure totalement plane (ou du moins, dessinée de façon planaire). Si des informations de stéréochimie sont dessinées, c'est ici qu'elles seront stockées en indiquant que la liaison définie sur la ligne est soit en avant (symbolisé par un 1) soit en arrière du plan (symbolisé par un 6).

Enfin, le *Mol block* se termine par « M END ».

² Attention à certains problèmes de numérotation qui peuvent apparaître lorsque plusieurs logiciels (notamment de dessin de molécules) sont utilisés consécutivement sur la même structure. La numérotation du SDF et celle apparaissant dans le logiciel pourront être différentes.

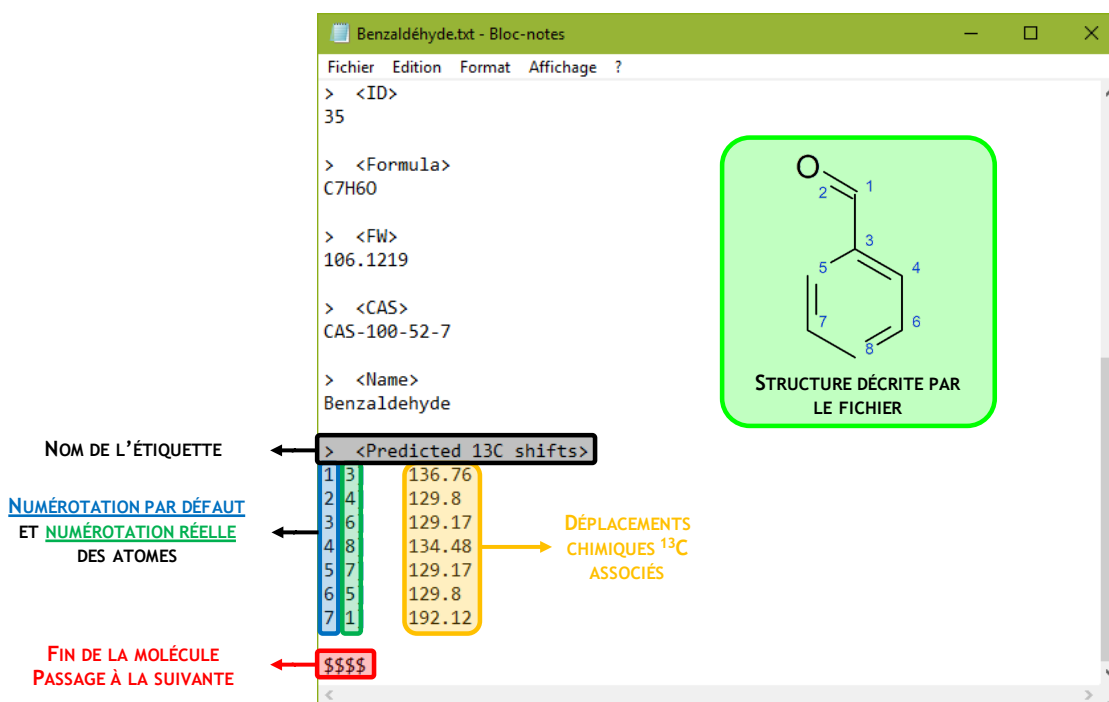


Figure 17: Anatomie d'un SDF : les étiquettes du benzaldéhyde.

Cependant, la liste des déplacements chimiques comporte uniquement le numéro d'atome de carbone et son δ_c associé. Le travail du script CTypeGen va donc être de classer chaque carbone en fonction de sa multiplicité. Le fichier contenant des informations sur la structure de la molécule au niveau du *Mol Block*, on peut - comme décrit dans le paragraphe précédent - savoir si un carbone est quaternaire, tertiaire, secondaire ou primaire en comptant le nombre de liaisons de l'atome concerné. A partir du SDF initial, en se basant sur la liste des déplacements chimiques originale présente sous l'étiquette « *Predicted 13C shifts* », ce programme va *in fine* créer un nouvel SDF dans lequel 4 listes vont être créées pour chaque molécule, soit une nouvelle étiquette pour chaque différent type de carbone (4 possibilités) répertoriant ses numéros et déplacements chimiques (Figure 19). La transformation est montrée de façon plus concrète en Figure 18 toujours à partir du benzaldéhyde du benzaldéhyde. On constate que 4 nouvelles étiquettes ont été créées après l'étiquette « ID », « *Quaternaries* », « *Tertiaries* », « *Secondaries* » et « *Primaries* », et que les différents atomes et leur déplacements correspondants ont été triés en fonction de leur « parité ». Même si la molécule ne comporte pas un certain type d'atomes, comme les secondaires et les primaires dans le cas du benzaldéhyde, l'étiquette sera tout de même créée, mais restera vide d'information.

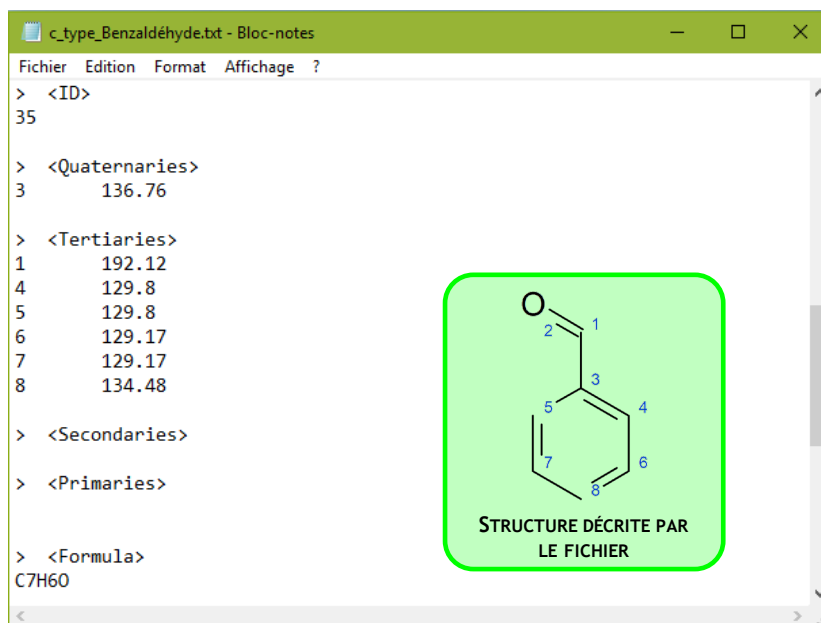


Figure 18: Extrait des étiquettes du benzaldéhyde après transformation par CTypeGen.

Ce traitement par CTypeGen permettra ensuite que le programme de *matching* traite l'information en fonction de la multiplicité (déduite des expériences DEPT) des atomes. Cette étape est bien sûr uniquement nécessaire une fois, *i. e.* lors de la création d'une nouvelle base de données (SDF) de déplacements chimiques triés par type de carbone. Elle évite de refaire un tri par type de carbone à chaque utilisation du logiciel, réduisant ainsi le temps de calcul nécessaire. Une fois établi le fichier « c_type » pourra ensuite être réutilisé à volonté.

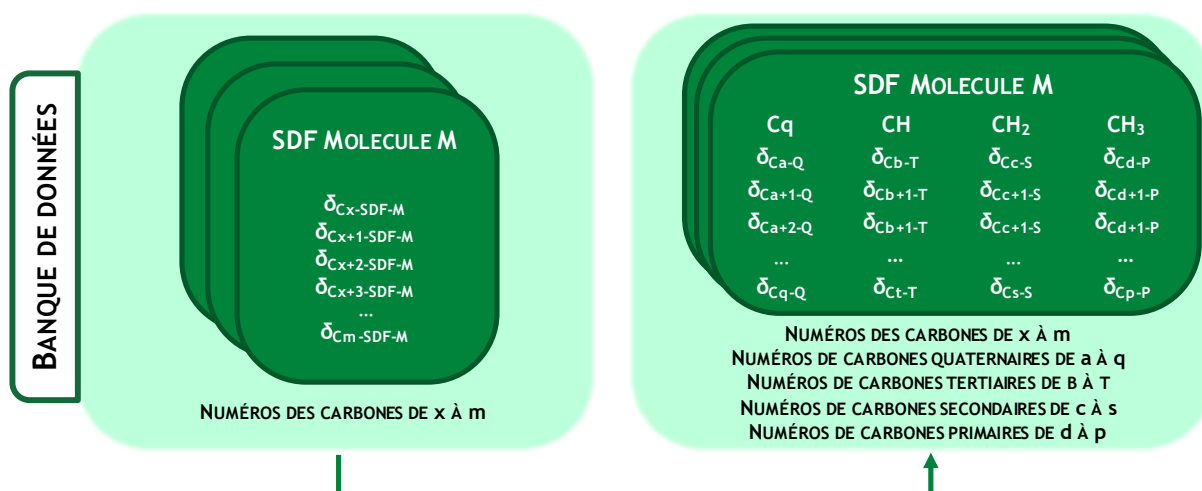


Figure 19: Tri des bases de données par type de carbone par le premier script CTypeGen.

Il convient de noter qu'afin de fonctionner correctement, le programme nécessite l'apport du SDF à retravailler/reclasser mais également d'un SDF de contenu identique à ceci près que les éléments relatifs à la stéréochimie des molécules en ont été supprimés (SDF *no stereo file*). Il est en effet très important que l'information stéréochimique soit absente (c'est-à-dire, au niveau informatique, que la 4^{ème} colonne du *bound block* ne comporte que des 0) avant le tri par type de carbone ne s'opère, car elle peut induire des erreurs lors de la lecture de la multiplicité de chaque atome, résultant en un tri erroné.

Le programme est pour l'instant optimisé pour fonctionner avec des bases de données créées avec ACD/Labs® [12] et avec des étiquettes nommées de façon spécifique. Il est donc possible qu'il ne fonctionne pas correctement sur des fichiers de bases de données au format SDF produites au moyen d'autres supports.

2.5.2. Matching : onglets *Inputs* et *Parameters*

Les deux autres onglets, i.e. *Inputs* et *Parameters* sont liés au processus de déréplication en lui-même.

L'onglet **Inputs** permet le chargement des fichiers d'entrée dans le programme, c'est-à-dire ceux nécessaires au bon fonctionnement de celui-ci (Figure 20). Au moins une base de données (.SDF) dont les δ_c des molécules sont triés par type de carbone et des données issues d'un spectre ^{13}C -RMN sont nécessaires. L'utilisateur peut également ajouter des données DEPT 135 et DEPT 90 s'il en dispose. Les données spectrales sont fournies sous forme de tableau .csv des déplacements chimiques carbones et de leur intensité respective (Figure 21).

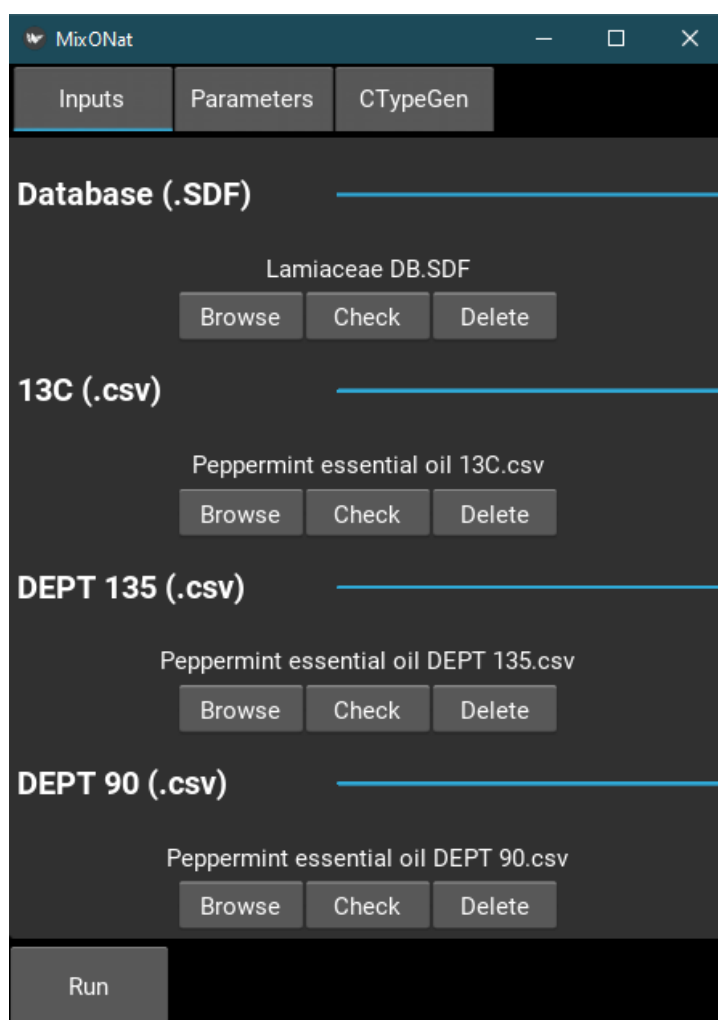


Figure 20: Onglet *Inputs* permettant de charger via un explorateur les différents fichiers d'entrée dans le programme. Base de données (SDF) et spectre 13C (.csv) sont obligatoires, DEPT 135 et DEPT 90 (.csv) optionnels.

	A	B	C	D	E	F	G
1	214.84,0.000478						
2	212.62,0.003509						
3	170.84,0.000577						
4	136.81,0.000560						
5	133.84,0.000280						
6	120.73,0.000728						
7	108.46,0.000895						
8	74.25,0.001400						
9	73.74,0.000811						
10	71.59,0.010642						

Peppermint essential oil 13C

Figure 21: Exemple d'un fichier .csv (*comma-separated values*) d'un spectre RMN-13C généré avec Microsoft Excel. Le déplacement chimique et son intensité sont séparés par une virgule.

Le dernier onglet **Parameters** (Figure 22) permet de sélectionner tous les paramètres nécessaires à la recherche effectuée :

- La marge autorisée (ou *tolerance*) ϵ , qui est choisie selon la précision de la banque de données utilisée, *i.e.* selon que les δ_c sont réels ou prédits. Lors de la comparaison des signaux expérimentaux (δ_{13C}) et de ceux contenus dans le fichier SDF (δ_{SDF}) pour chaque molécule, le logiciel considérera que δ_{13C} peut être associé à δ_{SDF} si $\delta_{SDF} - \epsilon < \delta_{13C} < \delta_{SDF} + \epsilon$. Le travail précédemment cité [16] sur la précision des bases de données créées avec ACD/Labs® [12] nous permettent de choisir la valeur de ce paramètre tel que $\epsilon = 1,3$ ppm s'il s'agit d'une base de données de déplacements chimiques prédits. Elle peut cependant être ajustée facilement si une base de données différente est utilisée, soit en tapant directement la valeur souhaitée, soit en augmentant ou diminuant la valeur grâce à des boutons + et -.

- L'incrémentation de ϵ peut être activée ou désactivée grâce à une case à cocher. Si activée, le programme cherchera d'abord à associer les différents déplacements chimiques en considérant que $\epsilon = 0,0$ ppm puis, augmentera cette valeur de façon incrémentale par paliers de 0,1 ppm jusqu'à atteindre la valeur ϵ fixée par l'utilisateur. Cela permet de faire en sorte que le processus de *matching* associe d'abord les déplacements chimiques les plus proches ensembles. Ce paramètre peut néanmoins être désactivé. Dans ce cas, l'algorithme associera un δ_{SDF} avec le premier δ_{13C} compris dans un intervalle de $\pm \epsilon$. Ce paramètre est activé par défaut et il est conseillé de le laisser comme tel, surtout lors de l'utilisation de bases de données expérimentales.

- Le facteur d'alignement, ou *DEPT alignment*, des données DEPT (135 et 90) permet d'associer un δ_{13C} au signal lui correspondant dans le spectre DEPT 135 ou DEPT 90. En effet, les déplacements chimiques pouvant être légèrement décalés d'un spectre à l'autre, et non nécessairement de façon uniforme entre DEPT 135 et DEPT 90, l'utilisateur doit renseigner cette valeur d'alignement. Par défaut, elle est basée à 0,02 ppm et se fait automatiquement, de façon incrémentale (de 0,00 ppm jusqu'à la valeur sélectionnée). La qualité des résultats est dépendante de ce facteur d'alignement puisqu'il va permettre de déduire la multiplicité des atomes.

- Un paramètre « carbones équivalents » peut être activé ou désactivé. Si activé, il permet au programme d'utiliser autant de fois un δ_{13C} que la valeur du δ_{SDF} auquel il est associé est présente dans la liste de déplacements chimiques de la molécule du SDF. S'il est désactivé, un δ_{13C} ne peut être utilisé qu'une seule fois et ce, peu importe le nombre de δ_{SDF} équivalents. Ce qu'il faut avoir à l'esprit en utilisant ce paramètre est d'abord que si l'on travaille avec des bases de données prédites, l'équivalence d'un ou plusieurs carbones est aussi prédite et donc, peut être incorrecte : des signaux prédits comme équivalents peuvent ne pas l'être en réalité et inversement. La seconde chose à laquelle il faut penser est que l'autorisation des « carbones équivalents » peut favoriser des molécules plus volumineuses comme des multimères, des molécules dotées de symétrie, mais également des longues chaînes carbonées dans lesquelles un seul signal sera utilisé une multitude de fois, et donc dont le score sera artificiellement augmenté.

- Un filtre de masse moléculaire permet de choisir soit un ou plusieurs masses moléculaires précises, soit un intervalle de masse (± 1 g/mol). Seules les molécules répondant à ces critères seront traitées et triées dans les résultats. Ce filtre est optionnel.

- L'utilisateur peut également choisir combien de résultats lui seront présentés, afin d'éviter que l'intégralité de la base de données n'apparaisse. De la même façon, il peut choisir le nombre de molécules qui s'afficheront par page de résultats lors de la sauvegarde de la recherche.

- Grâce à un explorateur de fichiers, il est finalement possible de choisir l'endroit où les résultats seront sauvegardés ainsi que leur nom de fichier.

MixONat

Inputs Parameters CTypeGen

Tolerance (ppm) - 1.3 +

Tolerance incrementation ☒

DEPT 135 alignment - 0.02 +

DEPT 90 alignment - 0.02 +

Equivalent carbons ☐

Molecular weight ☐ Specific values ☒ MW1; MW2; MW3...
☐ Interval ☐

Number of results - 50 + SDF file size : 980

Number of results per page - 25 +

Results directory path Select a path Browse

Run

Figure 22: Onglet *Parameters* permettant de modifier les paramètres du programme. Des valeurs par défauts sont proposées à l'utilisateur qui peut choisir de les modifier en fonction des informations qu'il possède.

2.6. Fonctionnement du *matching* de l'algorithme MixONat

Le programme de *matching* commence par trier les carbones du spectre RMN ^{13}C en fonction de leur multiplicité. Le tri se fait en fonction des données DEPT fournies par l'utilisateur et prend en compte le facteur d'alignement DEPT choisi. S'il n'y a pas de fichier DEPT, les carbones restent indifférenciés. L'ajout d'un DEPT 135 permet de différencier les carbones secondaires, ayant une intensité négative, des carbones tertiaires et primaires, ayant une intensité positive. Les carbones présents dans le spectre ^{13}C initial mais absent du spectre DEPT 135 sont considérés comme quaternaires. Si les données d'un spectre DEPT 90 sont également fournies, il est alors possible de différencier carbones tertiaires et primaires. Le programme commence d'abord par séparer les carbones secondaires, « négatifs » dans le DEPT 135, puis les carbones tertiaires, présents dans le DEPT 90. Enfin, les carbones primaires sont les carbones « positifs » du DEPT 135, absents du DEPT 90, et les carbones quaternaires sont les carbones restant du spectre RMN- ^{13}C initial (Figure 23).

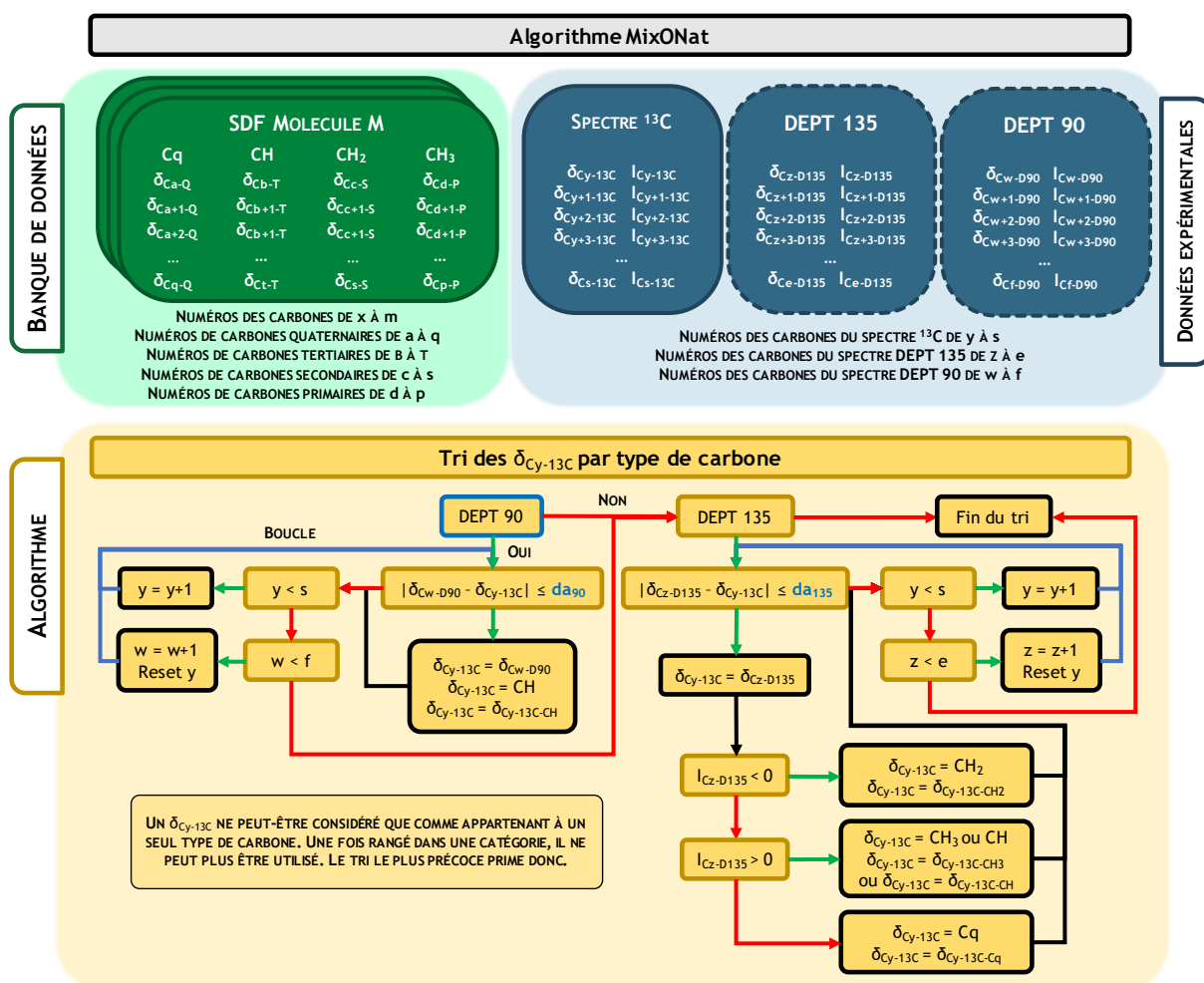


Figure 23: Fonctionnement de l'algorithme MixONat : données d'entrée et tri par type de carbone.

Une fois le processus de tri terminé, le *matching* commence : les associations se font par type de carbone. C'est-à-dire que les δ_c considérés comme ceux de carbones quaternaires dans le spectre ne peuvent être associés qu'avec des δ_c quaternaires du SDF. Les listes de déplacements chimiques expérimentales et celles du SDF sont triées par ordre croissant avant tout autre étape. Commenant par la première molécule du SDF (ID 1),

Bruguère Antoine | Mise au point d'une méthode d'analyse
dérépliquative par RMN du carbone 13 86

considérant son premier δ_{SDF} , le programme cherche un $\delta_{13\text{C}}$ qui satisfait la marge autorisée, paramétrée par l'utilisateur. En fonction du choix de ce dernier, la recherche peut se faire de façon incrémentale, et donc chercher préférentiellement les correspondances les plus proches de la valeur prédite ou expérimentale. De même, un $\delta_{13\text{C}}$ peut être utilisé plusieurs fois ou non en fonction du paramètre « carbones équivalents » renseigné. Quand tous les δ_{SDF} de la molécule 1 ont été passés en revue, un score et l'écart aux valeurs attendues (*vide infra*) sont calculés pour la molécule et l'information est stockée.

Le score correspond au rapport entre le nombre de δ_{C} associés pour une molécule et le nombre total de δ_{C} que comporte cette molécule. L'écart (ou *deviation*) représente la différence absolue cumulée entre signaux associés, c'est-à-dire $|\delta_{\text{SDF}} - \delta_{13\text{C}}|$ pour chacun des signaux associés. L'algorithme répète ces opérations pour chaque molécule du SDF qui correspond aux paramètres de masse moléculaire éventuellement sélectionnés par l'utilisateur (**Figure 24**).

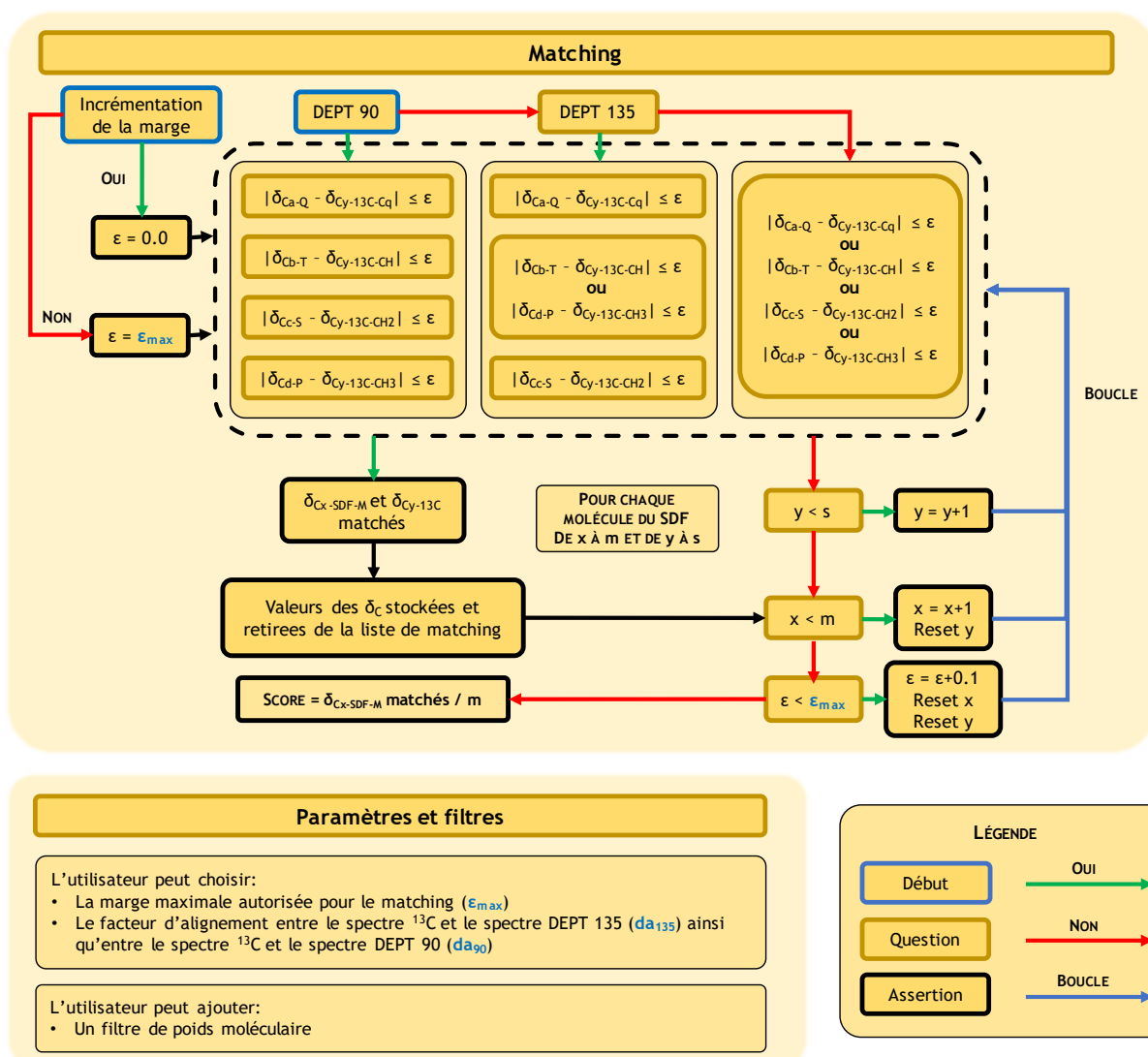


Figure 24: Fonctionnement de l'algorithme MixONat : processus de *matching* et paramètres.

Avant de présenter les résultats finaux, l'algorithme procède à une correction locale des associations. En effet, cet algorithme est dit « glouton », c'est-à-dire qu'il cherche à remplir les conditions imposées à un niveau local,

sans « vision globale » du process. Le glouton ne reviendra ainsi pas sur sa décision afin d’optimiser les résultats. En effet, il arrive que des déplacements chimiques soit associés car ils répondent à toutes les conditions fixées par l’utilisateur sans que les associations soient optimisées. Cela augmente alors artificiellement le taux d’écart sur une molécule. Voici un exemple concret permettant d’appréhender cette problématique (**Figure 25**). Le ϵ est fixé à 1,30 ppm : les δ_{13C} du spectre 111,24 et 111,92 ont été respectivement associés aux δ_{SDF} 111,57 et 110,67 d’une molécule du SDF. L’algorithme a bien associé en priorité les déplacements chimiques qui étaient les plus proches $|111,24 - 111,57| = 0,33$ alors que $|111,92 - 111,57| = 0,35$. Cependant, si on regarde la différence absolue cumulée (ou *deviation*) associée à ce *matching*, on se rend compte que $|111,24 - 111,57| + |111,92 - 110,67| = 1,58$ ppm, alors que si les signaux associés avaient été inversés, $|111,92 - 111,57| + |111,24 - 110,67| = 0,92$ ppm. Dans ce cas, dans une même fenêtre ϵ choisie, le *matching* des signaux n’était pas optimal. Le but de cet algorithme de correction locale des associations est donc le suivant : par groupes de signaux, dans une même fenêtre de ϵ ppm, il va chercher à optimiser les valeurs associées afin de réduire la différence absolue cumulée au maximum. Le score de la molécule ne sera pas modifié, mais l’écart le sera, ce qui permettra à plusieurs molécules possédant le même score d’être différenciées.

δ_c SDF	δ_c Spectre ^{13}C	Ecart	δ_c SDF	δ_c Spectre ^{13}C	Ecart
111,57	111,24	0,33	111,57	111,92	0,35
110,67	111,92	1,25	110,67	111,24	0,57
		1,58			0,92

Figure 25: Exemple de correction locale des associations : l’association proposée par l’algorithme en sortie de *matching* à gauche et la solution optimale corrigée par le script à droite.

2.7. Présentation des résultats de l’algorithme MixONat

Une fois le processus d’association terminé, le nombre de résultats choisis par l’utilisateur apparaîtra dans une nouvelle fenêtre. Les structures y apparaissent par score décroissant et écart croissant (**Figure 26**). En plus de la structure du composé, sont affichés son nom, sa masse molécule, son score et son écart. Sur la structure, les carbones identifiés apparaîtront surlignés en rouge.

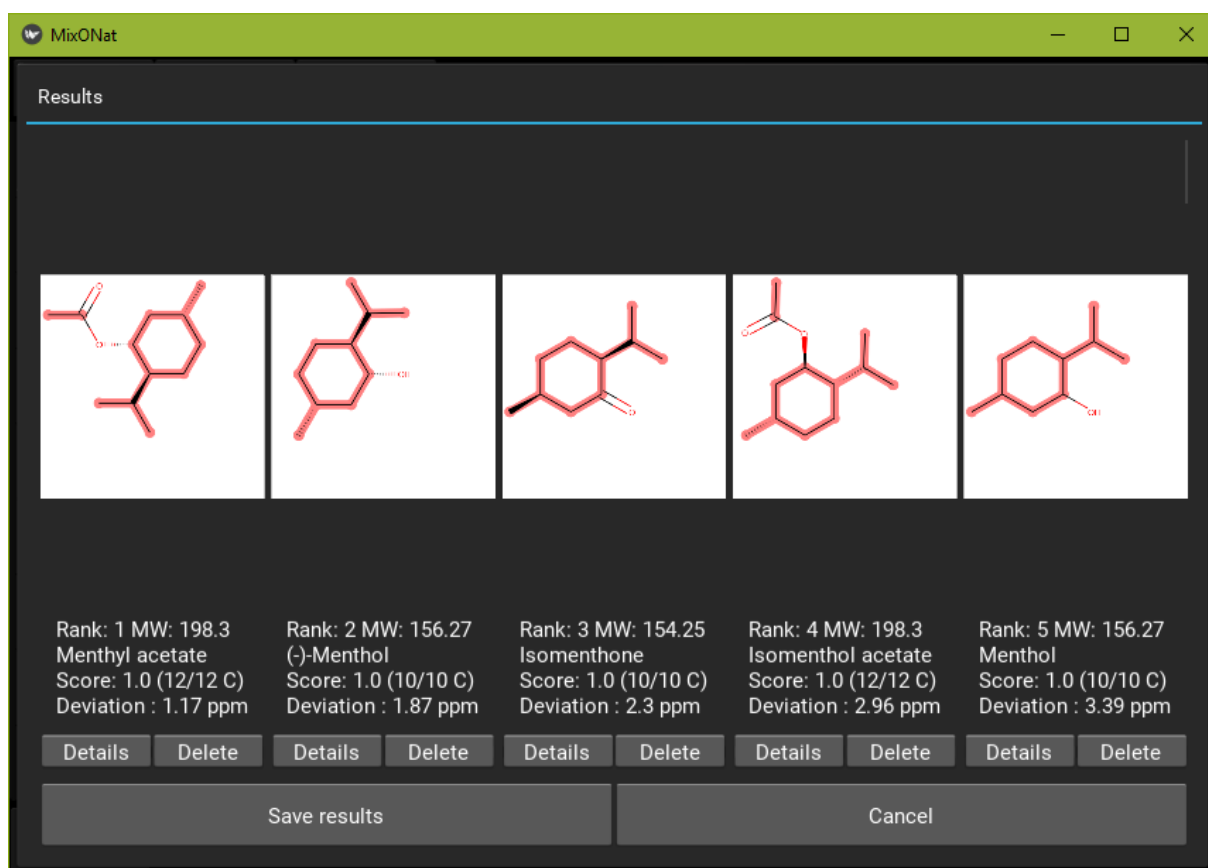


Figure 26: Fenêtre de présentation générale des résultats interactifs dans MixONat.

Il est possible de cliquer sur chaque structure pour avoir accès à des détails supplémentaires comme la numérotation des atomes sur la structure et la liste complète des δ_{SDF} triés par type de carbone (Figure 27). Un graphique représentant les $\delta_{13\text{C}}$ associés et leurs intensités respectives est également affiché sous la forme d'un spectre ^{13}C reconstitué. Sur cette présentation, les signaux apparaissent d'une couleur différente en fonction du type de carbone. Il est également possible de zoomer sur certaines zones du spectre et de l'exporter en tant qu'image. Cette fonctionnalité permet à l'utilisateur de rapidement repérer des déplacements chimiques dont l'intensité n'est clairement pas du même ordre de grandeur que les autres, et qui sont donc possiblement mal associés. L'utilisateur peut donc faire le lien entre structure, type de carbone, valeur de déplacement chimique et intensité, et ainsi prendre une décision *post-matching*. Après analyse des résultats par l'utilisateur, il est possible, à partir de la liste de déplacements chimiques précédemment mentionnée, de retirer un déplacement chimique qui avait été associé ou, au contraire, d'ajouter un déplacement chimique qui aurait dû l'être. Cela permet par exemple de retirer des déplacements chimiques d'une intensité anormale, repérés sur le graphique. Le chercheur peut aussi s'en servir pour ajouter un carbone qui n'avait pas été initialement sélectionné dans le spectre RMN ^{13}C , par exemple à cause de sa faible intensité. Le déplacement chimique étant bien présent, il peut donc être associé manuellement avec un carbone de la molécule et ainsi augmenter son score. En effet, ces modifications se répercuteront sur la structure (carbones associés en rouge) et le graphique, mais également sur le score, et donc le classement de la molécule en question. La mise à jour se fera automatiquement, sans avoir à relancer le programme. Enfin, il est également possible de supprimer une molécule entière des résultats lorsque l'utilisateur considère qu'elle n'a pas sa place dans les résultats, par exemple d'un simple point de vue chimiotaxonomique.

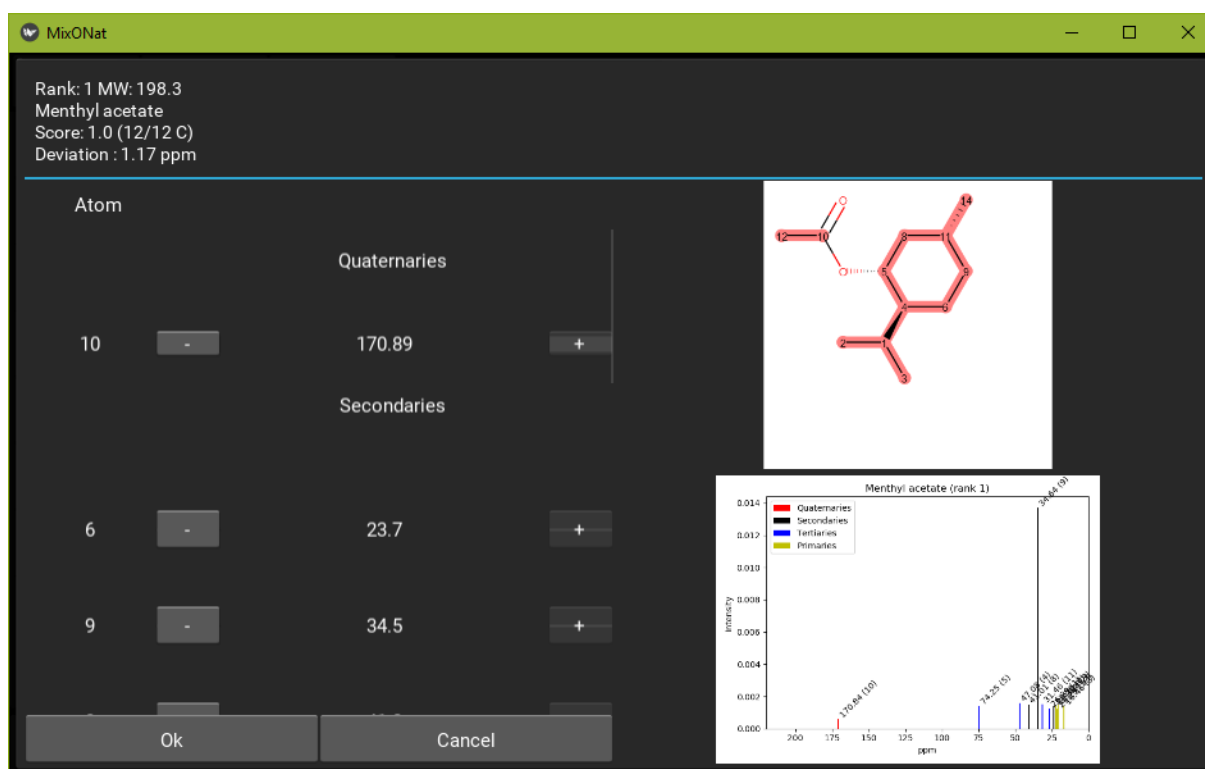


Figure 27: Fenêtre de présentation spécifique d'un résultat dans MixONat.

Une fois que le chercheur est intervenu et a modifié (ou non) les résultats à l'aide de ses connaissances analytiques et phytochimiques, ces derniers peuvent être sauvegardés sous la forme de 2 fichiers (**Figure 28**). Un fichier texte (.txt) qui résume les paramètres choisis pour la recherche et présente chaque molécule et les informations lui correspondant : nom, CAS, masse moléculaire, rang, écart, δ_c associés par type de carbone, etc... (**Figure 29**) Un fichier image (.png) représente les structures des molécules, avec les carbones associés surlignés en rouge, leur nom, masse moléculaire, score et écart (**Figure 30**). L'interface graphique permet de pouvoir lancer immédiatement une autre recherche en modifiant les paramètres ou les fichiers d'entrée souhaités.

Données de sortie

Un fichier texte comprenant:

- Nom des différents fichiers d'entrée (^{13}C , fichiers DEPT, SDF)
- Les paramètres utilisés pour la recherche (ϵ_{max} , da_{135} , da_{90})
- Le rang de la molécule, son ID, nom, numéro CAS, poids moléculaire et score
- La différence absolue cumulée entre les δ_c matchés (écart)
- La valeurs des δ_c du spectre (et leurs intensités) alignés avec les valeurs des δ_c du SDF avec lesquels ils ont été associés.

Un fichier image comprenant:

- La structure numérotée de la molécule sur laquelle les carbones matches apparaissent en rouge
- Le rang de la molécule, son nom (ou CAS ou ID), son poids moléculaire, score et écart

Figure 28: Fonctionnement de l'algorithme MixONat : données de sortie.

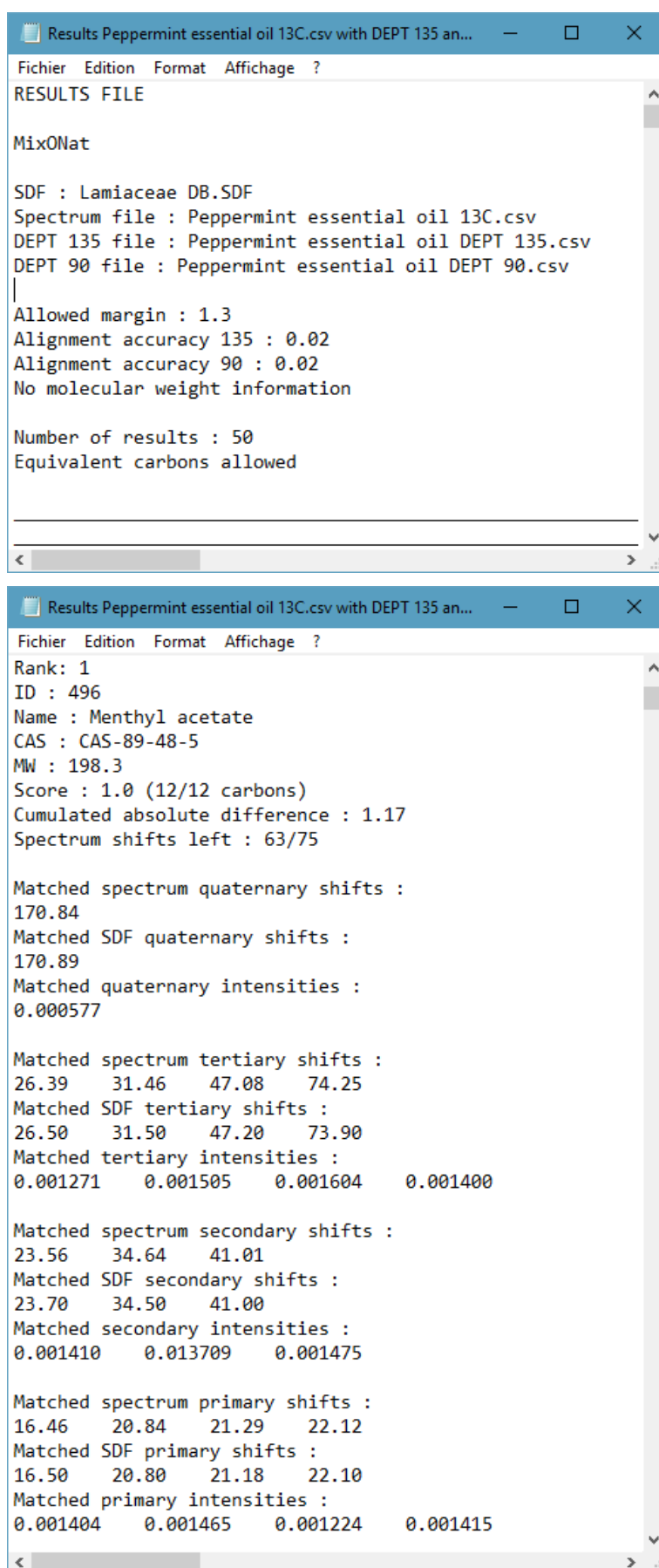


Figure 29: Extrait du fichier texte de résultats présentant les paramètres ainsi que le premier résultat d'une analyse

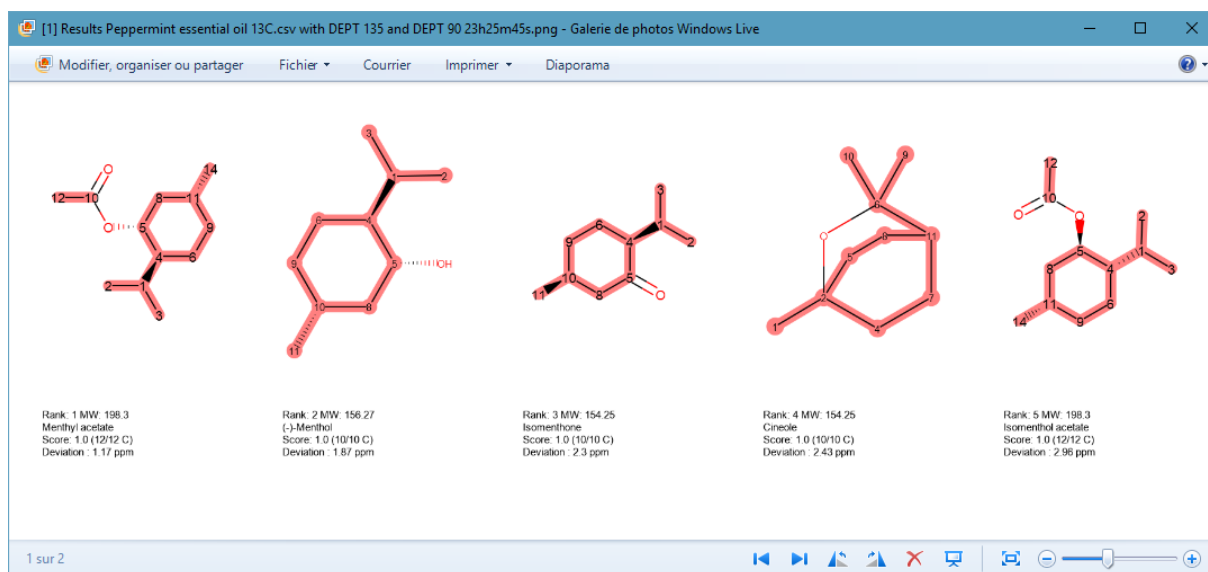


Figure 30: Extrait du fichier image de résultats présentant les 5 premiers résultats de la même analyse.

3. Validation de la méthode

L'algorithme étant opérationnel, il nous fallait en prouver l'efficacité sur des exemples de matrices de complexités variables, tant en considérant le nombre de molécules constitutives que leur diversité structurale. Ces exemples ont aussi été choisis car il s'agit d'extraits de compositions connues de plantes d'usage commercial, *i.e.* valorisées sur un plan médicinal, alimentaire ou cosmétique. Ont été ainsi choisis : l'huile essentielle de *Mentha piperita*, un extrait d'alcaloïdes totaux isoquinoléiques de *Papaver somniferum* (capsules), un extrait polaire et un extrait apolaire de *Rosmarinus officinalis* (sommités fleuries), ainsi que l'extrait standardisé E392 issu de la même plante [26], et enfin un extrait brut et une fraction enrichie de *Garcinia mangostana* (péricarpes des fruits).

Pour chacun de ces exemples, la démarche a été la même. Dans un premier temps, l'extrait a été caractérisé de façon non ambiguë, à l'aide de toutes les méthodes nécessaires (HPLC, SM, RMN). Puis, les résultats donnés par notre algorithme ont été comparés à la composition réelle de ces extraits. Enfin, ces résultats ont été scrupuleusement analysés afin de juger de la robustesse du programme : nous avons notamment tenté de comprendre pourquoi les molécules avaient été classées de la sorte par MixONat, pourquoi telle molécule était correctement retrouvée et pas telle autre, *etc...*

3.1. Huile essentielle de menthe poivrée

Un des extraits nous a permis de valider la méthode [27] est l'huile essentielle de *Mentha piperita*, utilisée traditionnellement en aromathérapie [28]. Comme pour chacun des autres exemples, nous avons d'abord pris le soin de caractériser la composition de l'huile essentielle par GC-MS et RMN ¹³C. Il a ensuite été possible de d'évaluer la pertinence des hypothèses proposées par le logiciel MixONat en les comparant avec les molécules présentes dans le mélange. Cet exemple a également été l'occasion de tester l'utilité du paramètre d'incrémentation de la fenêtre de recherche implémenté dans le logiciel.

3.1.1. Partie expérimentale

L'huile essentielle commerciale (Cooper) a été analysée par GC-MS (GCMS-QP2010 SE, Shimadzu) avec une colonne Phenomenex Zebron ZB-5 (30 m x 0,25 mm x 0,25 µm). L'hélium était le gaz vecteur avec un débit de 1,50 mL/min. La température de la colonne démarrait à 60°C pendant 10 min, et augmentait graduellement jusqu'à 180°C avec un taux de 2°C/min, pour finalement rester à 180°C pendant 5 min. L'analyse SM utilisait un système d'impact électronique (- 70eV), la source ionique était à 220°C et l'interface à 200°C. 1 µL de l'échantillon a été injecté (1/50 dilution dans du méthanol), avec un split ratio de 10.

Quelques gouttes de l'huile essentielle (environ 90 mg) furent également solubilisées dans du CDCl₃ pour analyse RMN-¹³C (1024 scans), DEPT 135 (512 scans) et DEPT 90 (512 scans). Les analyses RMN ont été conduites sur un spectromètre JEOL 400MHz YH.

La base de données « Lamiaceae » a été utilisée lors de l'analyse dérépliquative avec MixONat. Il s'agit de la même base prédite de 982 molécules que celle utilisée dans l'exemple du romarin, créée par recherche dans SciFinder [19] des produits décrits comme faisant parti de la famille des Lamiaceae. Le programme MixONat a été paramétré

de façon standard, à savoir une marge autorisée de 1,3 ppm (avec incrémentation), alignements des DEPT à 0,02 ppm et carbones équivalents autorisés.

3.1.2. Résultats et discussion

a) Composition de l’huile essentielle de menthe poivrée

L’analyse GC-MS (Figure 32 et Tableau 6) et la comparaison des δ_c avec ceux de la littérature (Tableau 7) ont permis de caractériser la composition de l’huile essentielle.

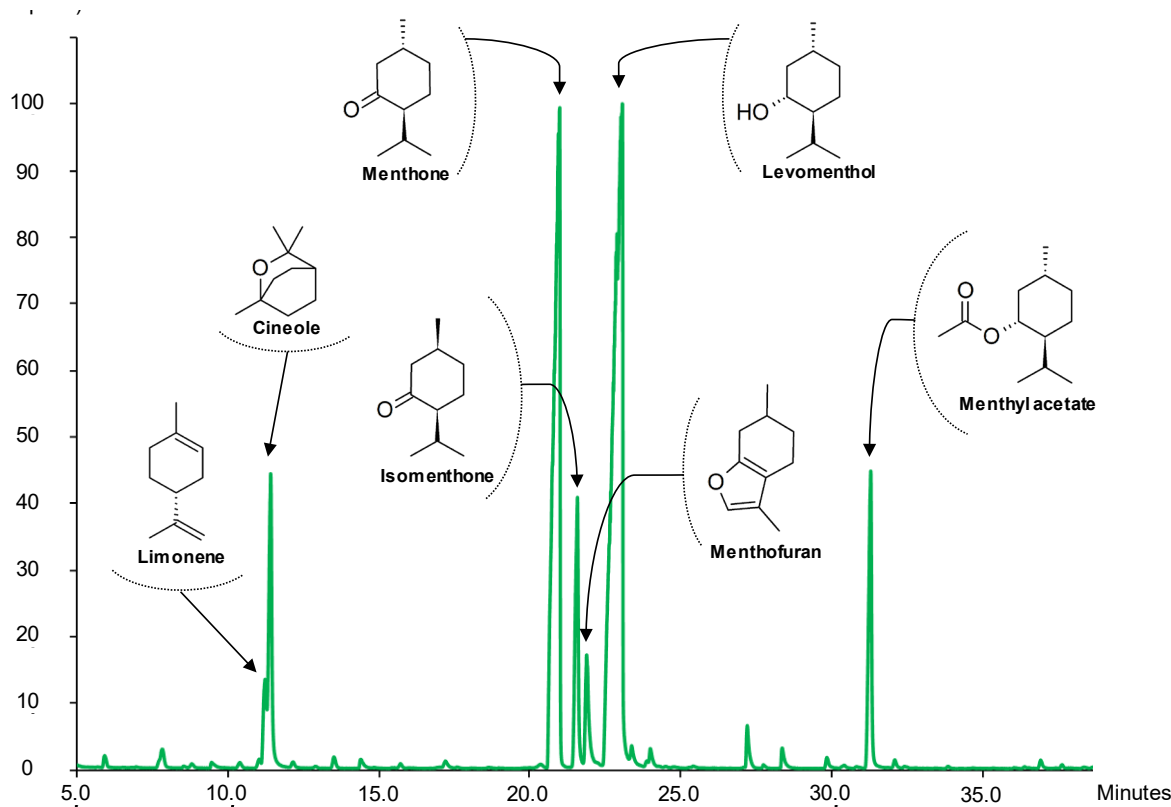


Figure 31: Chromatogramme GC-FID de l’huile essentielle de *Mentha piperita* et structures identifiées.

Tableau 6: Pourcentage relatif en GC-FID des différentes molécules contenues dans l’huile essentielle de *Mentha piperita*.

	Nom	%
1	(-)-Menthol	39,6
2	Menthone	27,3
3	Cineole	6,1
4	Menthyl acetate	5,2
5	Isomenthone	4,0
6	Limonene	2,8

Tableau 7: Comparaison des déplacements chimiques RMN ^{13}C entre l'huile essentielle de menthe et les données de la littérature [29] pour le menthol, la menthone, le cinéole, l'acétate de menthyle, l'isomenthone et le limonène.

	Menthol	Extrait		Menthone	Extrait		Cinéole	Extrait	
	71,5	71,5		36,0	35,6		73,6	73,7	
	50,1	50,2		51,3	50,9		31,6	31,6	
	23,1	23,2		212,9	212,6		22,9	22,9	
	34,5	34,5		56,3	56,0		33,0	33,0	
	31,6	31,7		28,3	28,0		-	-	
	45,0	45,2		34,4	34,0		-	-	
	22,2	22,3		22,8	22,4		69,8	69,9	
	25,8	25,9		26,4	26,0		28,9	29,0	
	21,0	21,0		21,7	21,3		-	-	
	16,1	16,2		19,2	18,8		27,6	27,7	

	Acétate de menthyle	Extrait		Isomenthone	Extrait		Limonène	Extrait	
	47,1	47,1		214,7	214,8		133,7	133,8	
	74,2	74,3		57,2	57,3		120,8	120,7	
	41,0	41,0		48,0	48,0		30,7	30,7	
	31,5	31,5		34,4	34,3		41,2	41,2	
	34,4	34,3		29,4	29,5		28,1	28,0	
	23,7	23,6		26,9	27,0		30,9	30,9	
	26,5	26,4		26,9	27,0		23,5	23,6	
	20,7	20,8		21,5	21,4		150,1	-	
	16,5	16,5		20,9	20,9		20,8	20,8	
	22,0	22,0		19,9	19,9		108,5	108,5	
	170,5	170,8		-	-		-	-	
	21,3	21,3		-	-		-	-	

On peut observer sur le tableau de comparaison δ_c (**Tableau 7**), un décalage homogène de 0,3 ou 0,4 ppm pour tous les signaux de la menthone lié certainement à un problème de référencement du spectre RMN décrit dans la littérature. Le signal manquant dans le spectre RMN ^{13}C de l'huile essentielle est un carbone quaternaire du limonène : on peut donc en déduire qu'il est surement confondu avec le bruit de fond, le limonène étant la molécule en plus faible concentration dans l'extrait (2,8 %).

b) Résultats de déréplication proposés par MixONat

Les résultats obtenus sont présentés en **Figure 32** et la position des molécules d'intérêt est récapitulée dans le **Tableau 8**.

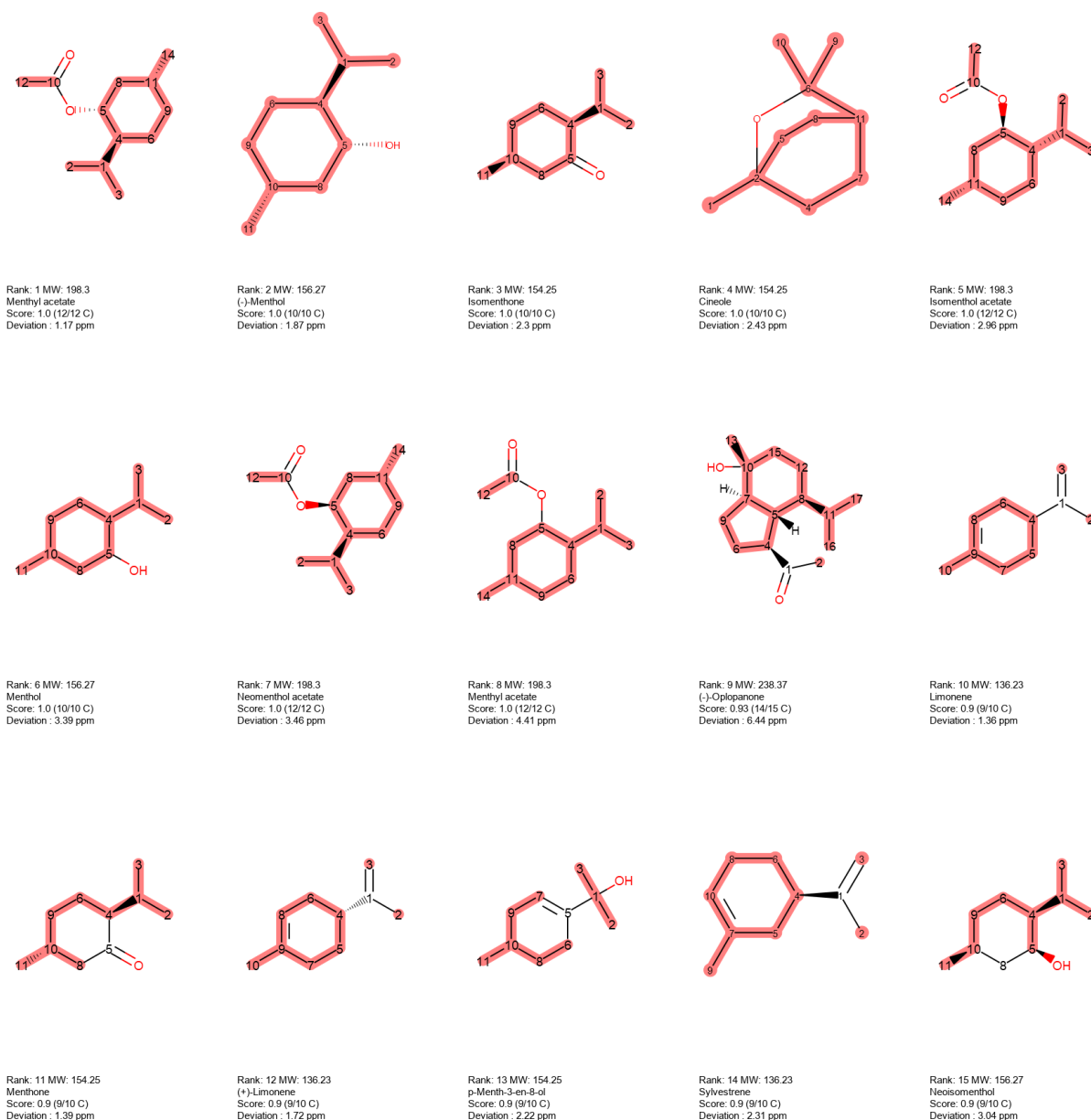


Figure 32: 15 premiers résultats de l'analyse déréplicative de l'huile essentielle de *Mentha piperita*.

Tableau 8: Nom, pourcentage relatif dans l'extrait, rang et score des molécules correctes.

	Name	%	Rang /980	Score /1,00
1	(-)-Menthol	39,6	2	1,00
2	Menthone	27,3	11	0,90
3	Cineole	6,1	4	1,00
4	Menthyl acetate	5,2	1	1,00
5	Isomenthone	4,0	3	1,00
6	Limonene	2,8	10	0,90

On remarque que 4 des 6 molécules présentes dans l'huile essentielle obtiennent un score parfait, tandis qu'un carbone n'a pas été associé pour chacune des 2 molécules restantes, résultant en un score de 90%. Après analyse des données spectrales, on peut s'apercevoir que le signal manquant de la menthone correspond à une prédiction trop éloignée de la réalité : la cétone est prédite à δ_c 210,82 ppm, alors qu'elle apparaît réellement à δ_c 212,62

ppm sur le spectre, soit $\Delta\delta_c = 1,8$ ppm ($> 1,3$ ppm, paramètre standard). Ce signal peut donc être rajouté manuellement grâce à l'interface graphique, ce qui donnera un score parfait à la molécule. Concernant le signal manquant du limonène, il s'agit d'un carbone quaternaire n'apparaissant pas sur le spectre, prédit à 149,65 ppm, et correspondant au carbone à 150,10 de la littérature. Comme précisé plus haut, ce signal est bel est bien absent du spectre.

Malgré ces signaux manquants, les molécules correctes arrivent dans les 11 premières positions sur les 980 structures de la banque de données. On remarque que les autres structures proposées sont des molécules extrêmement proches, puisqu'il s'agit de diastéréoisomères :

- de l'acétate de menthyle : acétate d'isomenthol (rang 5, score 1,00), acétate de néomenthol (rang 7, score 1,00), acétate de menthyle sans stéréochimie (rang 8, score 1,00).
- du (-)-menthol : menthol sans stéréochimie (rang 6, score 1,00), néoisomenthol (rang 15, score 0,90), isomenthol (rang 16, score 0,90).
- du limonène : (+)-limonène (rang 12, score 0,90).

Les hypothèses formulées par le programme MixONat sur cet exemple permettent d'arriver très rapidement à identifier les molécules entrant dans la composition du mélange analysé. Les limites de la méthode observées sur l'huile essentielle de menthe poivrée sont liées aux limites de détection en RMN- ^{13}C , et à la prédiction de déplacement chimique éloignée de la réalité.

Par ailleurs, cet exemple a également servi à mettre en évidence le rôle d'un paramètre évoqué précédemment : l'incrémentation de la marge autorisée. Cette fois ci, l'analyse sur l'huile essentielle a été effectuée deux fois, uniquement avec des données RMN- ^{13}C , sans ajout de *data* DEPT. Une première fois en activant l'incrémentation du facteur ϵ , puis une seconde fois en le désactivant. Les résultats sont comparés dans le **Tableau 8**.

Tableau 9: Effet de l'incrémentation de la marge autorisée lors de la recherche.

	Nom	%	Rang avec incrémentation /980	Score /1,00	Rang sans incrémentation /980
1	(-)-Menthol	39,6	3	1,00	11
2	Menthone	27,3	44	0,90	53
3	Cineole	6,1	4	1,00	7
4	Menthyl acetate	5,2	1	1,00	10
5	Isomenthone	4,0	2	1,00	9
6	Limonene	2,8	45	0,90	49

On remarque que l'inactivation de l'incrémentation, ce qui correspond à considérer que le premier déplacement chimique satisfaisant les conditions de la marge fixée est le bon, fait diminuer le rang de tous les terpènes réellement présents (de 3 à 9 places). Cela permet de renforcer l'hypothèse que la prédiction faite par le logiciel ACD/Labs® [12] est de bonne qualité, particulièrement lorsque les molécules d'intérêt sont bien décrites. Par conséquent, les banques de données sur lesquelles ACD/Labs® se base pour faire sa prédiction sont certainement

mieux fournies pour cette classe structurale que pour d'autres. La prédiction étant fiable, essayer d'abord d'associer les déplacements chimiques possédant la plus petite différence possible favorise donc les molécules de l'huile essentielle et pénalise les molécules non présentes. Bien que cela impacte rarement le score donné à la molécule, cela a un effet sur le second facteur qui sert à classer les molécules entre elles : l'écart (ou *deviation*). Il est donc recommandé aux utilisateurs de MixONat de laisser ce paramètre d'incrémentation activé lors des analyses avec le programme. Lors de l'utilisation de banques de données expérimentales, ce facteur est évidemment capital, puisque les données du mélange et celles de la référence sont normalement identiques.

3.2. Article 3 : MixONat, a software for mixtures dereplication based on ^{13}C -NMR experiments.

3.2.1. Résumé de l'article 3

Les autres exemples qui ont permis la validation de la méthode sont présentés dans l'article suivant qui reprend les points principaux de ce travail [27].

Les **alcaloïdes isoquinoléiques du pavot somnifère** sont largement utilisés dans le domaine médical comme antalgiques. A partir d'un extrait aqueux acide de *Papaver somniferum*, après alcalinisation, une extraction liquide-liquide par du dichlorométhane a permis l'obtention de l'extrait DCM ensuite analysé. Une analyse HPLC, RMN et LDI de cet extrait a mené, notamment par comparaison avec des standards, à l'identification formelle des composés présents : papavérine et noscapine.

En vue du travail de déréplication à l'aide de notre programme, une base de données prédite « Papaveraceae » a été créée à partir des 174 produits naturels décrits dans SciFinder [19] comme étant isolés de la famille des Papaveraceae. Cette base de données fut utilisée en référence lors de l'analyse du spectre RMN- ^{13}C de l'extrait, avec les paramètres de recherche standard (marge autorisée de 1,3 ppm, incrémentation activée, carbones équivalents autorisés). La papavérine et la noscapine apparaissent dans les 4 premières positions. L'ajout de données DEPT 135 et DEPT 90 permet de les obtenir dans les 3 premières positions mais surtout, de pénaliser le score des molécules non présentes dans l'extrait. La molécule placée en position 2 a pu également être éliminée, notamment sur la base de critères chimiotaxonomiques. En effet, les autres molécules suggérées dans les premières positions (papavérine et noscapine) sont des molécules caractéristiques du pavot somnifère (*Papaver somniferum*) et non du pavot de Californie (*Eschscholtzia californica*) duquel l'eschscholtzine, suggérée en position 2, est caractéristique. L'élimination de l'eschscholtzine laissant ainsi les molécules attendues en positions 1 et 2.

Les **feuilles du romarin officinal** sont traditionnellement utilisées pour soulager indigestion et trouble spasmodiques du tractus gastrointestinal [29]. Des extraits de romarin sont également connus comme additif alimentaire comme l'E392 [26]. Une extraction sous pression des feuilles, successivement par du dichlorométhane puis du méthanol, a permis d'obtenir 2 extraits ensuite analysés. Parallèlement, un extrait E392 a été réalisé par extractions successives des feuilles par du cyclohexane puis de l'éthanol ; l'extrait éthanolique correspond à l'extrait E392. Des analyses par HPLC, SM et RMN ont permis de caractériser les molécules présentes au sein de chacun des extraits. L'extrait dichlorométhanique contient triterpènes et diterpènes : acide ursolique, acide oléanolique, acide bétulinique, acide micromérique, et acide carnosique. L'extrait méthanolique contient de l'acide

rosmarinique ainsi que du saccharose. L'extrait E392 est très similaire à l'extrait dichlorométhanique : il est composé d'acide ursolique, d'acide oléanolique, d'acide bétulinique, d'acide micromérique, et de carnosol. Il convient de noter que l'acide carnosique s'oxyde spontanément pour former du carnosol et autres produits de dégradation.

Une base de données prédite « Lamiaceae » a été créée par recherche dans SciFinder [19] des produits décrits comme faisant parti de la famille des Lamiaceae, donnant 982 molécules. Les 3 extraits ont été analysés par RMN-¹³C, DEPT 135 et DEPT 90, puis les données obtenues ont été analysées grâce à MixONat avec les paramètres standards. L'analyse de l'extrait méthanolique par l'algorithme place l'acide rosmarinique en position 4 et le saccharose en position 5, tous deux avec un score de 83%. Les molécules mieux classées sont l'acide caféique, le tyrosol, et un ester d'acide caféique et de saccharose. Il est intéressant de noter que ces produits naturels sont en fait des fragments ou des esters des composés réellement dans l'extrait, l'acide rosmarinique étant, par exemple, un ester d'une unité d'acide caféique. Pour les extraits dichlorométhanique et E392, un filtre de poids moléculaire a également été appliqué : afin d'éliminer une grande partie des composés volatiles (mono et sesquiterpènes) reportés dans la famille des Lamiaceae, seuls ceux ayant une masse supérieure à 250 g/mol ont été considérés. Pour l'extrait E392, le programme propose l'acide bétulinique, l'acide ursolique, l'acide micromérique, l'acide oléanolique et l'acide carnosique en position 4, 6, 11, 17 et 22 respectivement, tous avec un score supérieur à 75%. Les autres suggestions sont des molécules possédant des squelettes similaires de triterpènes ou diterpènes. Il suffit donc à l'utilisateur de comparer les 22 premières positions aux données de la littérature pour connaître la composition de l'extrait. En effet, une fois cette position atteinte, 5 molécules ont été confirmées dans l'extrait, et plus de 95% des signaux de l'extrait ont été attribués. Alternativement, une analyse LC-MS donnant les poids moléculaires des composés majoritaires, à savoir 330, 332, 454 et 456 g/mol, permet aussi un filtrage plus sévère de la base de données et de pouvoir apprécier les molécules correctes dans les 5 premières positions. Des résultats très similaires en termes de classement et de score ont été obtenus pour l'extrait dichlorométhanique, hormis le fait que le carnosol remplace l'acide carnosique.

Les **péricarpes des fruits du mangoustanier** sont connus pour leur richesse en antioxydants, notamment en xanthones. Notre laboratoire porte une attention particulière aux molécules de la famille des xanthones car elles semblent avoir un effet sur la voie de l'*Unfolded Protein Response* (UPR) [30-32], une voie d'adaptation au stress cellulaire, la rendant une cible attrayante dans de nombreux domaines comme la protection des cultures, ou le traitement de maladies dégénératives ou métaboliques chez l'Homme. Après extraction sous pression des péricarpes broyés par du cyclohexane, l'extrait obtenu a été analysé par HPLC et RMN afin de le caractériser. Les composés majoritaires d'après l'HPLC-UV sont l'alpha-mangostine (69%) et la gamma-mangostine (14%). La gartanine (5%) et la garcinone E (5%) sont également présentes, le reste des molécules, à savoir les 8-deoxygartanine (4%), bêta-mangostine (1%), gudraxanthone (1%) et 9-hydroxycalabaxanthone (< 1%), est très minoritaire. L'extrait cyclohexanique fût ensuite fractionné par chromatographie flash en phase normale, puis une des fractions obtenues a été analysée par HPLC et RMN. La fraction se compose de gartanine (49%), 8-deoxygartanine (30%), bêta-mangostine (15%) et gudraxanthone (6%).

La base de données « *Garcinia* », contenant les prédictions RMN- ^{13}C des 718 produits naturels reportés dans le genre *Garcinia* d'après le DNP, a été utilisée pour le travail de déréplication. Les paramètres sélectionnés sont ceux par défaut pour analyser les données ^{13}C -RMN, DEPT 135 et DEPT 90. Dans l'extrait, l'alpha-mangostine, la garcinone E, la gamma-mangostine, la gartanine et la 8-deoxygartanine arrivent en position 1, 4, 10, 11 et 20 respectivement. On peut noter que les 25 premières propositions sont également des xanthones prénylées, ayant un score supérieur à 90%. Sur l'analyse de la fraction, la 8-deoxygartanin, la bêta-mangostine, la gartanine et la gudraxanthone furent classées respectivement en 1, 2, 10 et 24 avec un score parfait pour les 2 premières molécules, et un score avoisinant les 90% pour les 2 autres. Les autres propositions sont également des xanthones prénylées, ce qui permet d'arriver très rapidement à savoir quel type de structure est présent dans l'extrait. La **Figure 33** illustre la proximité structurale des différentes propositions faites par MixONat.

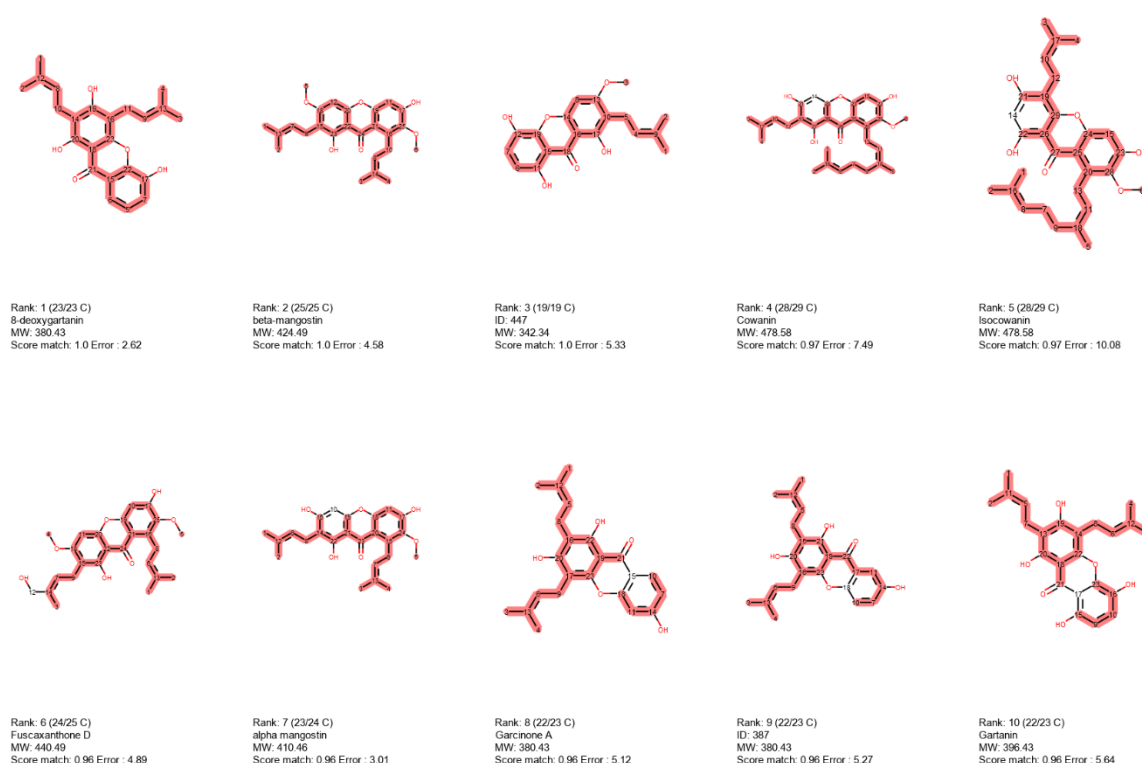


Figure 33: Résultats 1-10 pour la déréplication de la fraction enrichie de *Garcinia mangostana*.

3.2.2. Article 3

13 **Abstract**

14 Whether chemists or biologists, researchers dealing with metabolomics require tools to
15 decipher complex mixtures. As a part of metabolomics and initially dedicated to identifying
16 bioactive natural products, dereplication aims at reducing the usually time-consuming process
17 of known compounds isolation. Mass spectrometry and nuclear magnetic resonance are the
18 most commonly reported analytical tools during dereplication analysis. Though low sensitive,
19 ^{13}C -NMR has many advantages for such a study. Noteworthy, it is nonspecific allowing
20 simultaneous high-resolution analysis of any organic compounds including stereoisomers.
21 Since NMR spectrometers nowadays provide useful dataset in a reasonable time frame, we have
22 embarked upon writing a software dedicated to ^{13}C -NMR dereplication. The present study
23 describes the development of a freely distributed algorithm, namely MixONat and its ability to
24 help researchers decipher complex mixtures. Based on Python 3.5, MixONat analyses a $\{^1\text{H}\}$ -
25 ^{13}C NMR spectrum optionally combined with DEPT-135 and 90 data - to distinguish carbon
26 types (*i.e.* CH_3 , CH_2 , CH and C) - as well as a MW filtering. The software requires predicted
27 or experimental carbon chemical shifts (δc) databases and displays results that can be refined
28 based on user interactions.

29 As a proof of concept, this ^{13}C -NMR dereplication strategy was evaluated on mixtures of
30 increasing complexity and exhibiting pharmaceutical (poppy alkaloids), nutritional (rosemary
31 extracts) or cosmetics (mangosteen peel extract) applications. Associated results were
32 compared with other methods commonly used for dereplication. MixONat gave coherent results
33 that rapidly oriented the user towards the correct structural types of secondary metabolites,
34 allowing the user to distinguish between structurally close natural products, including
35 stereoisomers.

36 Whether for natural products (NPs) identification from extracts or fractions,^{1, 2} for predicting
37 the composition of biological samples³ or for crude reaction analysis in organic synthesis,⁴
38 chemists require tools to decipher complex mixtures. Since 1990, regarding secondary
39 metabolites in the context of drug discovery, the idea of making structural assumptions about
40 the composition of a complex mixture without any NPs isolation have emerged as the so-called
41 term dereplication.⁵ Initially, the latter aimed at avoiding bioguided fractionation steps as well
42 as isolation and structural determination of well-known NPs. As a part of metabolomics,⁶
43 dereplication usually requires mass spectrometry (MS) and nuclear magnetic resonance (NMR)
44 detection together with analytical and preparative chromatographic separation, most often in a
45 hyphenated manner.⁷ Then, comparison between spectral data and NPs databases (DBs) allows
46 formulating hypotheses as to the mixture composition.⁸ More recently, aiming at limiting long
47 chromatographic steps or because the sample itself (*e.g.* amount, composition) does not allow
48 them, dereplication was successfully performed directly on crude extracts and fractions.⁹⁻¹¹

49 Nowadays, LC-MS² and the molecular networking initiative seem to gain importance in NPs
50 dereplication¹² though actually both LC-MS and NMR exhibit advantages and drawbacks. The
51 LC step before MS analyses requires an optimization of suitable chromatographic conditions
52 that can waste valuable time. Moreover, while MS provides a higher sensitivity, a standard
53 ionization protocol may not be suitable for a wide range of structurally different NPs. Then,
54 even if high-resolution mass spectra give the molecular formula, the number of possible
55 compounds may remain important. Therefore, MS/MS experiments allow their differentiation
56 according to its fragmentation patterns. Introduced in 2013 by the Dorrestein group, molecular
57 networking based on the LC-MS² analysis of different but linked samples is now widely used
58 as an efficient dereplication method.¹³ In this process, similar MS² behavior of structurally
59 closed compounds leads to their gathering in clusters of nodes based on reference standards
60 fragmentations.¹⁴ The related Global Natural Product Social Molecular Networking (GNPS) is

61 a web-based facility allowing sharing of raw, processed or identified MS² data.¹⁵ Conversely, if
62 a few milligrams of the mixture are available (*i.e.* 1-20 mg according to the sample), NMR is
63 easy to implement, as it does not need any specific sample preparation, except for solubilization
64 in a suitable deuterated solvent. Last but not least, NMR detects all organic compounds and
65 allows differentiation of stereoisomers, which may be very challenging or even impossible
66 through MS analysis^{2, 16-18}.

67 That is why both 1D and 2D NMR experiments are often used in metabolomics¹⁶ and
68 dereplication applications, in addition or as an alternative to MS analysis.^{19, 20} Metabolomics
69 routinely uses ¹H NMR together with spectral deconvolution²¹ to make hypotheses about the
70 attendance of primary and, from time to time, secondary metabolites in a given organism, organ
71 or tissue. The good sensitivity of ¹H spectra allows short acquisition times, but complex
72 mixtures generate ¹H chemical shifts (δ_H) overlapping because of the low spectral range of ¹H-
73 NMR which impairs correct spectra interpretation. Even if {¹H}-¹³C NMR exhibits many
74 advantages including a large spectral dispersion almost preventing from overlapping signals, it
75 is hardly ever used for metabolomics purposes as longer acquisition times are required due to
76 ¹³C low natural abundance.^{22, 23} Eventually on-line experimental NMR DBs such as HMDB²⁴
77 or BMRB²⁵ may be searched either manually or automatically in order to identify metabolites
78 through 1D δ associations and/or 2D correlations matching.

79 In the field of NPs research, NMR-based dereplication processes have actually emerged
80 together with MS-based methods. Indeed hyphenated LC-NMR was first applied to NPs
81 mixtures in the early 90s.²⁶ In this particular case, NPs were not characterized in crude mixtures
82 strictly speaking but on the basis of 1D and 2D NMR data of an isolated compound after an
83 appropriate HPLC chromatography. To our knowledge, dereplication using ¹³C NMR was
84 initiated even earlier, *i.e.* in 1982 on essential oils²⁷, the process being automated later, in
85 1995.²⁸ The associated algorithm works with an in-house DB containing the δ_c of volatile

86 mono- and sesquiterpenes recorded in CDC13. Then, the SISCONST program was developed
 87 and applied for the dereplication of a mixture of terpenes in 2001, using an experimental DB of
 88 3800 mono- and sesquiterpenes. SISCONST required δ_c and carbon multiplicities deduced from
 89 DEPT experiments. Evaluated on volatile oils, it correctly identified major compounds (>
 90 2.2%).^{29, 30} Indeed, as far as ^{13}C NMR dereplication is concerned, carbon multiplicity appears
 91 as quite a discriminant filter. Based on a *Garcinia* genus (Clusiaceae) DB (772 NPs)³¹ this
 92 funnel effect is illustrated in Table 1.

93
 94 **Table 1. Funnel effect of multiplicities obtained by DEPT-135 additional experiment on**
 95 **NPs from *Garcinia* DB (772 NPs). A. Number of NPs sharing the same combination of carbon**
 96 **type. B. Number of time that one combination leads to a group of the size displayed on the right**
 97 **column.**

A.								B.	
Cq	CH ₂	CH+CH ₃	Nb of NPs	Cq	CH ₂	CH+CH ₃	Nb of NPs	Nb of occurrences	NPs in the group
13	2	9	21	12	3	9	8	164	1
15	3	10	16	13	2	10	8	49	2
11	1	7	15	14	4	10	8	24	3
15	6	17	15	16	4	13	8	18	4
12	2	9	13	17	6	15	8	10	5
13	2	8	12	11	0	7	7	8	7
16	6	16	12	11	1	6	7	7	6
13	1	10	10	11	4	15	7	7	8
13	1	9	10	13	3	7	7	3	9
15	2	11	10	14	4	11	7	3	10
12	1	10	9	15	1	12	7	2	12
15	5	18	9	15	8	15	7	2	15
8	0	5	9	16	8	19	7	1	13
11	1	9	8	10	0	8	6	1	16
12	3	8	8					1	21

98
 99 Most recently, in 2014, an efficient strategy based on the fractionation of a crude extract
 100 leading to a dataset of ^{13}C -NMR spectra analyzed by a hierarchical clustering analysis was
 101 proposed. The clusters of δ_c are those of the major NPs from the mixture and may be identified
 102 using either an experimental or a predicted DB.³² The same researchers published a freely

103 available algorithm aiming at comparing δ_c of NPs from a crude alkaloid extract of boldo with
104 those of a predicted DB, taking into account signals intensities as well.⁹

105 Considering this, we describe here the development of a freely distributed algorithm, namely
106 MixONat (Mixture of Natural Products, Available in Supporting information / Sourceforge),
107 and its ability to dereplicate mixtures of NPs using a $\{^1\text{H}\}$ - ^{13}C NMR spectrum combined with
108 DEPT-135 and 90 from either experimental or predicted δ_c comparisons. As a proof of concept,
109 the procedure was applied to various vegetal fractions or crude extracts of pharmaceutical
110 (poppy alkaloids), nutritional (rosemary extracts) or cosmetics (mangosteen peel extract)
111 interest.

112 EXPERIMENTAL SECTION

113 **Chemical and reagents.** Papaverine hydrochloride was purchased from Sigma-Aldrich
114 (Saint Quentin Fallavier, France). It was solubilized in water and the solution was alkalinized
115 to pH 12 using a 1M NaOH solution. It was then extracted 3 times with dichloromethane
116 (DCM). DCM phases were gathered, dried and further analyzed. Rosmarinic acid standard was
117 bought from Molekula (St Jean de Soudain, France).

118 **Plant material, extraction and fractionation and analyses.** *Papaver somniferum*: Poppy
119 pods (Sample PS-201810, confidential breeding) were extracted by an aqueous solution (acetic
120 acid 1%). The extract was filtrated on paper then alkalinized to pH 12 using a 1 M NaOH
121 solution. It was then extracted 3 times with DCM. DCM phases were gathered, dried and
122 analyzed by HPLC-UV (Agilent HP 1100 Series, Agilent Technologies, Les Ulis, France) using
123 a Gemini C18 column (150 x 4.6 mm, 3 μm , 100 \AA , Phenomenex, Le Pecq, France) with mobile
124 phase A = 0.005 M sodium 1-heptanesulfonate buffer adjusted to pH 2.6; phase B = methanol.
125 With a flow of 1.2 mL/min, the gradient was programmed as follows: t = 0 min, 75% A, 25%
126 B; t = 15 min, 45% A, 55% B; t = 18 min, 45% A, 55% B; t = 20 min, 75% A, 25% B. The
127 extract was also analyzed using LDI-MS in positive reflectron mode on a Biflex III time of

128 flight (TOF) mass spectrometer (Bruker Daltonik, Bremen, Germany) equipped with a 337 nm
129 pulsed nitrogen laser (model VSL-337i, Laser Sciences Inc., Boston, MA). A mass range of 40-
130 2000 Da was chosen for spectra acquisition. Acceleration voltage was set to 19 kV, pulse ion
131 extraction was 200 ns and laser frequency was 5 Hz. Applied laser energy ranged from 65 to
132 75% (86.5-93.1 μ J). Additionally, ^{13}C -NMR (17 408 scans), DEPT-135 (8 704 scans) and
133 DEPT-90 (8 704 scans) spectra were obtained in CDCl_3 .

134 *Rosmarinus officinalis*: Rosemary sample was bought from IPHYM laboratories (Jonage,
135 France) in October 2018. 10 g of plant material was extracted using pressurized liquid
136 extraction (Speed Extractor E-914, Büchi, Essen, Germany) first with DCM (3 cycles, 10 min
137 each) under 100°C and 100 bars, to obtain 1.8 g of extract (18%), and then with methanol,
138 during (3 cycles, 10 min each) under 100°C and 100 bars, to obtain 1.7 g of extract (17%). The
139 E392 was obtained following one of the protocols described by the European Food Safety
140 Authority (EFSA).³³ 2 g of dried rosemary leaves were extracted by 50 mL of cyclohexane
141 (ultrasonic bath 10 min) and then by 50 mL of ethanol (ultrasonic bath 10 min). 45 mg (2%) of
142 cyclohexanic extract and 180 mg (9%) of ethanolic extract were obtained. The methanol extract
143 was analyzed by HPLC (Prominence-i LC-2030C, Shimadzu, Noisiel, France) coupled with
144 ELSD (SEDEX 90 LT-ELSD, SEDERE) using a Luna C18 column (250 x 4.6 mm, 5 μ m, 100Å,
145 Phenomenex) with mobile phase A = water + 0.1% formic acid; phase B = methanol. With a
146 flow of 1 mL/min, the gradient was programmed as follows: t = 0 min, 70% A, 30% B; t = 5
147 min, 65% A, 35% B; t = 10 min, 55% A, 45% B; t = 30 min, 30% A, 70% B; t = 31 min, 0% A,
148 100%; t = 36 min, 0% A, 100% B. The sample was also analyzed using HPLC-MS (Esquire
149 3000 Plus, Ion trap, Bruker) in the same LC conditions. Furthermore, ^{13}C -NMR (2 048 scans),
150 DEPT-135 (1 024 scans) and DEPT-90 (1 024 scans) spectra of the methanol extract (50 mg)
151 were obtained in MeOD. The dichloromethanic extract and E392 extracts were analyzed with
152 the same HPLC-ELSD instrument, on the same Luna C18 column but with mobile phase A =

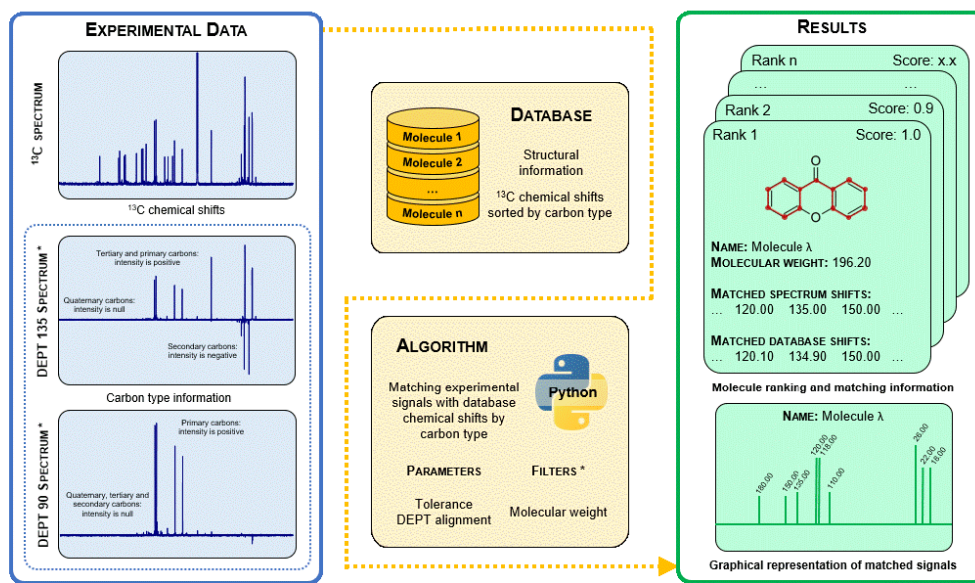
153 water + 0.1% formic acid; phase B = acetonitrile. With a flow of 1 mL/min, the gradient was
154 programmed as follows: t = 0 min, 40% A, 60% B; t = 25 min, 25% A, 75% B; t = 40 min, 25%
155 A, 75% B; t = 41 min, 0% A, 100% B; t = 46 min, 0% A, 100% B. The dichloromethanic extract
156 was also analyzed using the same HPLC-MS instrument, in the LC conditions previously
157 mentioned. Furthermore, ¹³C-NMR (10 000 scans), DEPT-135 (5 000 scans) and DEPT-90 (5
158 000 scans) spectra of the dichloromethanic extract (50 mg) were obtained in DMSO-d₆. The
159 E392 extract (50 mg) was also analyzed by NMR, ¹³C-NMR (1 024 scans), DEPT-135 (512
160 scans) and DEPT-90 (512 scans) in DMSO-d₆.

161 *Garcinia mangostana*: 1 kg of fresh mangosteen fruits, imported from Vietnam, were
162 purchased in an Asian grocery store (Tang Frères, Paris, France) in September 2017. A sample
163 was identified by PR as *Garcinia mangostana* fruits and a voucher specimen (GM-201701) was
164 kept at the laboratory's herbarium. The pericarps were separated from the rest of the fruit, dried
165 and grinded. 45 g of grinded pericarps were extracted using pressurized liquid extraction (Speed
166 Extractor E-914, Büchi) with cyclohexane (3 cycles, 8 min each) under 100°C and 100 bars, to
167 obtain 5.6 g of extract (12.4%). The cyclohexane extract was analyzed by HPLC-UV (Waters
168 2695 with a diode array detector Waters 2996, Guyancourt, France) using a Gemini C18 column
169 (150 x 4.6 mm, 3µm, 100Å, Phenomenex) with mobile phase A = water + 0.1% formic acid;
170 phase B = acetonitrile. With a flow of 0.75 mL/min, the gradient was programmed as follows:
171 t = 0 min, 40% A, 60% B; t = 30 min, 10% A, 90% B; t = 32 min, 0% A, 100% B; t = 37 min,
172 0% A, 100% B. Additionally, ¹³C-NMR (10 000 scans), DEPT-135 (8 000 scans) and DEPT-90
173 (8 000 scans) spectra were obtained in CDCl₃.

174 4 g of the cyclohexane extract were fractionated with a CombiFlash Rf-200 system (Teledyne
175 Isco, Lincoln, NE, USA) equipped with binary pumps, multiwavelength UV detectors, and
176 fraction collectors using a Si-HC 50µm Si-OH 120g column (Interchim, Montluçon, France)
177 with mobile phase A = petrol ether; phase B = 60% petrol ether, 30% chloroform, 10% acetone.

178 With a flow of 60 mL/min, the gradient was programmed as follows: t = 0 min, 100% A, 0%
179 B; t = 2 min, 100% A, 0% B; t = 7 min, 80% A, 20% B; t = 12 min, 80% A, 20% B; t = 17 min,
180 60 % A, 40% B; t = 22 min, 60% A, 20% B; t = 27 min, 40 % A, 60% B; t = 42 min, 40% A,
181 60% B; t = 50 min, 0% A, 100% B; t = 60 min, 0% A, 100% B. 7 fractions were obtained: F1
182 (52.4 mg, 9-hydroxycalabaxanthone³⁴), F2 (90.9 mg, garcinone E³⁵), F3 (1.8 g, α -mangostin³⁴),
183 F4 (483.0 mg, γ -mangostin³⁴), F5 (105.0 mg), F6 (381.2 mg) and F7 (312.3 mg). F6 was
184 analyzed by HPLC-UV, in the same conditions as the crude extract. Additionally, ¹³C-NMR (12
185 000 scans), DEPT-135 (6 000 scans) and DEPT-90 (6 000 scans) spectra were obtained in
186 CDCl₃.

187 310.7 mg of F6 were purified by flash chromatography (CombiFlash Rf), using a Si-OH 40 μ m
188 40g column (Macherey-Nagel, Hoerd, France) with mobile phase A = petrol ether; phase B =
189 50% chloroform, 50% ethyl acetate. With a flow of 30 mL/min, the gradient was programmed
190 as follows: t = 0 min, 100% A, 0% B; t = 5 min, 100% A, 0% B; t = 10 min, 90% A, 10% B; t
191 = 20 min, 90% A, 10% B; t = 22 min, 86% A, 14% B; t = 33 min, 86% A, 14% B; t = 35 min,
192 80% A, 20% B; t = 45 min, 80% A, 20% B; t = 50 min, 70% A, 30% B; t = 60 min, 70% A,
193 30% B; t = 65 min, 0% A, 100% B. From t = 70 min to t = 75 min, the solvent system was
194 switched to 100% ethyl acetate. 8-deoxygartanin³⁶ (23 mg), gartanin³⁶ 49 mg), β -mangostin³⁷
195 (25 mg) and gudraxanthone³⁴ (12 mg) were purified.
196



197 * Optional

198 **Figure 1.** General process of the MixONat program. The program matches chemical shifts from
 199 the user's experimental (^{13}C and optionally DEPT-135 and DEPT-90) with those of a database
 200 collecting selected molecules δ_{C} .

202 **NMR spectrometer.** NMR analyses were conducted on a JEOL 400MHz YH spectrometer
 203 (JEOL Europe, Croissy-sur-Seine, France).

204 **DBs: general information.** To create a DB of NPs and their δ_{C} that can be used by the
 205 dereplication program, the first step is to gather the structures of the compounds of interest (*e.g.*
 206 NPs from a genus or a botanical family), either by drawing them with a dedicated software (*e.g.*
 207 ChemDraw³⁸, ChemSketch³⁹) or downloading them from various available DBs (*e.g.*
 208 SciFinder⁴⁰, Dictionary of Natural Products³¹). Once the individual files (.mol, .cdx, .sk2) are
 209 gathered in a structure data file (.sdf), their δ_{C} are predicted using a NMR prediction software
 210 (*i.e.* ACD/Labs Spectrus processor and C/H-NMR Predictor).^{41, 42} The C-typeGen program
 211 included in the software (Figure S1) creates a suitable DB: it reads the SDF and sorts each

212 chemical shift by carbon type. A new SDF is then created. The latter contains, for each
213 compound of the DB, the predicted δ_C values organized as methyl, methylene, methine or
214 quaternary carbons. The creation of such a DB is required for the MixONat algorithm to work
215 properly.⁴³

216 **Specific DBs.** Garcinia DB was created by gathering the structures of molecules described in
217 the *Garcinia* genus from the Dictionary of Natural Products³¹, leading to a SDF containing 718
218 NPs. Lamiaceae DB was built by searching for compounds described in the Lamiaceae family
219 on SciFinder, resulting to a database of 982 NPs. Papaveraceae DB was also created using
220 SciFinder, gathering molecules isolated from the Papaveraceae, resulting in a 174 NPs DB. The
221 CH-NMR-NP DB containing the experimental δ_C of 32 854 NPs was created from the data
222 available (*i.e.* structure, name, molecular formula, molecular weight, source, ^{13}C chemical
223 shifts, deuterated solvent, reference) on the dedicated JEOL website⁴⁴.

224 **MixONat software.** The algorithm was implemented in the Python 3.5 programming
225 language.⁴⁵ The open source cheminformatics package RDKit was used to draw the molecular
226 structures and read SDF files⁴⁶. A detailed version can be found in Supporting information.

227 **MixONat software: Inputs and parameters.** A graphical user interface (GUI) was designed
228 with Kivy, an open Python library source compatible with Linux, Windows and OS X.

229 The home tab of the MixONat software (Figure S2) allows the selection of the input files by
230 the user, using a file browsing function. The input files must include at least a DB (.sdf), sorted
231 by the C-typeGen program and ^{13}C -NMR data, imported as a table (.csv) of δ_C values and
232 intensities. The users are also encouraged to provide DEPT-135 and 90 data (.csv).

233 The second tab (Figure S3) displays all the different parameters that can be adjusted:

- 234 - The tolerance (ϵ) reflects the accuracy of the used database, as the program compares each
235 chemical shift in the experimental spectrum ($\delta_{13\text{C}}$) with each chemical shift in the SDF

(δ_{SDF}) for each molecule. It considers that $\delta_{13\text{C}}$ matches with δ_{SDF} if $\delta_{\text{SDF}} - \varepsilon < \delta_{13\text{C}} < \delta_{\text{SDF}} + \varepsilon$. The default value for ε is 1.3.^{42,47}

- The tolerance incrementation can be turned ON or OFF. If ON, the program will start the matching process with $\varepsilon = 0.0$ and will then increment this value by steps of 0.1 ppm, until it reaches the ε value chosen by the user. This will match first the chemical shifts that are closest together. If this parameter is OFF, the algorithm will match a δ_{SDF} with the first $\delta_{13\text{C}}$ falling into the $\pm \varepsilon$ interval. It is recommended to leave this parameter ON, especially when using experimental DBs.
- The DEPT alignment parameter is necessary to associate a $\delta_{13\text{C}}$ to its corresponding carbon in the DEPT-135 and 90 spectra. Since the chemical shifts are never exactly the same from a spectrum to another, the user has to indicate the accuracy of the alignment in each one of his spectra, the default value is 0.02 ppm. The quality of the results will be very dependent on the quality of this alignment.
- The equivalent carbon factor can be turn ON or OFF. Turning it ON will allow the same $\delta_{13\text{C}}$ to be matched multiple times if several identical δ_{SDF} are found (equivalent carbons in the database). If OFF, a $\delta_{13\text{C}}$ can only be matched once, even if equivalent carbons are found in the database.
- The molecular weight filter will only show results if they correspond to the ones requested by the user. One can either select specific mass, or a range of values. The algorithm will search for molecules with the indicated molecular weight(s).

The third and last tab is the C-typeGen program that allows the user to create DBs compatible with the MixONat dereplication process. It sorts δ_{SDF} by carbon types.

MixONat algorithm: Matching process (Figure 1). First, the program will start sorting each carbon of the ^{13}C NMR spectrum depending on its type. This sorting is different, depending on the DEPT files provided by the user and according to the chosen DEPT alignment. If there is

no DEPT data, then the carbon types are not differentiated. If a DEPT 135 has been given, carbons exhibiting negative intensities will be considered as methylenes whereas other ones will be considered as methines or methyls. Carbons of the ^{13}C NMR spectrum not detected in DEPT experiments will be considered as quaternaries. Adding a DEPT-90 spectrum will allow to distinguish methines from methyls.

In a second step, the matching process consists, for each compound of the DB, in the comparison of $\delta_{13\text{C}}$ with δ_{SDF} . It is done by list of carbon types. All δ are first sorted by descending numerical order before the matching process. During the latter, the algorithm allows or not multiple uses of $\delta_{13\text{C}}$, depending on the equivalent carbon parameters. When all the δ_{SDF} have been considered, the score and error of the molecule are calculated and stored. The score is defined as the number of $\delta_{13\text{C}}$ matched with δ_{SDF} out of the number of carbons of the compound. The error is the cumulated absolute difference between matched signals (*i.e.* $\sum |\delta_{\text{SDF}} - \delta_{13\text{C}}|$).

MixONat software: Interactive results. At the end of the matching process, a fixed number of molecules is displayed provided they reach a minimal score set by the user. Compounds are ranked by decreasing score and increasing error, with their structure, name, molecular weight, score and error (Figure S4). On the structure, matched carbons are highlighted in red. For each molecule, it is possible to open a window showing the numbered structure with matched δ_{SDF} . Matched δ_{SDF} and intensities are also shown along as a graphical representation of a reconstructed ^{13}C -NMR spectrum. On the latter, carbon types have different colors and signals are numbered according to the structure whilst the chemical shift list displayed. This easily shows if intensities of matched carbons are homogenous, hence hypothetically being signals belonging to the same molecule (Figure S5). It is thus possible for the user to link the information gathered from the structure, the spectrum and the chemical shifts and eventually decide to remove or add chemical shifts, then modifying the score of the selected molecule.

286 This function can be, for example, used to remove a carbon matched with an abnormal intensity
287 (meaning it probably belongs to another molecule), or to add a quaternary carbon that was not
288 matched because it was not picked on the spectrum due to its low intensity, or predicted a bit
289 too far away, etc.⁴⁸ If a carbon is added or removed, the spectrum and the highlighted carbons
290 on the structure will be updated accordingly. It is also possible to delete a molecule altogether
291 if necessary (*e. g.* for chemotaxonomy considerations). Once the results have been checked by
292 the user, they can be saved as a text and image file (Figure S6). The text file will summarize the
293 parameters used for the research and show the information for each molecule (*i.e.* name, rank,
294 matched carbons by carbon type). The image file contains compounds structures with matched
295 carbons highlighted in red, as well as their name, molecular weight, score and error.

296 RESULTS AND DISCUSSION

297 In order to evaluate the relevance of the present ¹³C-NMR dereplication strategy using the
298 MixONat software and the pertinence of the results, we applied it to various herbal extracts. A
299 comparison was done with other methods commonly used for dereplication.

300 **Poppy alkaloids.** Poppy alkaloids are widely used all over the world for pharmaceutical
301 applications, mainly as painkillers. They are extracted from various varieties of *Papaver*
302 *somniferum*. If opium was initially the raw material, poppy straw is nowadays preferred by
303 industrials. Different chemical varieties are exploited, containing mainly either morphine,
304 thebaine, codeine, noscapine or other isoquinoline alkaloids.⁴⁹ The present ¹³C-NMR
305 dereplication process was thus evaluated for its ability to discriminate between different
306 chemotypes. The chosen extract contained papaverine and noscapine as major NPs (Figures
307 S7). At first, only δ_C from the ¹³C spectrum were considered (Figure S8) and “equivalent
308 carbons” were authorized. Using Papaveraceae DB (174 NPs), papaverine was predicted at rank
309 1 (Score 1.00) when (-)-noscapine [syn. (-)-narcotine] was suggested at the 4th position (score
310 0.77) (Figure S9). Indeed, for this alkaloid, 5 signals did not match as their chemical shifts were

311 predicted more than 1.3 ppm over their expected value in CDCl₃ (Table S1). The way the
312 algorithm works sometimes hinders matching. For example, as far as δ_{13C} at 62.5 and 61.0 ppm
313 in the extract were concerned, the algorithm started to match the higher one at $\delta_C = 62.5$ ppm
314 with the closer predicted δ_{SDF} at 62.0 ppm. Then, the second one at 61.0 ppm was too far away
315 from the predicted δ_{SDF} at 63.1 ppm. However, as MixONat software offers an interactive
316 interface, the user may correct the matching after careful analysis of results. In this example,
317 dehydrocavidine (score 0.81) and berberrubine were proposed at positions 2 and 3 respectively
318 (Figure S9). Additional DEPT-135 and 90 experiments were thus registered and used during the
319 dereplication process to discriminate carbons type. Papaverine remained at the 1st position and
320 (-)-noscapine moved to the 3rd one. Dehydrocavidine switched to rank 4 just after the latter with
321 a score decreasing to 0.71 because of 2 more unmatched carbons (Figure S10). Finally, (-)-
322 eschscholtzine appeared at rank 2. However, the authorization of “equivalent carbon” by the
323 software artificially doubled the score of this symmetrical NP. Together with chemotaxonomic
324 consideration, the user would eliminate such a hypothesis. Finally, as expected, MixONat
325 software successfully helped the user to identify papaverine and (-)-noscapine in this poppy
326 extract.

327 One can wonder about the use of predicted chemical shifts. In this example, the comparison
328 of papaverine δ_{13C} with the ones described in literature (CDCl₃) was not convincing.⁵⁰ Thus,
329 the ¹³C-NMR spectrum of a commercial reference was recorded in CDCl₃, leading to a perfect
330 matching (Table S2).

331 **Rosemary phenols.** Rosemary leaf is a medicinal plant traditionally used to relieve symptoms
332 of dyspepsia and treat mild spasmodic disorders of the gastrointestinal tract.⁵¹ According to the
333 European pharmacopeia, it contains more than 3% of hydroxycinnamic derivatives such as
334 rosmarinic acid. Triterpenes and tricyclic phenolic diterpenes (*e.g.* carnosic acid and carnosol)
335 are also described as major NPs. The latter are found in large amount in extracts used as food

336 antioxidants (*i.e.* E392 additive).³³ After successive extraction of *Rosmarinus officinalis* leaf by
337 DCM and MeOH, their major NPs were investigated using the Lamiaceae DB, ¹³C-NMR
338 spectrum, DEPT-135 and 90 experiments and the MixONat software. “Equivalent carbons”
339 were authorized. Results were compared with LC-ELSD-DAD-MSⁿ data.

340 The MeOH extract contained rosmarinic acid and a disaccharide, suggested as sucrose (Figure
341 S11 and Table S3). Out of 982 compounds in the DB, the ¹³C-NMR dereplication process
342 predicted rosmarinic acid (score 0.83) and sucrose (score 0.83) at ranks 4 and 5 respectively
343 (Figures S12-13). It should be noted that rosmarinic acid is an ester including a caffeic acid
344 moiety, predicted at rank 1. The NP proposed in position 3 is an ester of caffeic acid and sucrose.
345 3 δ_{13C} of rosmarinic acid were not matched by the algorithm. First, the methine at 117.1 ppm
346 was set aside due to the way the algorithm works, as explained earlier. However, the user might
347 correct the matching after ¹³C-NMR spectrum examination and associate the signals at 115.1,
348 115.2, 116.2, 116.4 and 117.5 together with the predicted ones at 115.1, 116.2, 116.5, 117.1,
349 and 117.5. Then, the signal at 129.3 ppm was a little bit too far from its prediction at 130.6 ppm.
350 The use of the interactive interface would also correct this to reach as score of 0.94 and the 2nd
351 rank. As far as the predicted δ_{SDF} at 173.8 was concerned, it was absent from the spectra (Figure
352 S13), probably due to a matrix effect, as already observed.⁵² Finally, with the help of the
353 MixONat software, the user would easily identify the two major compounds of this extract.

354 Both the rosemary DCM extract and the so-called E392 antioxidant additive share similar
355 major metabolites, *i.e.* the triterpenes ursolic, oleanolic, betulinic and micromeric acids and the
356 phenolic diterpenes carnosic acid and its derivatives (Figure S14-15 and S19, Table S4). Indeed,
357 while triterpenes are stable NPs, carnosic acid will spontaneously oxidized in carnosol and
358 degradation products,⁵³ impacting the anti-oxidant activity of the whole extract.⁵⁴ Thus, the
359 present ¹³C-NMR dereplication process was investigated as a mean to predict the presence of
360 carnosic acid or its degradation products.

361 The E392 extract obtained by ethanolic extraction of *R. officinalis* leaves after a delipidating
 362 step contains mainly carnosic acid together with the previously cited triterpenes (Figure S14).
 363 As Lamiaceae plants (*e.g.* mint, lavender, sage, thyme) are well-known to biosynthesize mono-
 364 and sesquiterpenes from essential oils, such low molecular weight NPs constitute a large part
 365 of the Lamiaceae DB. To focus on the non-volatile NPs from the rosemary dry extracts, a
 366 molecular weight filter was used and only compounds beyond 250 Da were selected. Amongst
 367 the 982, MixONat predicted the presence of betulinic, ursolic, micromeric, oleanolic and
 368 carnosic acids in ranks 4, 6, 11, 17 and 22 respectively, with scores ranging from 0.90 to 0.75
 369 (Figure S16). Among the first 24 suggested NPs, 10 are triterpenes from either ursane, lupane
 370 or oleanane types. At ranks 1-3, were respectively suggested (+)-lupeol, viminalol and (+)-
 371 betulin whose structure are very close to betulinic and ursolic acids. 10 are diterpenes, 6 of them
 372 sharing the same abietane skeleton as carnosic acid and its derivatives. At this stage, the user
 373 can opt for a careful comparison between E392 δ_C and literature data for these first ranked NPs.
 374 However, LC-MS² analysis of E392 gave us the molecular weight of the major NPs, *i.e.* 330,
 375 332, 454 and 456 Da (Table S4). Finally, using these specific values as a filter, MixONat
 376 software managed to successfully identify betulinic acid (score 0.9), ursolic acid (score 0.87),
 377 micromeric acid (score 0.87), oleanolic acid (score 0.83) and carnosic acid (score 0.75) in the
 378 first 5 ranks (Figure S17). None of the compounds has reached a score of 1. Proposed in the
 379 first position, betulinic acid (C₃₀) had 2 carbons that were incorrectly predicted ($\Delta\delta_C > 1.3$ ppm):
 380 C-3 (78.4) and C-19 (48.6 ppm). The quaternary C-4 was predicted at 38.1 ppm but its signal
 381 was hidden by the ones from C-1 and C-13. C-3 and C-4 were not matched neither in ursolic,
 382 micromeric and oleanolic acids for the same raisons. For the latter, other unmatched δ_C were
 383 either due to missing signals or inaccurate predictions (Table S5). Back to the antioxidant
 384 diterpenes, carnosic acid reached position 5 because, as a minor compound, signal for
 385 quaternaries C-9, C-11 and C-12 were missing. Furthermore, the δ_{SDF} of C-14 and C-20 were

386 overpredicted (> 1.3 ppm). Nevertheless, it came in ahead the DB's diterpenes, *i.e.* 6,7-
387 dehydrocarnosic acid (rank 6, 0.7), carnosol (rank 8, 0.6) or carnosic acid quinone (rank 10,
388 0.6) demonstrating that this process allows to decipher carnosic acid, carnosol or their
389 degradation products in such a complex mixture.

390 The same method applied on the rosemary DCM extract containing the same triterpenes but
391 carnosol instead of carnosic acid (Figure S19, Table S4) led to the same conclusions as carnosol
392 reached rank 1 (Score 0.95, Figure S20) followed by the triterpenes. Only C-13, predicted at
393 135.1 ppm, was not matched with δ_{13C} at 134.3 ppm in the spectrum as the latter was previously
394 suggested as C-8 (δ_{SDF} 133.6 ppm). But both C-8 and C-13 chemical shifts (δ_{13C} 131.6 and
395 134.3 ppm) appeared in the ^{13}C -NMR spectrum (Figure S21, Table S6).

396 **Mangosteen peel xanthenes.** Regarding the former example, one can wonder if a
397 fractionation step is required before ^{13}C -NMR dereplication complex mixtures, including
398 compounds exhibiting similar backbones. Thus, in the framework of a project that aimed at
399 discovering new Unfolded Protein Response (UPR) modulators,⁵⁵⁻⁵⁸ the methodology was
400 finally applied to a *G. mangostana* fruit peel apolar extract together with one fraction. This
401 plant is well-known to contain bioactive prenylated xanthenes. Indeed, one LC-ELSD analysis
402 revealed α -mangostin, γ -mangostin, gartanin, garcinone E and 8-deoxygartanin as major NPs
403 in the cyclohexanic crude extract whereas gudraxanthone, 9-hydroxycalabaxanthone and β -
404 mangostin appeared as minor products (Figure S22). The ^{13}C -NMR dereplication process was
405 then undertaken this extract using the aforementioned *Garcinia* DB. Equivalent carbons were
406 allowed. As a result, among the first 25 suggestions out of the 718 NPs of the DB, 23
407 compounds were xanthenes bearing at least one prenylated side chain (Figure S23). All these
408 compounds received a score higher than 0.9. It should be noted that all these proposals share
409 either the 1,3-dihydroxy,2-prenylxanthone scaffold or the 1,3,6,7-tetrahydroxy,8-
410 prenylxanthone core of the major products α -mangostin (68% of the extract, Figures S22-S23).

411 The latter reached the 1st position as all the predicted δ_{SDF} were matched with a δ_{C} by the
 412 MixONat software (Table S7). Then garcinone E, γ -mangostin, gartanin and 8-deoxygartanin
 413 were suggested in position 4, 10, 11 (score 0.96) and 20 (score 0.93) respectively. For all of
 414 them, the predicted δ_{SDF} that were not matched were either missing quaternary carbons (C-7 of
 415 garcinone E) or a CH not picked in the DEPT-90 experiment (C-2'' of γ -mangostin, C-7 of
 416 gartanin, C-6 and C-8 of 8-deoxygartanin) (Table S7 and Figure S25). Indeed, during the
 417 methines peak picking process, the threshold was set as to avoid signals due to the methyl
 418 groups (*i.e.* OMe at 62.2 ppm) of the major α -mangostin. This exemplifies the limitation of the
 419 present ^{13}C -NMR method to dereplicate the minor NPs in crude extracts. A way to circumvent
 420 such issue would consist in a coarse fractionation of complex crude extracts. The mangosteen
 421 peel extract was thus fractionated and the dereplication process applied again on a fraction
 422 containing minor xanthenes, *i.e.* gartanin, 8-deoxygartanin, β -mangostin and gudraxanthone
 423 (Figure S26). After such a concentration step, ^{13}C -NMR dereplication showed that 8-
 424 deoxygartanin and β -mangostin reached the two first position with a perfect match while
 425 gartanin and gudraxanthone were suggested at positions 10 and 24 respectively (score 0.96 and
 426 0.89, Figure S27). Concerning gartanin, its quaternary C-8a predicted at 107.8 ppm was not
 427 matched (δ_{C} 109.3 ppm). With the same causes producing the same effects, C-5 and C-8a of
 428 gudraxanthone predicted at 146.4 and 122.7 ppm from experimental data in DMSO-d₆ were
 429 not matched either. Finally, a DB of experimental chemical shifts, namely CH-NMR-NP DB
 430 was also used.⁴⁴ Amongst 32854 entries, β -mangostin, 8-deoxygartanin and gartanin reached
 431 the three first position with a perfect match whilst gudraxanthone was not suggested (Figure
 432 S28). The latter was suggested in the position 122 with a score of 0.79. Indeed, the NMR was
 433 described in DMSO-d₆ only^{34,59}, inducing discrepancies in δ_{C} values. This demonstrates,
 434 however, that when DBs containing NPs of interest and their experimental chemical shifts are
 435 available, their use obviously increases chances for better matches. Even if the MixONat

436 program gives satisfactory results with predicted DBs, dereplication based on NMR would then
437 greatly benefit from a repository of publicly accessible raw NMR data for all published NPs.^{60,}

438 ⁶¹

439

440 CONCLUSION

441 To decipher complex mixtures using ¹³C-NMR data, we propose here the freely available
442 software MixONat, a complementary tool to those already existing, notably based on LC-MS
443 profilings. Although initiated at the same time as MS methodologies, ¹³C-NMR-based
444 dereplications remained less popular, probably due to their lower sensitivity. However, NMR
445 spectrometers nowadays provide valuable dataset within a reasonable amount of time on
446 quantities of the order of 1-10 mg. Trying to integrate the best of former works,^{9, 28, 30, 32} the
447 present methodology allows to take into account the type of carbon through DEPT experiments
448 and classifies compounds from a specific DB according to decreasing scores. Signal intensities
449 are eventually monitored through an interactive interface. The MixONat software also requires
450 DBs that may contain predicted δ_{SDF} . Therefore, in the event of pre-profiling analyses, neither
451 reference compounds nor published data are initially required.

452 For all analyzed mixtures, MixONat suggested the correct NPs in first ranks. Even if one exact
453 structure is not ranked first every time, the program presents coherent results that rapidly direct
454 the user towards a particular structural type. Moreover, the software is able to distinguish
455 structurally close NPs, including stereoisomers. Then, manual comparison of the best-ranked
456 hypotheses with literature data may confirm identifications. Most of the time, no more than the
457 twenty first compounds need to be checked, out of several hundreds of molecules present in the
458 DBs. Interactive results greatly facilitate the work of finding the right NPs. Finally, one can
459 imagine the interest to associate ¹³C-NMR data and MixONat software to LC-MS² data to assert
460 the major compound of a mixture, including stereoisomers.

461

462 References

- 463 1. Wolfender, J.-L.; Litaudon, M.; Touboul, D.; Queiroz, E. F. *Nat. Prod. Rep.* **2019**, *36* (6),
464 855-868.
- 465 2. Hubert, J.; Nuzillard, J.-M.; Renault, J.-H. *Phytochem. Rev.* **2015**, *16* (1), 55-95.
- 466 3. Singh, U.; Baishya, B. *J. Magn. Reson.* **2019**, *301*, 19-29.
- 467 4. Kumar, N.; Devineni, S. R.; Singh, G.; Kadirappa, A.; Dubey, S. K.; Kumar, P. *J. Pharm.*
468 *Biomed. Anal.* **2016**, *119*, 114-121.
- 469 5. Beutler, J. A.; Alvarado, A. B.; Schaufelberger, D. E.; Andrews, P.; McCloud, T. G. *J. Nat.*
470 *Prod.* **1990**, *53* (4), 867-74.
- 471 6. Robinette, S. L.; Brüscheweiler, R.; Schroeder, F. C.; Edison, A. S. *Acc. Chem. Res.* **2012**,
472 *45* (2), 288-297.
- 473 7. Gaudencio, S. P.; Pereira, F. *Nat. Prod. Rep.* **2015**, *32* (6), 779-810.
- 474 8. Allard, P. M.; Peresse, T.; Bisson, J.; Gindro, K.; Marcourt, L.; Pham, V. C.; Roussi, F.;
475 Litaudon, M.; Wolfender, J. L. *Anal. Chem.* **2016**, *88* (6), 3317-23.
- 476 9. Bakiri, A.; Hubert, J.; Reynaud, R.; Lanthony, S.; Harakat, D.; Renault, J. H.; Nuzillard,
477 J. M. *J. Nat. Prod.* **2017**, *80* (5), 1387-1396.
- 478 10. Le Pogam, P.; Schinkovitz, A.; Legouin, B.; Le Lamer, A.-C.; Boustie, J.; Richomme, P.
479 *Anal. Chem.* **2015**, *87* (20), 10421-10428.
- 480 11. Schinkovitz, A.; Boisard, S.; Freuze, I.; Osuga, J.; Mehlmer, N.; Brück, T.; Richomme,
481 P. *Anal. Bioanal. Chem.* **2018**, *410* (24), 6187-6195.
- 482 12. Wang, M.; Carver, J. J.; Phelan, V. V.; Sanchez, L. M.; Garg, N.; Peng, Y.; Nguyen, D.
483 D.; Watrous, J.; Kapono, C. A.; Luzzatto-Knaan, T.; Porto, C.; Bouslimani, A.; Melnik, A.
484 V.; Meehan, M. J.; Liu, W.-T.; Crisemann, M.; Boudreau, P. D.; Esquenazi, E.; Sandoval-
485 Calderón, M.; Kersten, R. D.; Pace, L. A.; Quinn, R. A.; Duncan, K. R.; Hsu, C.-C.; Floros,
486 D. J.; Gavilan, R. G.; Kleigrew, K.; Northen, T.; Dutton, R. J.; Parrot, D.; Carlson, E. E.;
487 Aigle, B.; Michelsen, C. F.; Jelsbak, L.; Sohlenkamp, C.; Pevzner, P.; Edlund, A.; McLean,
488 J.; Piel, J.; Murphy, B. T.; Gerwick, L.; Liaw, C.-C.; Yang, Y.-L.; Humpf, H.-U.; Maansson,
489 M.; Keyzers, R. A.; Sims, A. C.; Johnson, A. R.; Sidebottom, A. M.; Sedio, B. E.; Klitgaard,
490 A.; Larson, C. B.; Boya, P. C. A.; Torres-Mendoza, D.; Gonzalez, D. J.; Silva, D. B.; Marques,
491 L. M.; Demarque, D. P.; Pociute, E.; O'Neill, E. C.; Briand, E.; Helfrich, E. J. N.; Granatosky,
492 E. A.; Glukhov, E.; Ryffel, F.; Houson, H.; Mohimani, H.; Kharbush, J. J.; Zeng, Y.; Vorholt,
493 J. A.; Kurita, K. L.; Charusanti, P.; McPhail, K. L.; Nielsen, K. F.; Vuong, L.; Elfeki, M.;
494 Traxler, M. F.; Engene, N.; Koyama, N.; Vining, O. B.; Baric, R.; Silva, R. R.; Mascuch, S.
495 J.; Tomasi, S.; Jenkins, S.; Macherla, V.; Hoffman, T.; Agarwal, V.; Williams, P. G.; Dai, J.;
496 Neupane, R.; Gurr, J.; Rodríguez, A. M. C.; Lamsa, A.; Zhang, C.; Dorrestein, K.; Duggan,
497 B. M.; Almaliti, J.; Allard, P.-M.; Phapale, P.; Nothias, L.-F.; Alexandrov, T.; Litaudon, M.;
498 Wolfender, J.-L.; Kyle, J. E.; Metz, T. O.; Peryea, T.; Nguyen, D.-T.; VanLeer, D.; Shinn,
499 P.; Jadhav, A.; Müller, R.; Waters, K. M.; Shi, W.; Liu, X.; Zhang, L.; Knight, R.; Jensen,
500 P. R.; Pálsson, B. Ø.; Poglian, K.; Linington, R. G.; Gutiérrez, M.; Lopes, N. P.; Gerwick,
501 W. H.; Moore, B. S.; Dorrestein, P. C.; Bandeira, N. *Nature Biotechnol.* **2016**, *34*, 828.
- 502 13. Yang, J. Y.; Sanchez, L. M.; Rath, C. M.; Liu, X.; Boudreau, P. D.; Bruns, N.; Glukhov,
503 E.; Wodtke, A.; de Felicio, R.; Fenner, A.; Wong, W. R.; Linington, R. G.; Zhang, L.;
504 Debonsi, H. M.; Gerwick, W. H.; Dorrestein, P. C. *J. Nat. Prod.* **2013**, *76* (9), 1686-1699.
- 505 14. Fox Ramos, A. E.; Le Pogam, P.; Fox Alcover, C.; Otogo N'Nang, E.; Cauchie, G.; Hazni,
506 H.; Awang, K.; Bréard, D.; Echavarren, A. M.; Frédérick, M.; Gaslonde, T.; Girardot, M.;
507 Grougnet, R.; Kirillova, M. S.; Kritsanida, M.; Lémus, C.; Le Ray, A.-M.; Lewin, G.;

508 Litaudon, M.; Mambu, L.; Michel, S.; Miloserdov, F. M.; Muratore, M. E.; Richomme-
 509 Peniguel, P.; Roussi, F.; Evanno, L.; Poupon, E.; Champy, P.; Beniddir, M. A. *Sci. Data* **2019**,
 510 6 (1), 15.
 511 15. Global Natural Products Social Molecular Networking, <https://gnps.ucsd.edu/> (Accessed
 512 July 2019).
 513 16. Marti, G.; Eparvier, V.; Moretti, C.; Susplugas, S.; Prado, S.; Grellier, P.; Retailleau, P.;
 514 Gueritte, F.; Litaudon, M. *Phytochemistry* **2009**, 70 (1), 75-85.
 515 17. Přichystal, J.; Schug, K. A.; Lemr, K.; Novák, J.; Havlíček, V. *Anal. Chem.* **2016**, 88 (21),
 516 10338-10346.
 517 18. Vignoli, A.; Ghini, V.; Meoni, G.; Licari, C.; Takis, P. G.; Tenori, L.; Turano, P.; Luchinat,
 518 C. *Angew. Chem. Int. Ed.* **2019**, 58 (4), 968-994.
 519 19. Simmler, C.; Graham, J. G.; Chen, S.-N.; Pauli, G. F. *Fitoterapia* **2018**, 129, 401-414.
 520 20. Nardella, F.; Margueritte, L.; Lamure, B.; Viéville, J. M. P.; Bourjot, M. *Phytochem. Lett.*
 521 **2018**, 26, 138-142.
 522 21. Guennec, A. L.; Giraudeau, P.; Caldarelli, S. *Anal. Chem.* **2014**, 86 (12), 5946-54.
 523 22. Clendinen, C. S.; Lee-McMullen, B.; Williams, C. M.; Stupp, G. S.; Vandenborne, K.;
 524 Hahn, D. A.; Walter, G. A.; Edison, A. S. *Anal. Chem.* **2014**, 86 (18), 9242-50.
 525 23. Clendinen, C. S.; Stupp, G. S.; Ajredini, R.; Lee-McMullen, B.; Beecher, C.; Edison, A.
 526 S. *Front. Plant Sci.* **2015**, 6, 611.
 527 24. The Human Metabolome Database, <http://www.hmdb.ca/> (Accessed October 2019).
 528 25. Biological Magnetic Resonance Data Bank, <http://www.bmrb.wisc.edu/> (Accessed October
 529 2019).
 530 26. Hostettmann, K.; Wolfender, J.-L.; Rodriguez, S. *Planta Med.* **1997**, 63 (1), 2-10.
 531 27. Formacek, V.; Kubeczka, K. H.; ¹³C-NMR analysis of essential oils. In *Aromatic Plants:*
 532 *Basic and Applied Aspects*, Margaris, N.; Koedam, A.; Vokou, D., Eds. Springer Netherlands:
 533 Dordrecht, 1982; pp 177-181.
 534 28. Tomi, F.; Bradesi, P.; Bighelli, A.; Casanova, J. *Magn. Reson. Anal.* **1995**, 1, 25-34.
 535 29. Fromanteau, D. L. G.; Gastmans, J.-P.; Vestri, S. A.; De P. Emerenciano, V.; Borges, J. H.
 536 G. *Comput. Chem.* **1993**, 17 (4), 369-378.
 537 30. Ferreira, M. J. P.; Costantin, M. B.; Sartorelli, P.; Rodrigues, G. V.; Limberger, R.;
 538 Henriques, A. T.; Kato, M. J.; Emerenciano, V. P. *Anal. Chim. Acta* **2001**, 447 (1-2), 125-134.
 539 31. Dictionary of Natural Products, <http://dnp.chemnetbase.com/faces/chemical/ChemicalSearch.xhtml> (accessed March 2019).
 540 32. Hubert, J.; Nuzillard, J. M.; Purson, S.; Hamzaoui, M.; Borie, N.; Reynaud, R.; Renault,
 541 J. H. *Anal. Chem.* **2014**, 86 (6), 2955-62.
 542 33. EFSA. *EFSA journal* **2008**, 721, 1-29.
 543 34. Ryu, H. W.; Curtis-Long, M. J.; Jung, S.; Jin, Y. M.; Cho, J. K.; Ryu, Y. B.; Lee, W. S.;
 544 Park, K. H. *Bioorg. Med. Chem.* **2010**, 18 (17), 6258-64.
 545 35. Suksamrarn, S.; Suwannapoch, N.; Ratananukul, P.; Aroonlerk, N.; Suksamrarn, A. *J Nat.*
 546 *Prod.* **2002**, 65 (5), 761-3.
 547 36. Groweiss, A.; Cardellina, J. H.; Boyd, M. R. *J. Nat. Prod.* **2000**, 63 (11), 1537-9.
 548 37. Likhitwitayawuid, K.; Phadungcharoen, T.; Krungkrai, J. *Planta Med.* **1998**, 64 (1), 70-2.
 549 38. PerkinElmer ChemDraw,
 550 [http://www.cambridgesoft.com/Ensemble_for_Chemistry/ChemDraw/ChemDrawProfessional](http://www.cambridgesoft.com/Ensemble_for_Chemistry/ChemDraw/ChemDrawProfessional/Default.aspx)
 551 [/Default.aspx](http://www.cambridgesoft.com/Ensemble_for_Chemistry/ChemDraw/ChemDrawProfessional/Default.aspx) (accessed March 2019).
 552 39. Softonic ChemSketch, <https://chemsketch.fr.softonic.com/> (accessed March 2019).
 553 40. CAS SciFinder, <https://www.cas.org/products/scifinder> (accessed March 2019).
 554 41. ACD/Labs NMR Spectroscopy Software, <https://www.acdlabs.com/products/adh/nmr/>
 555 (accessed March 2019).
 556

557 42. Bruguère, A.; Derbré, S.; Coste, C.; Le Bot, M.; Siegler, B.; Leong, S. T.; Sulaiman, S.
 558 N.; Awang, K.; Richomme, P. *Fitoterapia* **2018**, *131*, 59-64.
 559 43. Please note that the program has been optimized for DBs created with ACD/Labs and hence
 560 may not work properly with a different type of DB.
 561 44. JEOL. Natural product NMR-DB "CH-NMR-NP", <https://www.j-resonance.com/en/nmrdb/>
 562 (Accessed October 2019).
 563 45. Rossum, G. *Python reference manual*; CWI (Centre for Mathematics and Computer
 564 Science): 1995.
 565 46. Landrum, G. An overview of the RDKit, <https://www.rdkit.org/docs/Overview.html>
 566 (accessed May 2019).
 567 47. The looseness factor can be lowered when working with experimental DBs or increased if
 568 the DB was created with a less accurate prediction software.
 569 48. On ¹³C-NMR spectra, the intensity of the ¹³C chemical shifts is roughly proportional to the
 570 the concentration of the compound that contains these carbons. In a mixture, if the concentration
 571 of species are different enough, it seems easy to differentiate major and minor compounds.
 572 However, their intensity depends also on the type of carbons (i.e. methyle, methylene, methine,
 573 quaternary): it makes more difficult to distinguish quaternary carbons from major compounds
 574 from methyle from minor ones. All this without taking into account intra- or intermolecular
 575 equivalences. On a simple mixture of 3 compounds at various concentrations, we failed to find
 576 an automated universal statistical method to discriminate δ_c from individual compound. A
 577 paramagnetic relaxation agent [i.e. Cr(Acac)] was added to the mixture: it polluted the sample
 578 without any improvement.
 579 49. International narcotics control board. Narcotic drugs,
 580 [https://www.incb.org/documents/Narcotic-Drugs/Technical-Publications/2018/INCB-](https://www.incb.org/documents/Narcotic-Drugs/Technical-Publications/2018/INCB-Narcotics_Drugs_Technical_Publication_2018.pdf)
 581 [Narcotics_Drugs_Technical_Publication_2018.pdf](https://www.incb.org/documents/Narcotic-Drugs/Technical-Publications/2018/INCB-Narcotics_Drugs_Technical_Publication_2018.pdf) (accessed March 2019).
 582 50. Gilmore, C. D.; Allan, K. M.; Stoltz, B. M. *J. Am. Chem. Soc.* **2008**, *130* (5), 1558-1559.
 583 51. European Medicines Agency. *Rosmarini folium*,
 584 <https://www.ema.europa.eu/en/medicines/herbal/rosmarini-folium> (accessed April 2019).
 585 52. Akoury, E. *Am. Res. J. Chem.* **2017**, *1* (1), 17-23.
 586 53. Zhang, Y.; Smuts, J. P.; Dodbiba, E.; Rangarajan, R.; Lang, J. C.; Armstrong, D. W. *J.*
 587 *Agric. Food Chem.* **2012**, *60* (36), 9305-9314.
 588 54. Frankel, E. N.; Huang, S.-W.; Aeschbach, R.; Prior, E. *J. Agric. Food Chem.* **1996**, *44* (1),
 589 131-135.
 590 55. Bruguère, A.; Ray, A.-M. L.; Bréard, D.; Blon, N.; Bataillé, N.; Guillemette, T.;
 591 Simoneau, P.; Richomme, P. *Planta Med.* **2016**, *82*, S1-S381.
 592 56. Li, G.; Petiwala, S. M.; Pierce, D. R.; Nonn, L.; Johnson, J. J. *PLoS One* **2013**, *8* (12),
 593 e81572.
 594 57. Xu, X. H.; Liu, Q. Y.; Li, T.; Liu, J. L.; Chen, X.; Huang, L.; Qiang, W. A.; Chen, X.;
 595 Wang, Y.; Lin, L. G.; Lu, J. J. *Sci. Rep.* **2017**, *7* (1), 10718.
 596 58. Li, G.; Petiwala, S. M.; Nonn, L.; Johnson, J. J. *Biochem. Biophys. Res. Commun.* **2014**,
 597 *453* (1), 75-80.
 598 59. Gudraxanthone is not soluble in CDCl₃, methanol-d₄ nor acetone-d₆.
 599 60. McAlpine, J. B.; Chen, S.-N.; Kutateladze, A.; MacMillan, J. B.; Appendino, G.; Barison,
 600 A.; Benididir, M. A.; Biavatti, M. W.; Bluml, S.; Boufridi, A.; Butler, M. S.; Capon, R. J.;
 601 Choi, Y. H.; Coppage, D.; Crews, P.; Crimmins, M. T.; Csete, M.; Dewapriya, P.; Egan, J.
 602 M.; Garson, M. J.; Genta-Jouve, G.; Gerwick, W. H.; Gross, H.; Harper, M. K.; Hermanto,
 603 P.; Hook, J. M.; Hunter, L.; Jeannerat, D.; Ji, N.-Y.; Johnson, T. A.; Kingston, D. G. I.;
 604 Koshino, H.; Lee, H.-W.; Lewin, G.; Li, J.; Linington, R. G.; Liu, M.; McPhail, K. L.;
 605 Molinski, T. F.; Moore, B. S.; Nam, J.-W.; Neupane, R. P.; Niemitz, M.; Nuzillard, J.-M.

606 Oberlies, N. H.; Ocampos, F. M. M.; Pan, G.; Quinn, R. J.; Reddy, D. S.; Renault, J.-H.;
607 Rivera-Chávez, J.; Robien, W.; Saunders, C. M.; Schmidt, T. J.; Seger, C.; Shen, B.;
608 Steinbeck, C.; Stuppner, H.; Sturm, S.; Taglialatela-Scafati, O.; Tantillo, D. J.; Verpoorte,
609 R.; Wang, B.-G.; Williams, C. M.; Williams, P. G.; Wist, J.; Yue, J.-M.; Zhang, C.; Xu, Z.;
610 Simmler, C.; Lankin, D. C.; Bisson, J.; Pauli, G. F. *Nat. Prod. Rep.* **2019**, *36* (1), 35-107.
611 61. NMReDATA initiative. <http://nmredata.org/> (Accessed October 2019).

612

MixONat a software for mixtures dereplication based on ^{13}C -NMR experiments

Antoine Bruguière[†], Séverine Derbré^{†*}, Joël Dietsch^{†,‡}, Jules Leguy[§], Valentine Rahier[§], Quentin Pottier[†], Dimitri Bréard[†], Soprane Suor-Cherer[†], Guillaume Viault[†], Anne-Marie Le Ray[†], Frédéric Saubion[§], Pascal Richomme^{†*}

[†] SONAS, EA921, UNIV Angers, SFR QUASAV, Faculty of Health Sciences, Dpt Pharmacy, 16 Bd Daviers, 49045 Angers cedex 01, France

[‡] JEOL Europe SAS, 1 Allée de Giverny, 78290 Croissy-sur-Seine, France

[§] LERIA, EA2645, UNIV Angers, SFR MathSTIC, Faculty of Sciences, 2 boulevard Lavoisier, 49045 Angers cedex 01, France.

The screenshot shows the 'Third tab' of the MixONat software interface. The top bar is green and contains the 'MixONat' logo and three tabs: 'Inputs', 'Parameters', and 'CTypeGen'. The main area is dark grey and contains several sections:

- SDF File:** A section with a 'Browse' button and a 'Delete' button.
- SDF no stereo file:** A section with a 'Browse' button and a 'Delete' button.
- Output directory path:** A section with a 'Browse' button.
- Output file name:** A section with a text input field and a 'Generate c.type file' button.

At the bottom, there is a 'Run' button.

Figure S1. Third tab of the MixONat software corresponds to the C-TypeGen subprogram which sort chemical shifts according to carbon types. C-TypeGen requires the original DB (.sdf) and the same one without stereochemistry. A new SDF is created with, for each compound of the DB, predicted δ_c values organized as methyl, methylene, methine or quaternary carbons.

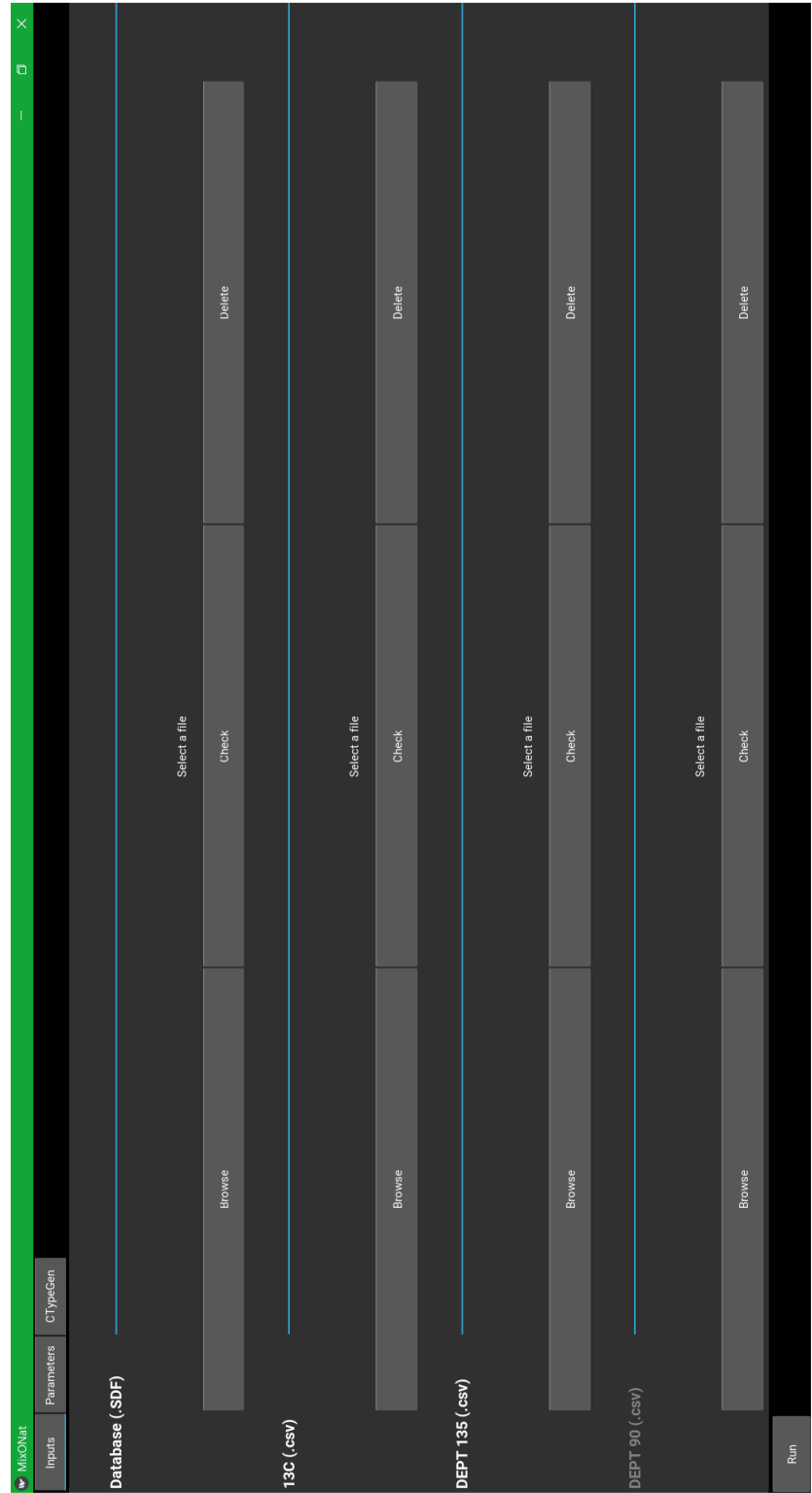


Figure S2. Home tab of the MixONat software. The input files must include at least one database (.sdf) processed by the C-typeGen routine, and ¹³C-NMR data, imported as a table (.csv) of δ_C values and intensities. DEPT-135 and 90 data can also be provided as tables (.csv) of δ_C values and intensities.

Parameter	Value
Tolerance (ppm)	1.3
Tolerance incrementation	<input checked="" type="checkbox"/>
DEPT 135 alignment	0.02
DEPT 90 alignment	0.02
Equivalent carbons	<input type="checkbox"/>
Molecular weight	MW1, MW2, MW3...
Number of results	50
Number of results per page	25
Results directory path	Select a path <input type="button" value="Browse"/>
<input type="button" value="Run"/>	

Figure S3. Second tab of the MixONat software. It displays all the different parameters that can be adjusted including the tolerance (ϵ) (Incrementation ON or OFF), the DEPT alignment parameter; the equivalent carbons option, the molecular weight filter and the minimal score required. The number of results and the directory path for saving results can also be selected in this tab.

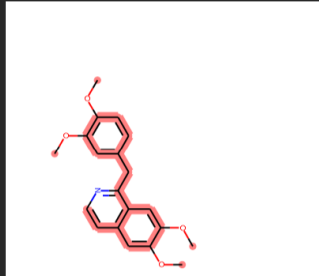
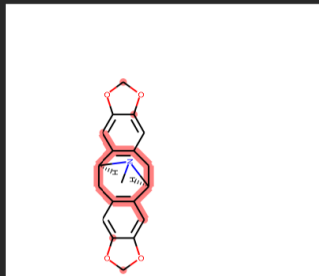
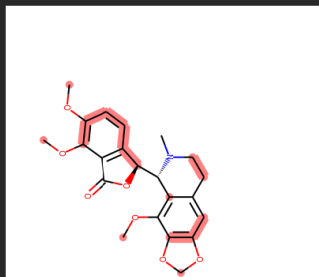
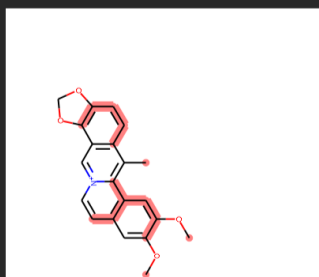
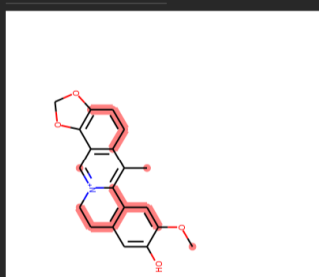
Results				
				
Rank: 1 (20/20 C) Papaverine MW: 339.45 Score match: 1.0 Error: 6.38	Rank: 2 (14/19 C) (+)-Eschscholtzine MW: 323.35 Score match: 0.74 Error: 8.2	Rank: 3 (16/22 C) (-)-Narcotine MW: 413.55 Score match: 0.73 Error: 4.52	Rank: 4 (15/21 C) Dihydroscavoline MW: 348.39 Score match: 0.71 Error: 6.59	Rank: 5 (14/20 C) Dihydroscavoline MW: 348.39 Score match: 0.7 Error: 8.77
Show shifts	Show shifts	Show shifts	Show shifts	Show shifts
Delete	Delete	Delete	Delete	Delete

Figure S4. Display of results for the ^{13}C -NMR dereplication (+ DEPT 90 and 135) of the poppy extract. Compounds of the DB are ranked by decreasing score and increasing error, with their structure, name, molecular weight, score and error. On the structure, matched carbons are highlighted in red.

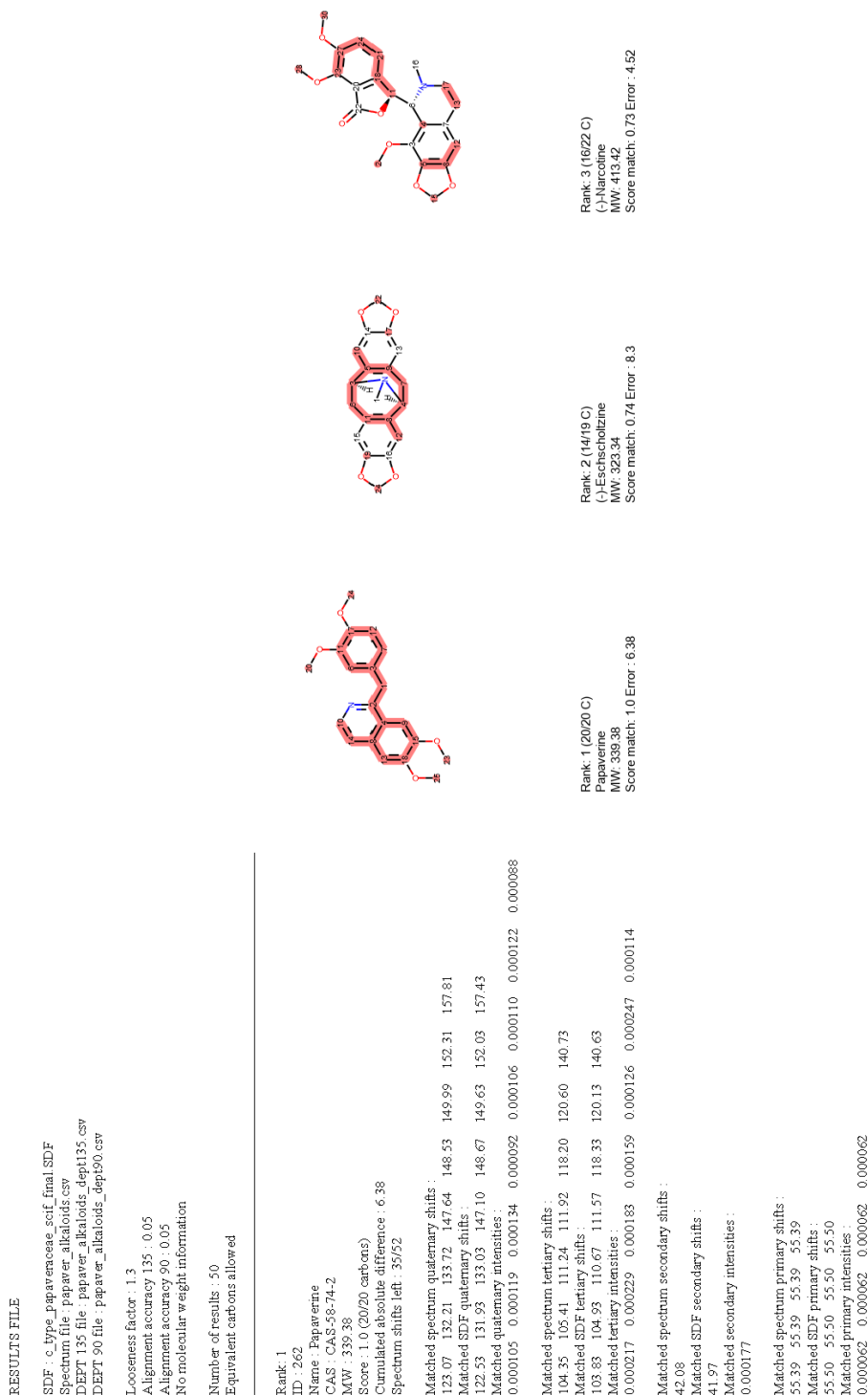


Figure S6. Part of displayed results for the ¹³C-NMR dereplication of the poppy extract (+ DEPT 90 and 135) (saved as text (Left) and image (Right) files.

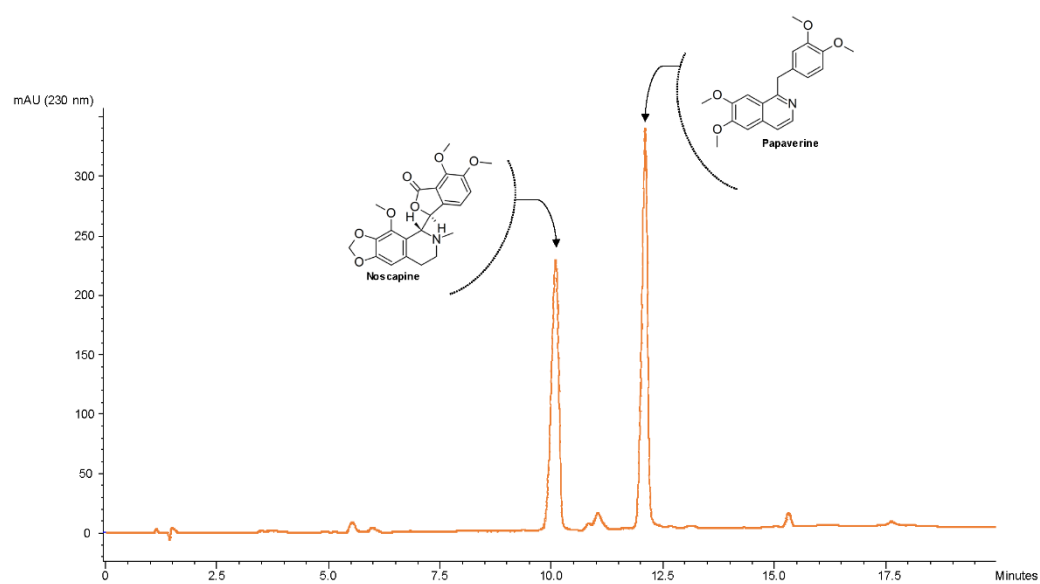


Figure S7. HPLC-UV (λ 230 nm) chromatogram of the *Papaver somniferum* alkaloids extract

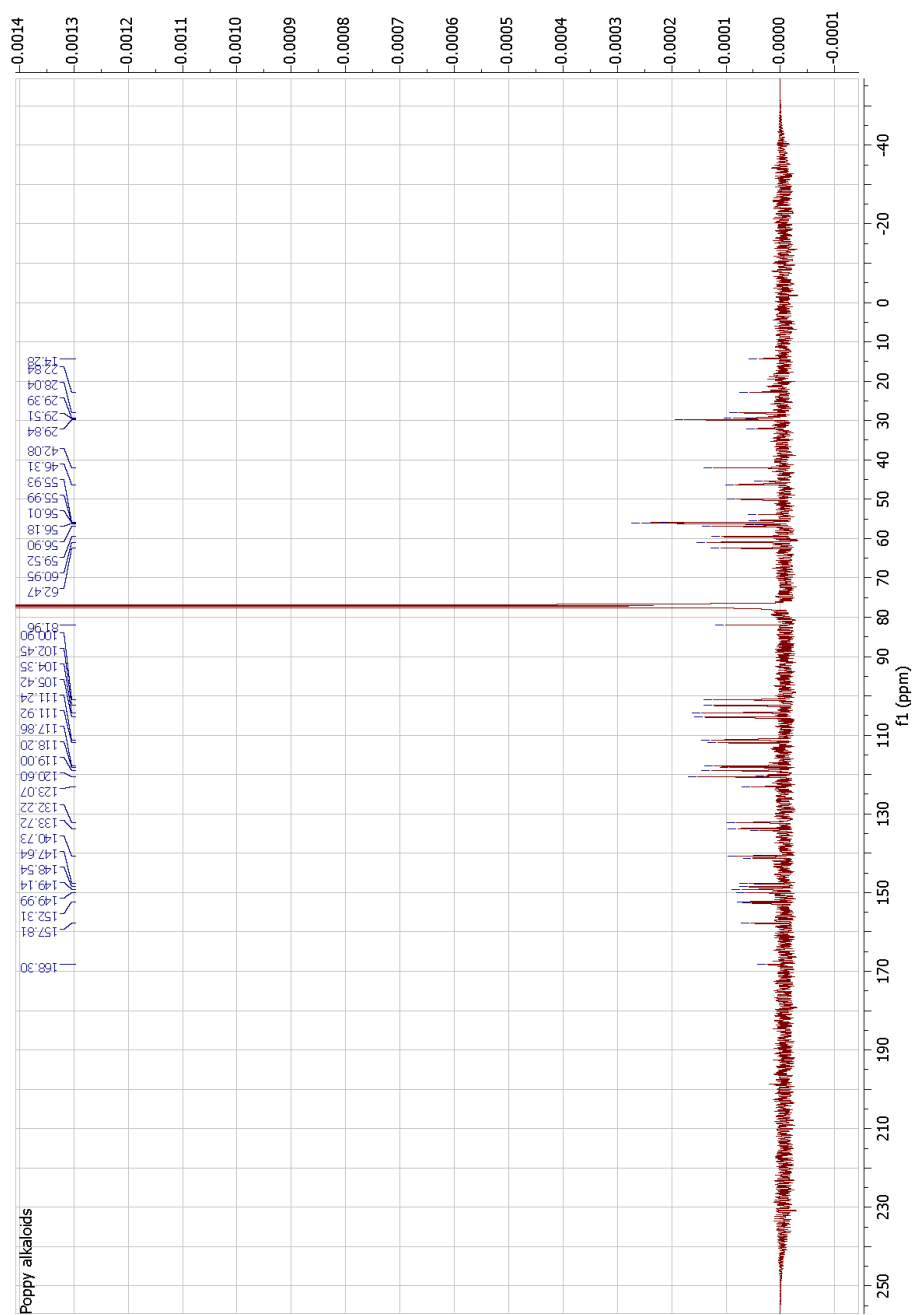


Figure S8. ^{13}C -NMR spectrum of poppy alkaloids extract recorded in CDCl_3 at 100 MHz

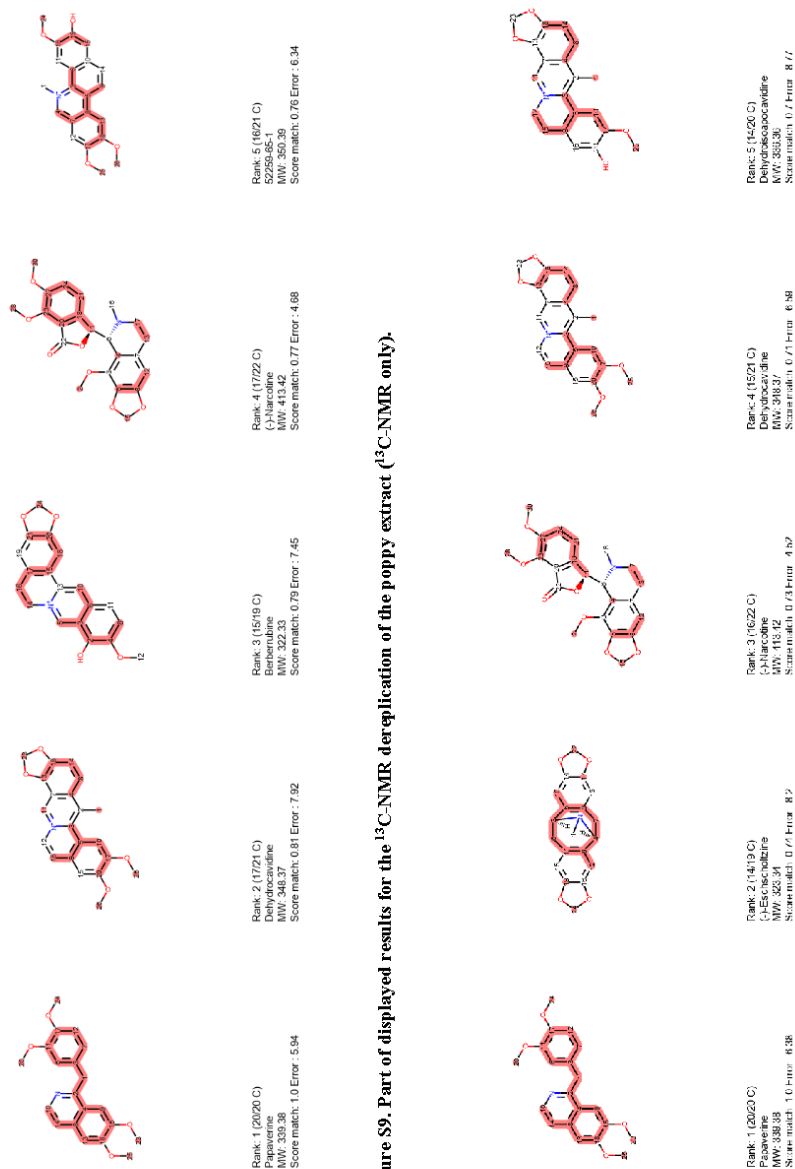


Figure S9. Part of displayed results for the ^{13}C -NMR dereplication of the poppy extract (^{13}C -NMR only).

Figure S10. Part of displayed results for the ^{13}C -NMR dereplication (+ DEPT- 90 and 135) of the poppy extract.

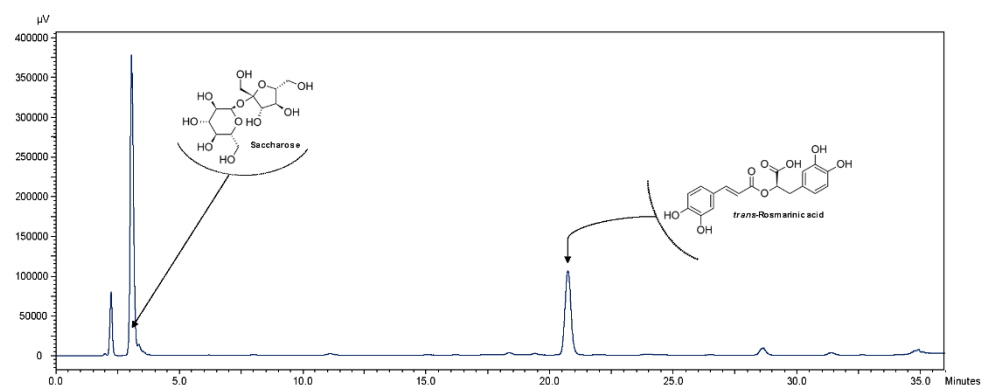


Figure S11. HPLC-ELSD chromatogram of the *Rosmarinus officinalis* MeOH extract.

RESULTS FILE

SDF : c_type_LAMIACEAE_Final.SDF
Spectrum file : Romarin MeOH.csv
DEPT 135 file : Romarin MeOH_dept135.csv
DEPT 90 file : Romarin MeOH_dept90.csv

Looseness factor : 1.3
Alignment accuracy 135 : 0.02
Alignment accuracy 90 : 0.02
No molecular weight information

Number of results : 25
Equivalent carbons not allowed

Rank: 4
ID : 969

Name : trans-Rosmarinic acid
CAS : CAS-20283-92-5
MW : 360.31

Score : 0.83 (15/18 carbons)
Cumulated absolute difference : 3.29
Spectrum shifts left : 14/29

Matched spectrum quaternary shifts :

127.82 144.88 145.95 146.70 149.43 168.92

Matched SDF quaternary shifts :

127.70 145.08 145.96 146.80 149.70 168.50

Matched quaternary intensities :

0.000325 0.000308 0.000299 0.000317 0.000258 0.000273

Not matched SDF quaternary shifts :

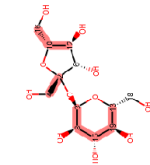
129.29 173.76

Matched spectrum tertiary shifts :

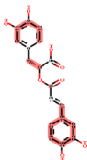
74.54 115.09 116.19 116.43 117.48 121.73 122.98 146.92

Matched SDF tertiary shifts :

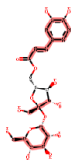
75.00 115.10 116.25 116.50 117.54 121.74 123.10 147.64



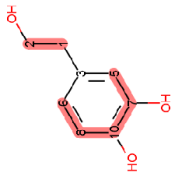
Rank: 5 (10/12 C)
D-(-)-Sucrose
MW: 342.3
Score match: 0.83 Error: 5.83



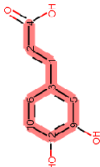
Rank: 4 (15/18 C)
trans-Rosmarinic acid
MW: 360.31
Score match: 0.83 Error: 3.29



Rank: 3 (10/21 C)
CAS: 1343877-80-6
MW: 504.44
Score match: 0.86 Error: 4.17



Rank: 2 (7/8 C)
3-Hydroxyphenol
MW: 154.16
Score match: 0.88 Error: 2.6



Rank: 1 (9/9 C)
Caffeic acid
MW: 180.16
Score match: 1.0 Error: 3.97

Matched tertiary intensities :
0.001348 0.000473 0.000491 0.000460 0.000472 0.000457 0.000437 0.000398
Not matched SDF tertiary shifts :
117.07

Matched spectrum secondary shifts :
38.51
Matched SDF secondary shifts :
37.85
Matched secondary intensities :
0.000328

Figure S12. Part of displayed results for the ^{13}C -NMR dereplication (+ DEPT 90 and 135) of the rosemary MeOH extract. Equivalent carbons were allowed.

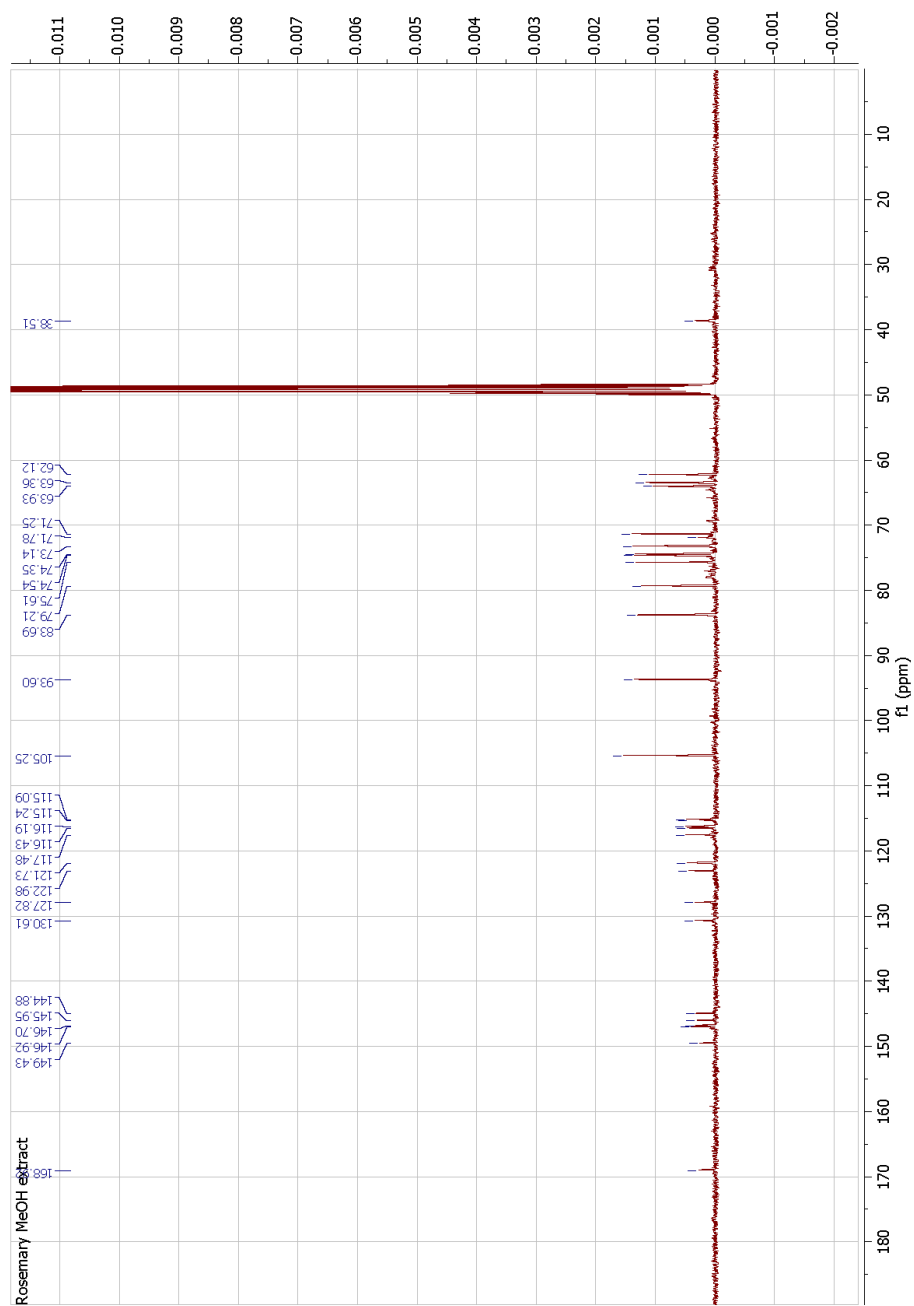


Figure S13. ^{13}C -NMR spectrum of rosemary leaf MeOH extract recorded in methanol- d_4 at 100 MHz

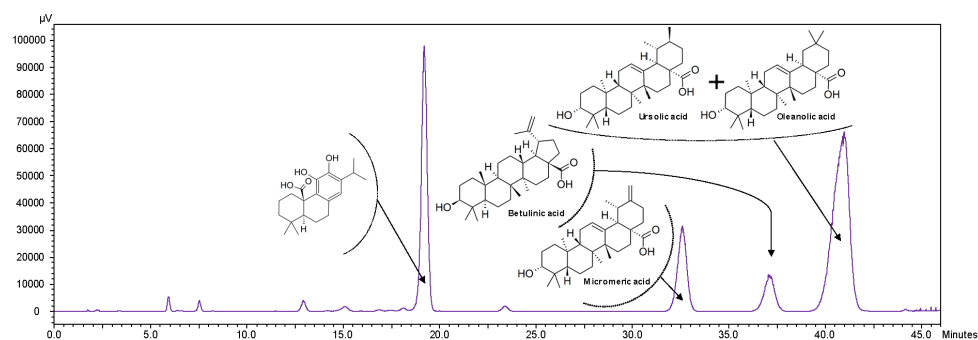


Figure S14. HPLC-ELSD chromatogram of the E392 extract.

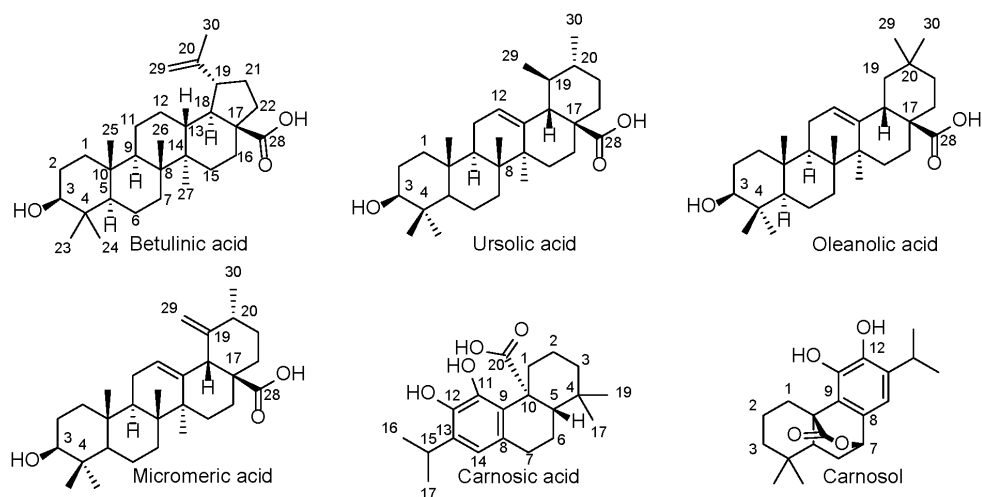


Figure S15. Structures and carbon numbering of triterpenes and carnolic acid in E392 extract

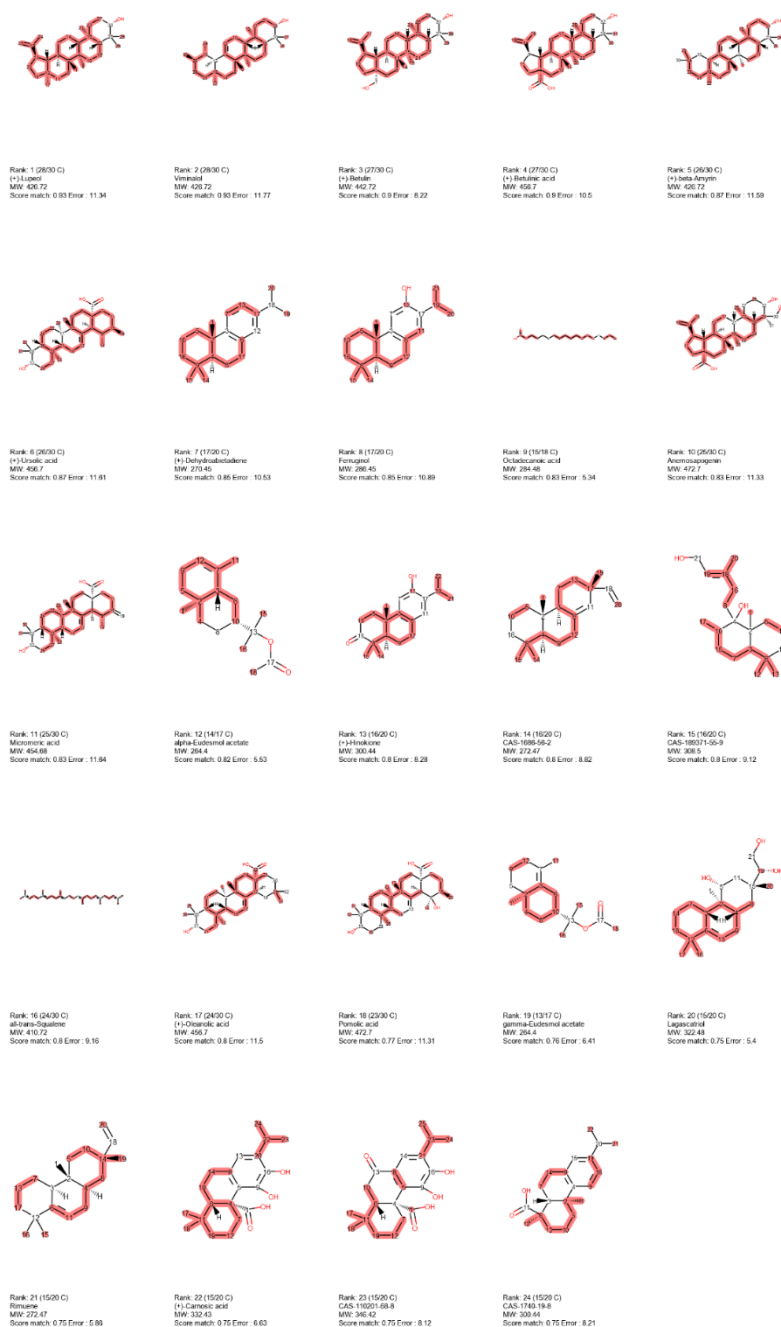


Figure S16. Part of displayed results for the ^{13}C -NMR dereplication (+DEPT 90 and 135) of the rosemary E392 extract. Equivalent carbons were allowed. A molecular weight filter (> 250 Da) was used.

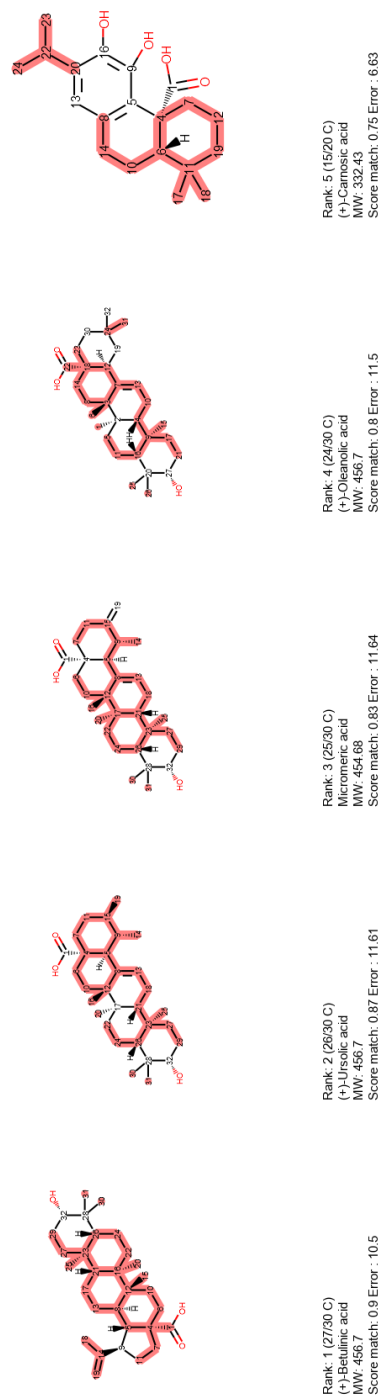


Figure SI 7. Part of displayed results for the ^{13}C -NMR dereplication (+ DEPT 90 and 135) of the rosemary E392 extract. Equivalent carbons were allowed. A molecular weight filter was applied, *i. e.* only NPs with 330, 332, 454 and 456 Da as a molecular weight were allowed.

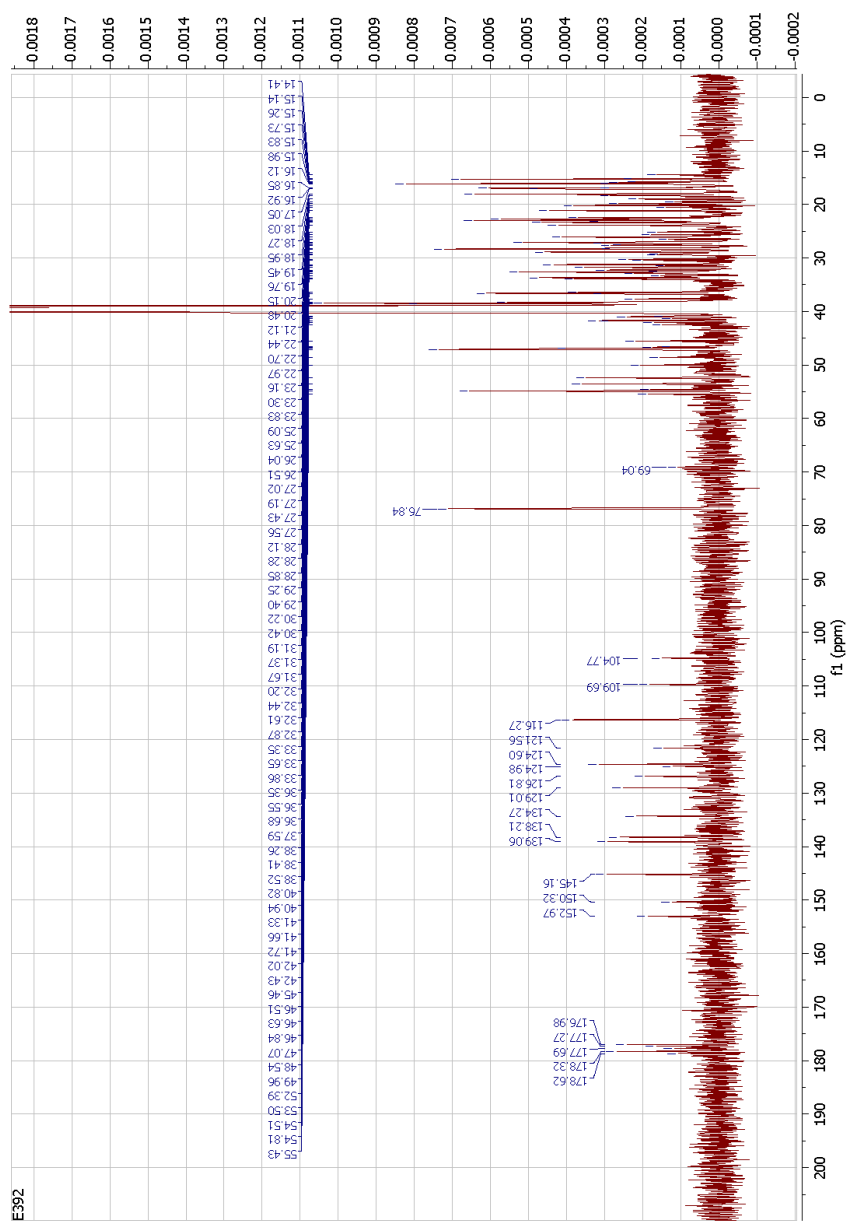


Figure S18. ^{13}C -NMR spectrum of rosemary E392 extract recorded in DMSO- d_6 at 100 MHz

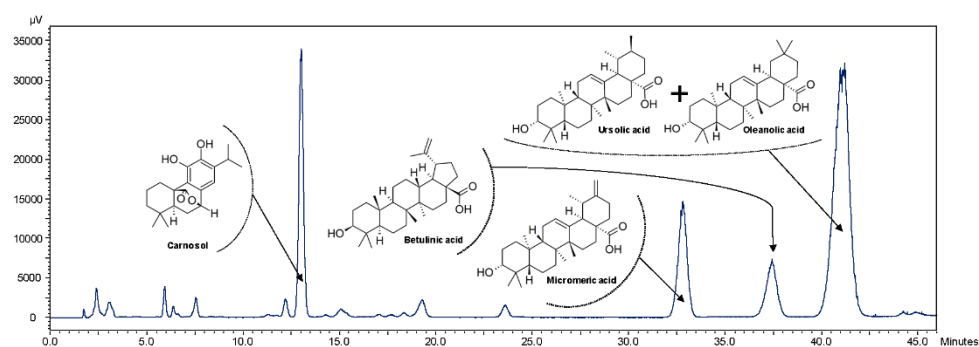


Figure S19. HPLC-ELSD chromatogram of the *Rosmarinus officinalis* DCM extract.

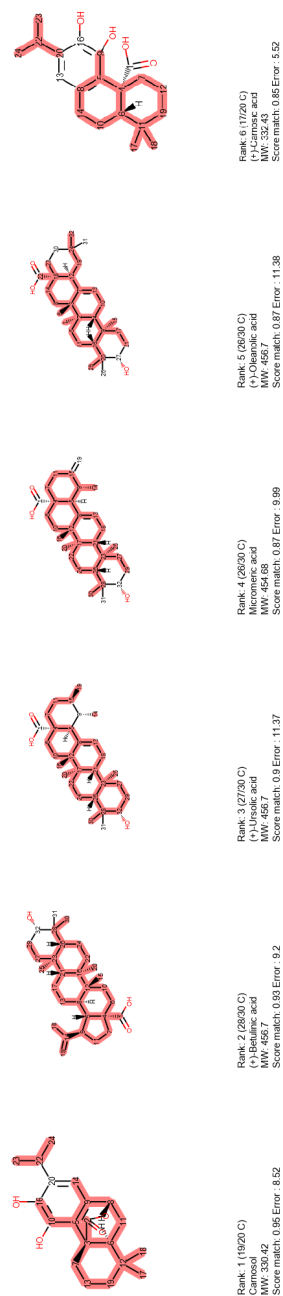


Figure S20. Part of displayed results for the ^{13}C -NMR dereplication (+ DEPT 90 and 135) of the rosemary DCM extract. Equivalent carbons were allowed. A molecular weight filter was applied, *i. e.* only NPs with 330, 332, 454 and 456 Da as a molecular weight were allowed.

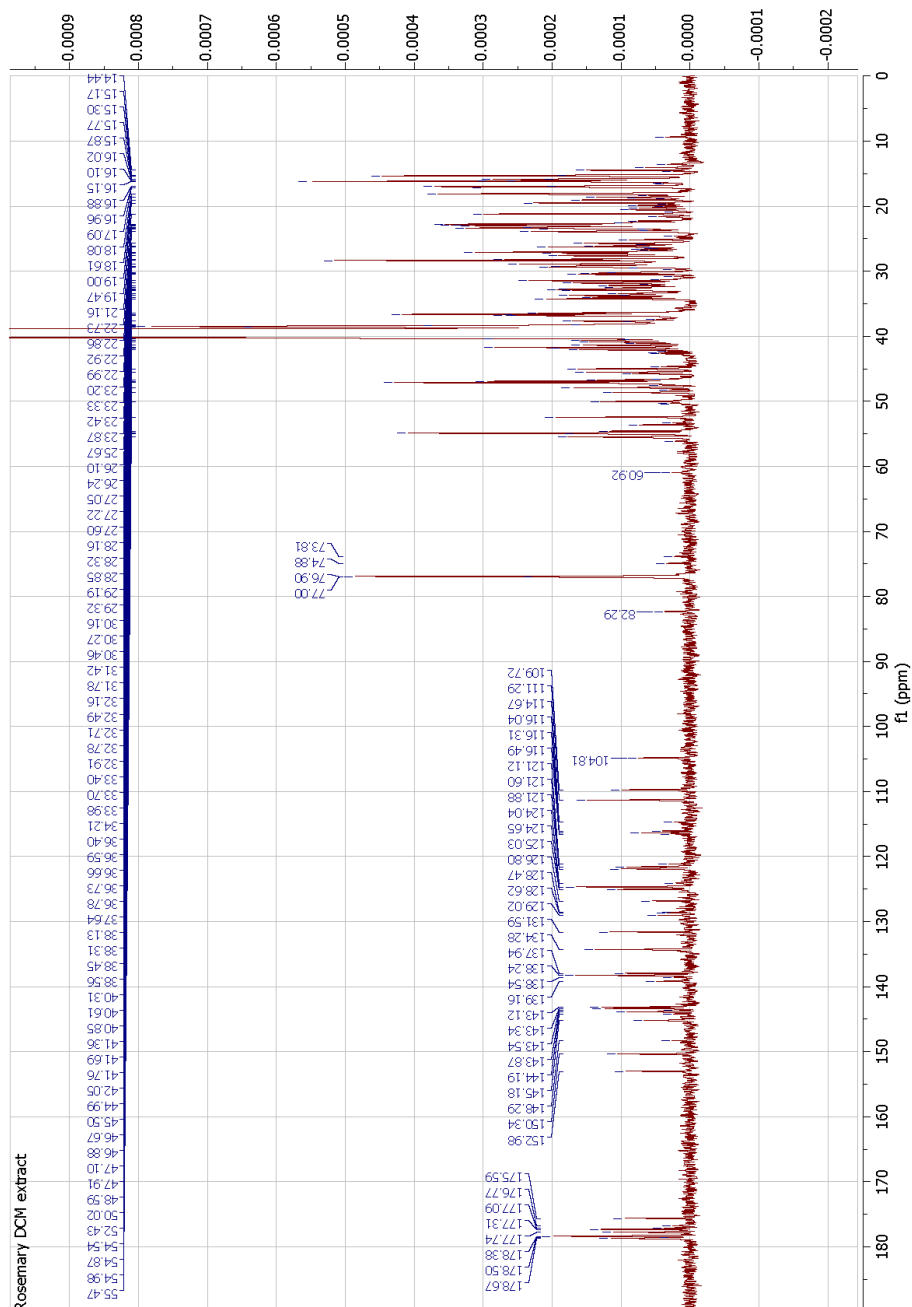
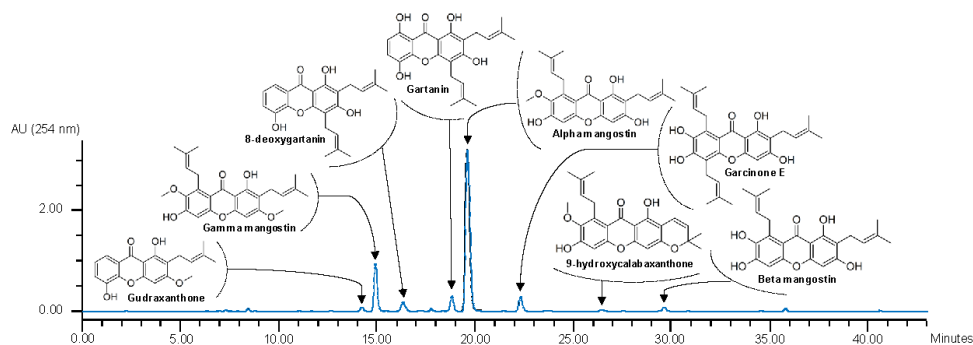


Figure S21. ^{13}C -NMR spectrum of rosemary leaf DCM extract recorded in DMSO- d_6 at 100 MHz

A.



B.

Name	Retention time (min)	Area (%)
α -mangostin	19.6	68
γ -mangostin	14.9	14
Gartanin	18.8	5
Garcinone E	22.2	5
8-desoxygartanin	16.3	3.5
β -mangostin	29.6	1.5
Gudraxanthone	14.2	1.0
9-hydroxycalabaxanthone	26.4	< 1%

C.

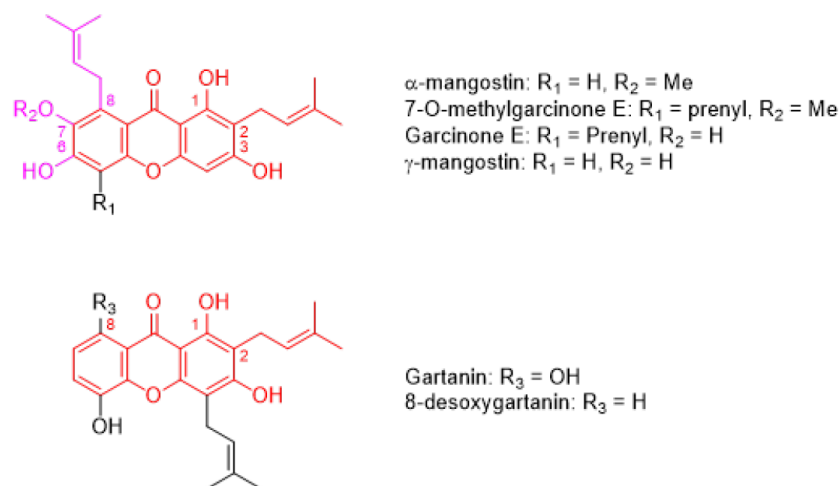


Figure S22. (A) HPLC-UV (λ 254 nm) chromatogram of the *Garcinia mangostana* (fruit peel) cyclohexanic extract (B) % (area) of each xanthone. (C) Structure and numbering of xanthones

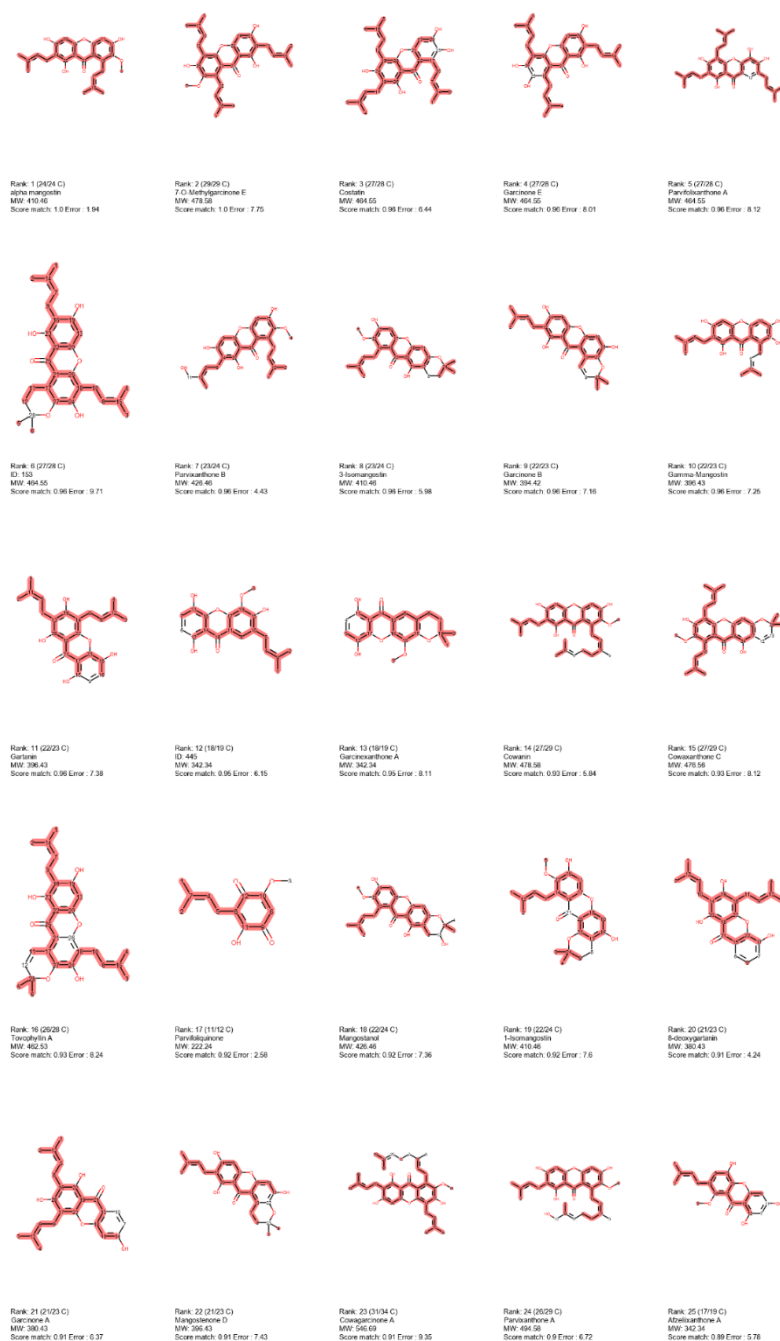


Figure S23. Part of displayed results for the ^{13}C -NMR dereplication (+ DEPT 90 and 135) of the mangosteen (fruit peel) cyclohexanic extract. Equivalent carbons were allowed.

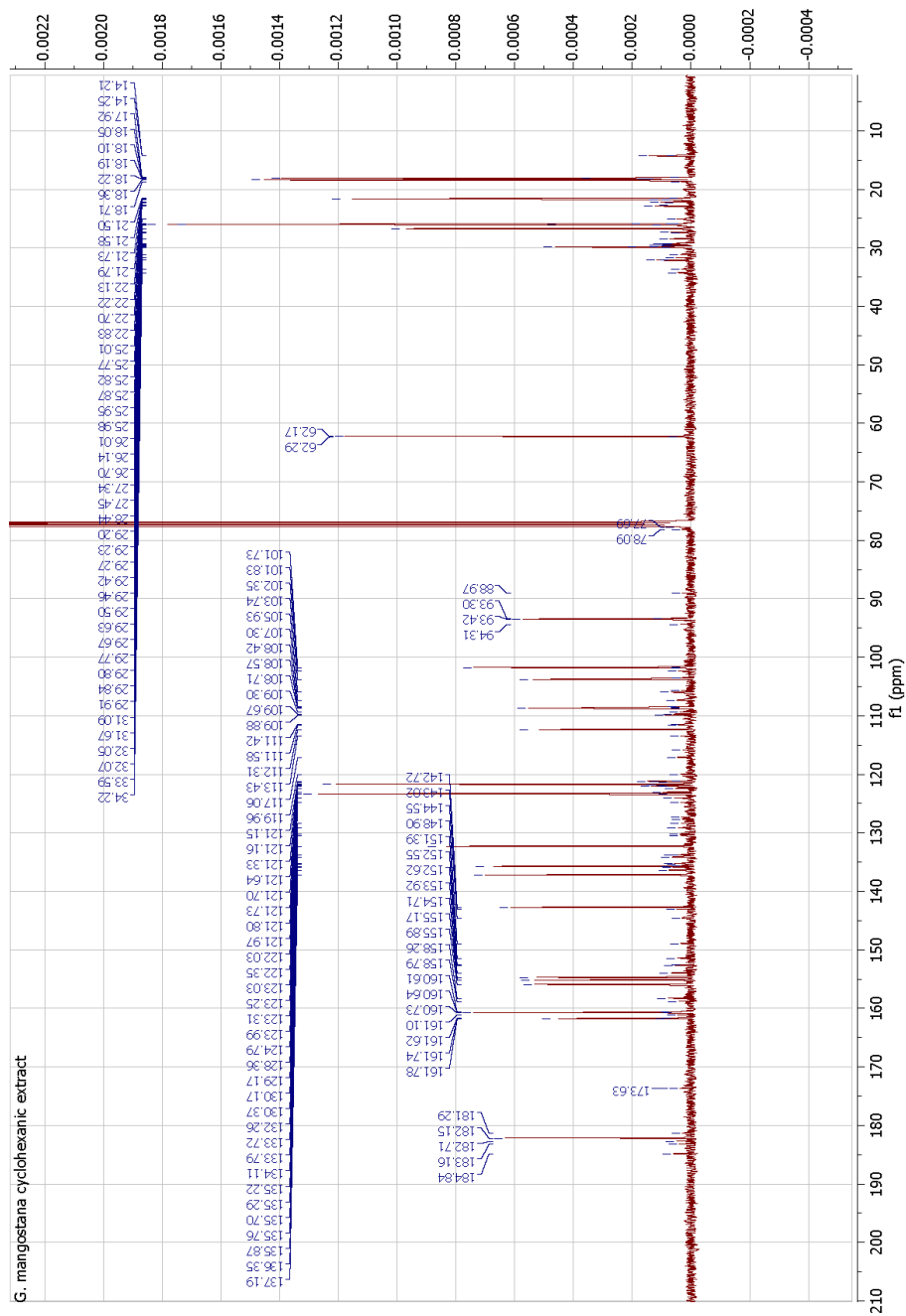


Figure S24. ¹³C-NMR spectrum of mangosteen (fruit peel) cyclohexanic extract recorded in CDCl₃ at 100 MHz

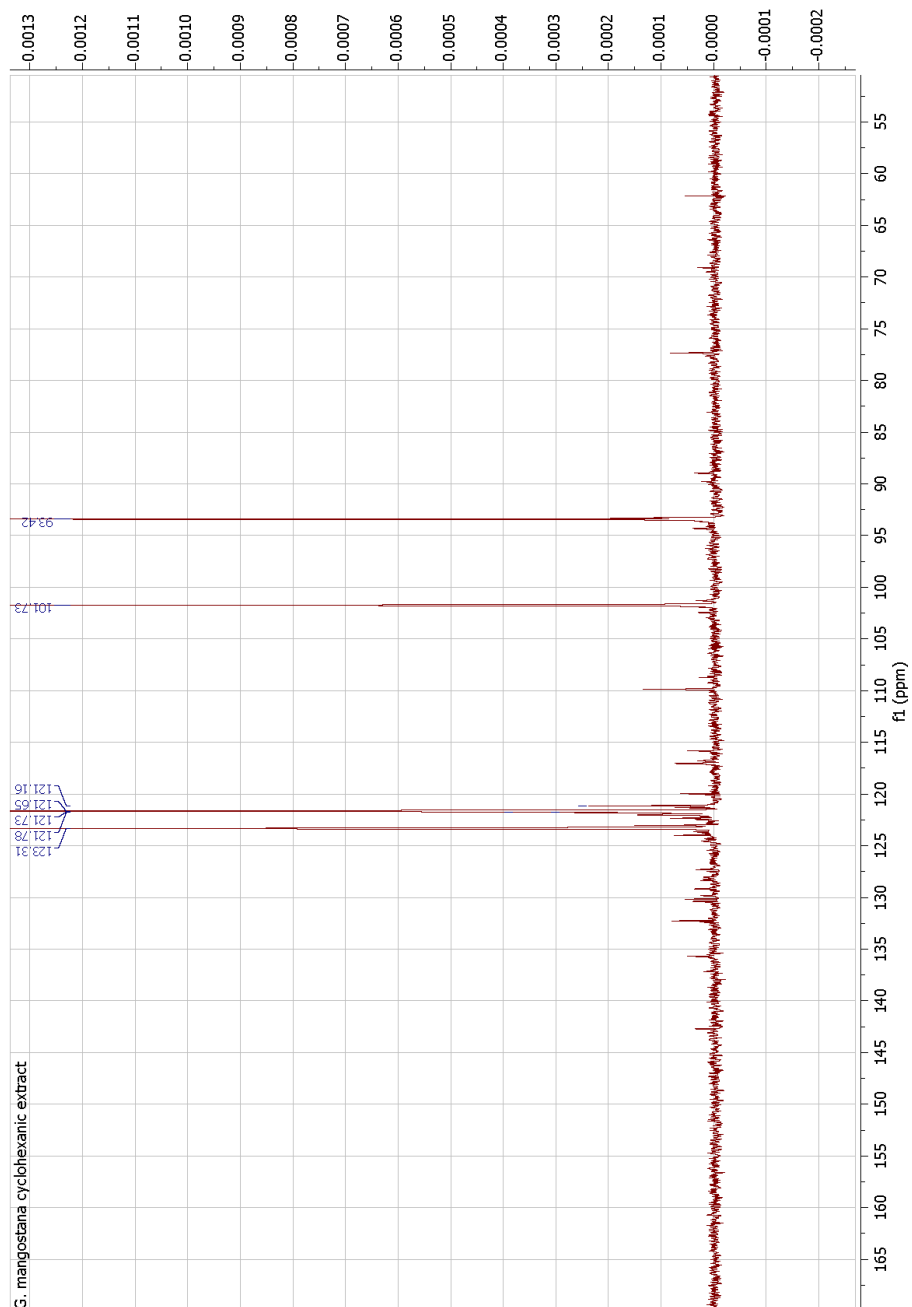


Figure S25. DEPT-90 spectrum of mangosteen fruit peel cyclohexanic extract recorded in CDCl_3 at 100 MHz.

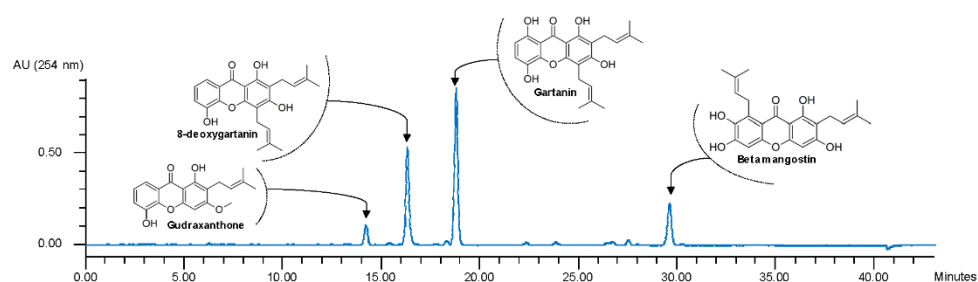


Figure S26. HPLC-UV (λ 254 nm) chromatogram of the fraction from the *Garcinia mangostana* (fruit peel) cyclohexanic extract containing minor xanones.

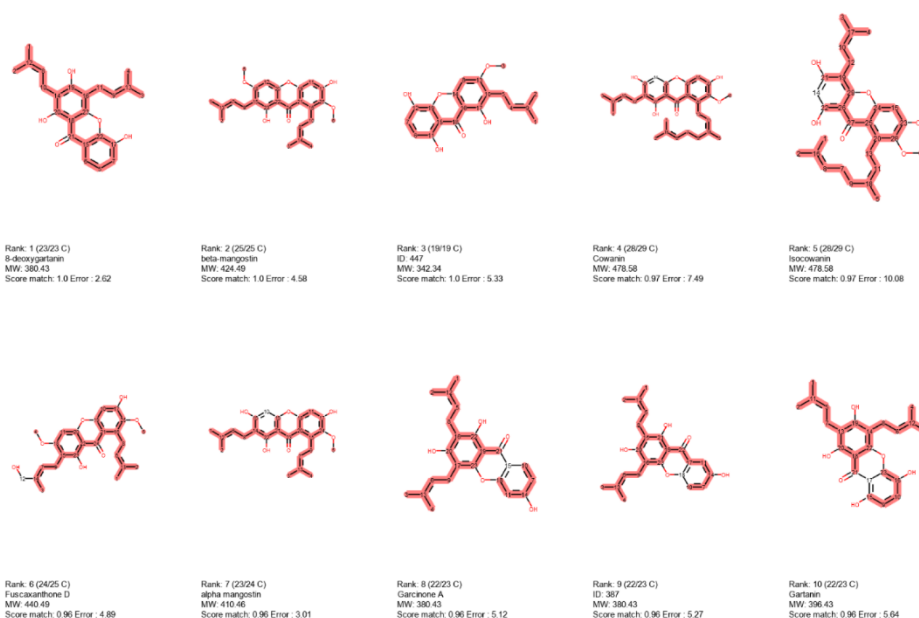
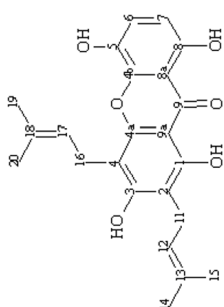
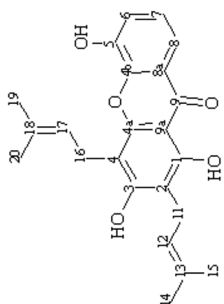


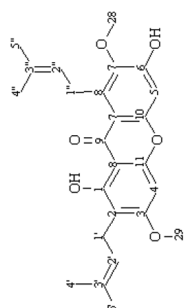
Figure S27. Part of displayed results for the ^{13}C -NMR dereplication (+ DEPT 90 and 135) of the mangosteen (fruit peel) fraction based on *Garcinia* DB. Equivalent carbons were allowed.



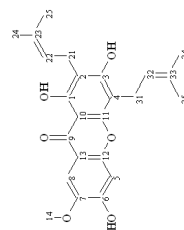
Rank: 3 (23/23 C)
Gartanin
MW: 396.44 Score: 1.0



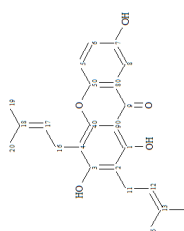
Rank: 2 (23/23 C)
8-Desoxygartanin
MW: 380.44 Score: 1.0



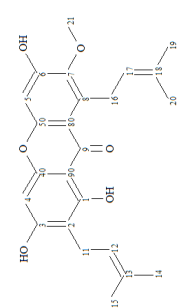
Rank: 1 (25/25 C)
Beta-mangostin
MW: 424.49 Score: 1.0



Rank: 6 (23/24 C)
cochinchinone E
MW: 410.5 Score: 0.96



Rank: 5 (22/23 C)
alpha-mangostin
MW: 380.4 Score: 0.96



Rank: 4 (23/24 C)
alpha-mangostin
MW: 410.5 Score: 0.96

Figure S28. Part of displayed results for the ^{13}C -NMR dereplication (+ DEPT 90 and 135) of the mangosteen (fruit peel) fraction based on CH-NMR-NP DB. Equivalent carbons were allowed.

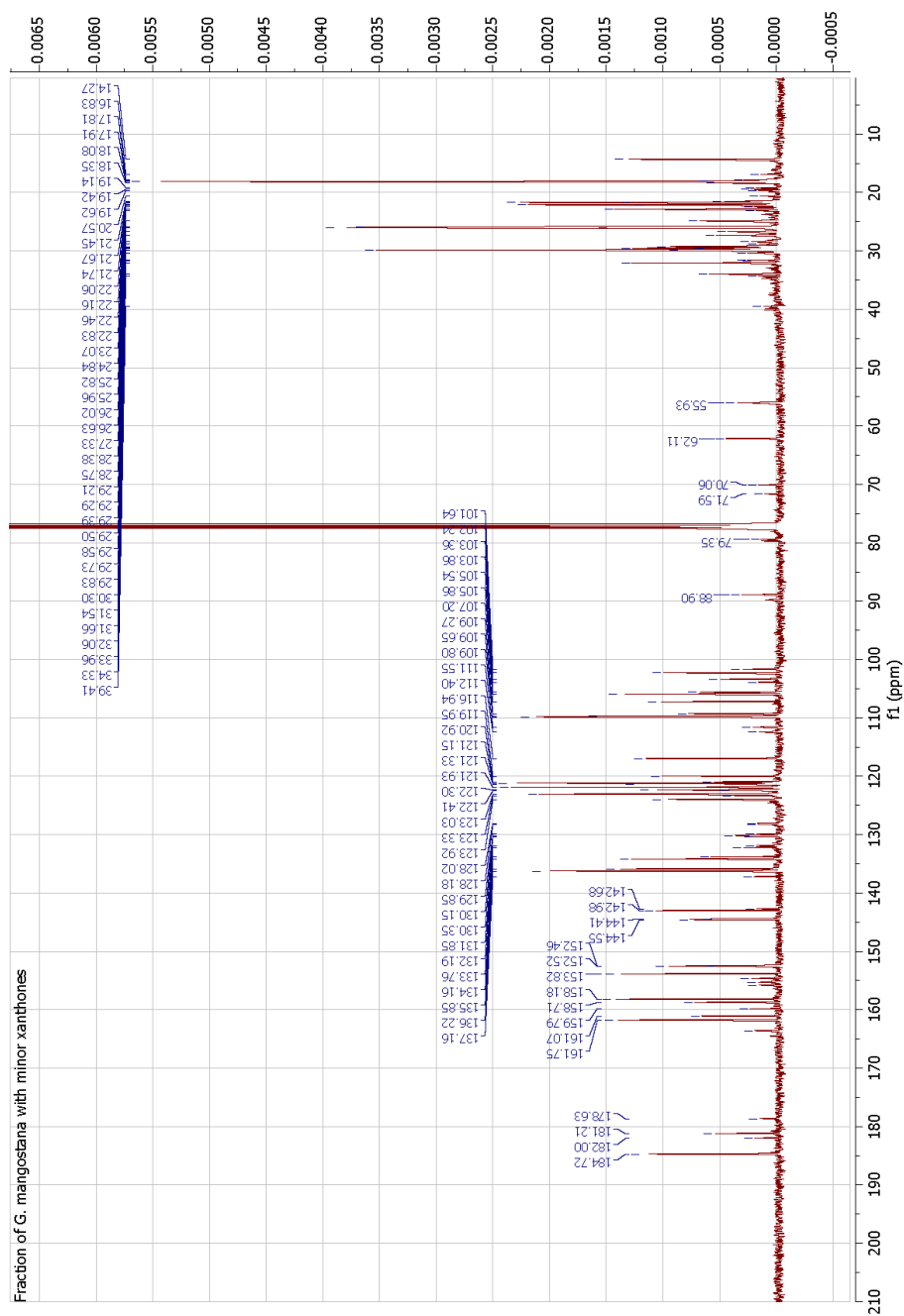


Figure S29. ^{13}C -NMR spectrum of the fraction from mangosteen (fruit peel) cyclohexanic extract containing minor xanthenes recorded in CDCl_3 at 100 MHz

Table S1. (-)-noscapine δ_C in the poppy extract and predicted values. $\Delta\delta_C > 1.3$ ppm are highlighted in yellow.

(-)-noscapine		
δ_{13C} in the extract (ppm) in CDCl ₃	Predicted δ_{SDF} (ppm)	$\varepsilon = \delta_{SDF} - \delta_{13C} $ (ppm)
168.3	166.9	1.4
152.3	152.6	0.3
148.0	148.5	0.0
147.8	147.8	0.0
141.2	145.8	4.6
140.7	140.5	0.3
134.1	134.1	0.0
132.2	129.7	2.5
120.3	120.6	0.3
119.0	119.5	0.5
118.2	118.6	0.4
117.9	117.2	0.7
102.5	102.4	0.0
100.9	100.8	0.1
82.0	82.3	0.3
62.5	62.0	0.5
61.0	63.1	2.1
59.5	59.4	0.1
56.9	56.7	0.2
50.0	49.1	0.9
46.3	44.3	2.0
22.8	23.5	0.7

Chemical shifts were only listed in descending order without any attribution.

Table S2. ^{13}C NMR spectrum of papaverine recorded in CDCl_3 at 100 MHz

Papaverine	
Carbon number (type)	δ_{isc} (ppm)
1 (Cq)	157.8
3 (CH)	141.0
4 (CH)	118.8
4a (Cq)	133.5
5 (CH)	105.3
6 (Cq)	152.4
7 (Cq)	149.8
8 (CH)	104.2
8a (Cq)	123.0
1' (Cq)	132.2
2' (CH)	111.8
3' (Cq)	149.0
4' (Cq)	147.5
5' (CH)	111.1
6' (CH)	120.5
OMe	56.1
OMe (X2)	55.9
OMe	55.8
CH_2	42.3

Table S3. Retention time (t_R) and ESI-MS² data for the major compounds in the rosemary leaf MeOH extract

Peak	t_R (min)	UV λ_{max} (nm)	(+)-ESI-MS m/z	(+)-ESI-MS ² m/z	(-)-ESI-MS m/z	(-)-ESI-MS ² m/z	Suggested molecular weight (Da)	Hypothetical structure
1	3.0	-	325 [M+H-H ₂ O] ⁺ 343 [M+H] ⁺ 360 [M+H+H ₂ O] ⁺	145 [M+H-H ₂ O- 180] ⁺ /127/163/97/ 264/292/307/247/109/2 23/186	683 [2M-H] ⁻ 341 [M-H] ⁻	341 [M+H] ⁻	342	Disaccharide (sucrose ?)
2	20.6	202, 329	743 [2M+Na] ⁺ 383 [M+Na] ⁺	383 [M+Na] ⁺	719 [2M-H] ⁻ 359 [M-H] ⁻	359 [M+H] ⁻	360	Rosmarinic acid

Table S4. Retention time (t_R) and ESI-MS² data for the major compounds in the rosemary leaf DCM and E392 extracts

Peak	t_R (min)	UV λ_{max} (nm)	(+)-ESI-MS m/z	(+)-ESI-MS ² m/z	(-)-ESI-MS m/z	(-)-ESI-MS ² m/z	Suggested molecular weight (Da)	Hypothetical structure ¹
1	12.7	282	331 [M+H] ⁺ 683 [2M+Na] ⁺	285 [M-CO ₂ H] ⁺ 267 [M+H-64] ⁺	329 [M-H] ⁻ 659 [2M-H] ⁻	285 [M-H-CO ₂] ⁻	330	Carnosol
2	19.1	284	333 [M+H] ⁺	287 [M-CO ₂ H] ⁺	331 [M-H] ⁻ 685 [2M-2H+Na] ⁻	287 [M-H-CO ₂] ⁻	332	Carnosic acid
3	31.9	-	455 [M+H] ⁺ 437 [M-H ₂ O+H] ⁺	409 [M-CO ₂ H] ⁺ 437 [M-H ₂ O+H] ⁺	453 [M-H] ⁻	405 [M-H-48] ⁻	454	Micromeric acid
4	36.7	-	457 [M+H] ⁺ 439 [M-H ₂ O+H] ⁺	393 [M+H-64] ⁺	455 [M-H] ⁻	455 [M-H] ⁻	456	Betulonic acid
5	40.2	-	457 [M+H] ⁺ 439 [M-H ₂ O+H] ⁺	411 [M-CO ₂ H] ⁺ 439 [M-H ₂ O+H] ⁺	455 [M-H] ⁻	455 [M-H] ⁻ 407 [M-H-48] ⁻	456	Ursolic acid / Oleanolic acid

¹ Kontogianni, V. G.; Tornie, G.; Nikolie, I.; Nerantzaki, A. A.; Sayyad, N.; Stosic-Grujicic, S.; Stojanovic, I.; Gerothanassis, I. P.; Tzakos, A. G. *Food Chem.* **2013**, 136, 120-129.

Table S5. δ_c matched to triterpenes and carnosic acid in the E392 extract along with their respective predicted values. ¹ $\Delta\delta_c > 1.3$ ppm are highlighted in yellow. Hidden and missing chemical shifts are highlighted in green and blue respectively.

Carbons numbering	Betulinic acid		Ursolic acid		Micromeric acid		Oleanolic acid		Carnosic acid	
	δ_{13C} matched in the extract (ppm) in DMSO-d6	Predicted δ_{SDF} (ppm)	δ_{13C} matched in the extract (ppm) in DMSO-d6 ²	Predicted δ_{SDF} (ppm)	δ_{13C} matched in the extract (ppm) in DMSO-d6 ³	Predicted δ_{SDF} (ppm)	δ_{13C} matched in the extract (ppm) in DMSO-d6	Predicted δ_{SDF} (ppm)	δ_{13C} matched in the extract (ppm) in DMSO-d6	Predicted δ_{SDF} (ppm)
1	38.3	38.2	38.3	39.0	38.3	38.5	38.3	39.1	33.9	33.9
2	27.0	26.5	27.6	28.0	27.6	27.4	27.6	27.9	19.8	19.3
3	76.9	78.4	76.9	78.2	76.9	78.4	76.9	78.3	40.9	41.6
4	38.7	38.1	38.4	39.2	38.4	38.9	38.4	39.3	33.6	34.5
5	53.5	54.0	54.8	55.6	54.8	55.2	54.8	56.0	53.5	53.9
6	19.8	19.7	19.8	19.4	18.3	18.5	19.8	19.5	20.5	20.4
7	33.9	34.1	33.3	33.3	33.4	33.0	33.9	33.6	32.4	32.1
8	41.3	41.1	-	39.8	41.3	40.2	-	39.8	129.0	129.2
9	50.0	50.3	48.5	48.0	48.5	47.9	48.5	48.2	-	121.7
10	36.6	37.0	36.6	37.2	36.6	37.4	36.6	37.5	46.8	47.3
11	20.5	21.0	23.8	23.5	23.8	23.8	23.8	23.8	-	142.5
12	25.1	25.9	125.0	125.4	124.6	124.6	121.6	121.7	-	141.3
13	38.5	38.6	139.1	139.0	139.0	139.0	145.2	144.7	134.3	133.9
14	42.4	42.8	42.4	42.4	42.4	42.3	42.4	42.6	121.6	119.1
15	31.2	30.7	29.2	28.7	29.3	28.2	29.3	28.3	26.0	27.2
16	32.4	32.7	25.1	24.8	25.1	24.2	27.0	27.0	22.7	22.4
17	55.4	56.5	47.1	48.1	-	49.0	46.8	47.3	22.7	22.4
18	48.5	48.7	53.5	53.4	52.4	52.8	40.8	41.7	32.9	2.8
19	47.2	48.6	38.6	39.4	40.8	39.8	-	45.3	23.0	22.1
20	150.3	151.0	38.5	39.2	153.0	151.9	31.4	31.0	177.7	180.6

² Chemical shifts described in Brandão, G., Kroon, E., Souza, D., Filho, J., Oliveira, A., 2013. Chemistry and Antiviral Activity of *Arrabidaea pulchra* (Bignoniaceae). Molecules 18, 9919–9932. <https://doi.org/10.3390/molecules18089919>.

³ Micromeric acid was purified and NMR experiments were registered in DMSO-d6.

21	30.2	30.5	31.2	31.1	31.2	30.7	-	34.7		
22	36.4	37.5	36.4	37.3	36.4	37.2	33.9	33.6		
23	27.4	27.3	27.4	27.4	27.6	27.4	27.4	27.4		
24	28.9	29.3	28.9	29.4	28.9	29.4	28.3	29.4		
25	16.0	16.0	15.1	15.0	15.1	14.9	15.1	14.9		
26	16.1	16.2	17.1	17.2	16.9	17.0	17.1	17.3		
27	15.1	14.9	23.3	23.9	23.3	24.0	25.6	26.1		
28	178.6	179.5	178.4	180.3	177.3	181.0	178.6	179.7		
29	109.7	109.8	16.9	17.4	104.6	106.2	-	30.2		
30	19.0	19.5	21.1	21.4	17.1	17.1	28.9	29.3		

Table S6. ^{13}C NMR spectrum of carnosol recorded in DMSO- d_6 at 100 MHz

Carnosol	
Carbon number (type)	$\delta_{13\text{C}}$ (ppm)
1 (CH_2)	29.5
2 (CH_2)	18.8
3 (CH_2)	40.8
4 (Cq)	34.4
5 (CH)	45.2
6 (CH_2)	29.1
7 (CH)	77.3
8 (Cq)	131.9
9 (Cq)	122.1
10 (Cq)	48.1
11 (Cq)	143.6
12 (Cq)	143.3
13 (Cq)	134.6
14 (CH)	111.6
15 (CH)	26.5
16 (CH_3)	22.9
17 (CH_3)	19.7
18 (CH_3)	23.1
19 (CH_3)	31.6
20 (Cq)	176.1

Table S7. δ_c matched to xanthenes in the mangosteen cyclohexanic extract along with their respective predicted values. $\Delta\delta_c > 1.3$ ppm are highlighted in yellow. Hidden and missing chemical shifts are highlighted in green and blue respectively.

Carbons numbering	α -mangostin ⁴			Garcinone E ⁵			γ -mangostin ⁶			Gartianin ⁷			8-deoxygartianin ⁷		
	δ_{13C} matched in the extract (ppm) in CDCl ₃	Predicted δ_{SDF} (ppm)	δ_{13C} matched in the extract (ppm) in CDCl ₃	δ_{SDF} (ppm)	Predicted δ_{SDF} (ppm)	δ_{13C} matched in the extract (ppm) in CDCl ₃	δ_{SDF} (ppm)	Predicted δ_{SDF} (ppm)	δ_{13C} matched in the extract (ppm) in CDCl ₃	δ_{SDF} (ppm)	Predicted δ_{SDF} (ppm)	δ_{13C} matched in the extract (ppm) in CDCl ₃	δ_{SDF} (ppm)	Predicted δ_{SDF} (ppm)	δ_{13C} matched in the extract (ppm) in CDCl ₃
1	160.7	160.7	161.1	160.9	162.8	161.7	162.8	158.8	158.8	158.5	158.5	158.8	158.8	158.7	158.7
2	108.4	108.4	109.3	109.3	108.6	108.6	108.6	111.4	111.4	110.6	110.6	109.3	109.3	109.1	109.1
3	161.6	161.6	161.8	162.4	161.8	161.8	161.8	161.8	161.8	162.0	162.0	161.1	161.1	160.9	160.9
4	93.4	93.3	93.4	93.3	93.0	93.4	93.0	107.3	107.3	106.8	106.8	109.3	109.3	109.1	109.1
4a	155.2	155.1	155.9	156.8	155.8	155.9	155.8	152.6	152.6	153.1	153.1	152.6	152.6	152.4	152.4
5	101.7	101.6	113.4	114.0	101.2	101.7	101.2	137.2	137.2	136.9	136.9	144.6	144.6	144.5	144.5
6	154.7	154.5	148.9	148.7	149.6	148.9	149.6	123.3	123.3	123.9	123.9	-	-	119.7	119.7
7	142.7	142.6	-	139.8	141.8	142.7	141.8	-	-	109.9	109.9	123.3	123.3	123.8	123.8
8	137.2	137.1	128.4	128.0	131.4	132.3	131.4	153.9	153.9	154.1	154.1	-	-	116.9	116.9
8a	112.3	112.3	111.4	110.7	110.8	111.4	110.8	108.4	108.4	107.8	107.8	121.0	121.0	120.9	120.9
9	182.2	182.1	183.2	183.2	183.3	183.2	183.3	184.8	184.8	185.4	185.4	181.3	181.3	181.1	181.1
9a	103.7	103.7	103.5	103.0	103.8	103.7	103.8	102.4	102.4	102.5	102.5	103.5	103.5	103.3	103.3
10a	155.2	155.8	152.6	152.4	153.6	153.9	153.6	144.6	144.6	143.9	143.9	143.0	143.0	144.3	144.3
1'	21.6	21.5	21.7	21.7	22.1	22.1	22.1	22.1	22.1	22.1	22.1	21.6	21.6	21.6	21.6
2'	121.6	121.4	121.7	122.2	123.6	123.3	123.6	121.8	121.8	121.9	121.9	121.2	121.2	121.2	121.2
3'	135.9	135.8	134.1	134.2	131.4	132.3	131.4	134.1	134.1	134.6	134.6	136.4	136.4	136.3	136.3
4'	18.1	18.0	18.1	18.0	18.1	18.1	18.1	18.0	18.0	18.0	18.0	18.1	18.1	18.0	18.0
5'	25.8	25.8	26.0	25.9	26.0	26.0	26.0	26.0	26.0	25.9	25.9	26.0	26.0	25.9	25.9
1''	26.7	26.6	26.0	26.1	26.5	26.7	26.5	21.8	21.8	22.0	22.0	22.1	22.1	22.1	22.1

⁴ Experimental δ_c in CDCl₃ can be found in Han, A.-R.; Kim, J.-A.; Lantvit, D. D.; Kardono, L. B. S.; Riswan, S.; Chai, H.; Caracache de Blanco, E. J.; Farnsworth, N. R.; Swanson, S. M.; Kinghorn, A. D., Cytotoxic xanthone constituents of the stem bark of *Garcinia mangostana* (Mangosteen). *J. Nat. Prod.* **2009**, *72* (11), 2028-2031.
⁵ Experimental δ_c in CDCl₃ can be found in Sakai, S.-I.; Katsura, M.; Takayama, H.; Aimi, N.; Chokethaworn, N.; Suttajit, M., The structure of garcinone E. *Chem. Pharm. Bull.* **1993**, *41* (5), 958-960.
⁶ To our knowledge, no experimental δ_c were described in CDCl₃.
⁷ Experimental δ_c in CDCl₃ can be found in Groweiss, A.; Cardellina, J. H.; Boyd, M. R., HIV-Inhibitory prenylated xanthenes and flavones from *MacLura tinctoria*. *J. Nat. Prod.* **2000**, *63* (11), 1537-9.

2''	123.3	123.2	121.7	122.4	-	124.6	123.3	123.9	121.8	122.2
3''	132.3	132.1	132.3	132.3	129.2	129.1	133.7	133.4	133.8	133.5
4''	18.2	18.2	18.1	18.1	18.1	18.1	18.2	18.1	18.1	18.0
5''	25.8	25.8	26.0	25.9	25.8	25.9	25.8	25.8	25.8	25.7
1'''			22.7	22.7						
2'''			121.2	120.8						
3'''			133.8	132.9						
4'''			18.2	18.0						
5'''			25.8	25.8						
OMe	62.2	62.1								

Detailed MixONat software.

Inputs and parameters. A graphical user interface (GUI) was designed with Kivy, an open Python library source compatible with Linux, Windows and OS X.

The home tab of the MixONat software (Figure S2) allows the selection of the input files by the user, using a file browsing function. The input files must include at least a DB (.sdf), sorted by the C-typeGen program, and ^{13}C -NMR data, imported as a table (.csv) of δ_{C} values and intensities. The users are also encouraged to provide DEPT-135 and 90 data (.csv).

The second tab (Figure S3) displays all the different parameters that can be adjusted:

- The tolerance (ϵ) reflects the accuracy of the used database, as the program compares each chemical shift in the experimental spectrum ($\delta_{13\text{C}}$) with each chemical shift in the SDF (δ_{SDF}) for each molecule. It considers that $\delta_{13\text{C}}$ matches with δ_{SDF} if $\delta_{\text{SDF}} - \epsilon < \delta_{13\text{C}} < \delta_{\text{SDF}} + \epsilon$. The default value for ϵ is 1.3 ppm.^{8,9}
- The tolerance incrementation can be turned ON or OFF. If ON, the program will start the matching process with $\epsilon = 0.0$ and will then increment this value by steps of 0.1 ppm, until it reaches the ϵ value chosen by the user. This will match first the chemical shifts which are closest together. If this parameter is OFF, the algorithm will match a δ_{SDF} with the first $\delta_{13\text{C}}$ falling into the $\pm \epsilon$ interval. It is recommended to leave this parameter ON, especially when using experimental DBs.
- The DEPT alignment parameter is necessary to associate a $\delta_{13\text{C}}$ to its corresponding carbon in the DEPT-135 and 90 spectra. Since the chemical shifts are never exactly the same from a spectrum to another, the user has to indicate the accuracy of the alignment in each one of his spectra, the default value is 0.02 ppm. The quality of the results will be very dependent to the quality of this alignment.
- The equivalent carbon factor can be turned ON or OFF. Turning it ON will allow the same $\delta_{13\text{C}}$ to be matched multiple times if several identical δ_{SDF} are found (equivalent carbons in the database). If OFF, a $\delta_{13\text{C}}$ can only be matched once, even if equivalent carbons are found in the database.
- The molecular weight filter will only show results if they correspond to the ones required by the user. One can either select specific mass, or a range of values. The algorithm will search for molecules with the indicated molecular weight(s).
- The number of results to display can be selected by the user, as well as the number of results per page when saving the files, making it easier to browse and analyze them.
- For each compound of the DB, a score is attributed and corresponds to the number of $\delta_{13\text{C}}$ matched with δ_{SDF} out of the number of carbons of the compound. A minimal score to reach can also be a parameter to limit the number of results. If a molecule does not reach the minimal score, it will not be displayed; the minimum being 0.0 and the maximum 1.0. By default, the minimal score is 0.0, in order to be able to see every result that fits the previous parameters.
- The directory of the output files can be selected by the user, using a browser function.

The third and last tab is the C-typeGen program that allows the user to create DBs compatible with the MixONat dereplication process. It sorts δ_{SDF} by carbon types. Via a browser, the user must select the input files required, meaning the original database and the same database without stereochemistry. It is important that the stereochemistry is removed before the sorting, because it can alter the way the program works, resulting in a wrong carbon type listing. Even if this is a mandatory step before proceeding to the dereplication, this program is only used when creating a new database and is otherwise not required for the search itself.

Matching process. First, the program will start by sorting each carbon of the ^{13}C NMR spectrum depending on their type. This sorting is different, depending on the DEPT files provided by the user and according to the chosen DEPT alignment. If there is no DEPT data, then the carbon types are not differentiated. If a DEPT 135 has been given, carbons that have a negative intensity will be considered as methylenes, carbons with a positive intensity will be considered as methines or methyles. The rest of the chemical shifts from the ^{13}C -NMR spectrum that are not visible in the DEPT-135 will be labeled as quaternaries. If DEPT-90 information is also provided, then the program will first consider that signals in the DEPT-90 are methines, before considering negative signals on the DEPT 135 as secondaries, positive signals on the DEPT 135 as primaries, and remaining carbons from the ^{13}C spectrum as quaternaries.

In a second time, the matching process consists, for each compound of the DB, in the comparison of $\delta_{13\text{C}}$ with δ_{SDF} . It is done by list of carbon types, meaning that the carbons labeled as quaternaries from the user spectra can only be matched

⁸ The tolerance can be lowered when working with experimental DBs or increased if the DB was created with a less accurate prediction software.

⁹ Bruguière, A.; Derbré, S.; Coste, C.; Le Bot, M.; Siegler, B.; Leong, S. T.; Sulaiman, S. N.; Awang, K.; Richomme, P. *Fitoterapia* **2018**, *131*, 59-64.

with quaternary carbons from each compound of the SDF. All δ are first sorted by descending numerical order before the matching process. The program starts by the first molecule of the SDF, with the first δ_{SDF} , and tries to find a $\delta_{13\text{C}}$ that fits in the tolerance interval. As a reminder, if incrementation is ON, the program will first look for perfect matches with $\varepsilon = 0.0$, and gradually go up each cycle until ε is the tolerance value chosen by the user. During the latter, the algorithm allows or not multiple uses of $\delta_{13\text{C}}$, depending on the equivalent carbon parameters. When all the δ_{SDF} have been considered, the score and error of the molecule are calculated and stored. The score is defined as the number of $\delta_{13\text{C}}$ matched with δ_{SDF} out of the number of carbons of the compound. a score of 1.0 thus means that all the carbons from the molecule have been matched; 0.50 means that only half of the signals were found. The error is the cumulated absolute difference between matched signals (*i.e.* $\sum |\delta_{\text{SDF}} - \delta_{13\text{C}}|$). The algorithm will repeat the matching process for each molecule in the SDF which has a molecular weight fitting the parameters chosen by the user

In some rare cases, the algorithm matches couples of δ_{SDF} and $\delta_{13\text{C}}$ values (within the same ε window) with a non-optimal solution, increasing the error value. For example, the matching process associated $\delta_{\text{SDF}1}$ with $\delta_{13\text{C}1}$ and $\delta_{\text{SDF}2}$ with $\delta_{13\text{C}2}$ but, under closer inspection, if $\delta_{\text{SDF}1}$ was matched with $\delta_{13\text{C}2}$ and $\delta_{\text{SDF}2}$ with $\delta_{13\text{C}1}$, the total error on the molecule would be lower, and thus the molecule can see its rank increased. Thus, the very end of the process, another algorithm will optimize the local error on matched values.

Interactive results. At the end of the matching process, the number of desired molecules is displayed if they reach the minimal score fixed by the user. Compounds are ranked by decreasing score and increasing error, their structure, name, molecular weight, score and error (Figure S4). On the structure, matched carbons are highlighted in red. For each molecule, it is possible to open a window that shows its numbered structure, and which δ_{SDF} have been matched. Along with it, matched chemical shifts and their corresponding intensity is shown as a graphical representation of a ^{13}C -NMR spectrum. On the reconstituted spectrum, each carbon type will give a different color and be numbered according to the structure and chemical shift list displayed. This easily shows if the intensities of matched carbons are homogenous, hence hypothetically being signals belonging to the same molecule (Figure S5). It is thus possible for the user to link the information gathered from the structure, the spectrum and the chemical shifts and use his knowledge of NMR spectroscopy to remove or add chemical shifts, artificially modifying the score of the selected molecule. This function can be, for example, used to remove a carbon matched with an abnormal intensity (meaning probably belonging to another molecule), or to add a quaternary carbon that was not matched because it was not picked on the spectrum due to its low intensity, or predicted a bit too far away, etc. If a carbon is added or removed, the spectrum and the highlighted carbons on the structure will be updated accordingly. It is also possible to delete a molecule altogether if necessary. Once the results have been improved by the user, they can be saved as a text and image file (Figure S6). The text file will summarize the parameters used for the research, show the information for each molecule (*i.e.* name, rank, matched carbons by carbon type). The image file contains the structure of the compounds with the matched carbons highlighted, as well as their name, molecular weight, score and error.

3.2.3. Éléments de discussion

Ce travail a montré que les hypothèses données par le programme MixONat permettent de toujours s'orienter vers un type structural et, la plupart du temps, vers les molécules correctes elles-mêmes, puisqu'elles apparaissent en règle générale dans les 25 premières positions. L'ajout des informations obtenues avec les expériences DEPT améliore également le score des molécules réellement présentes. Cela permet de considérablement réduire le nombre de molécules proposées au chercheur, puisqu'il a seulement à vérifier les données expérimentales pour une vingtaine de molécules au maximum afin de comparer les données spectrales.

Les problèmes qui peuvent ressortir de ces essais sont avant tout des problèmes intrinsèques aux analyses par RMN. La **détection des composés minoritaires** est limitée (concentrations trop faibles) et impose donc un fractionnement des extraits, comme dans l'exemple de l'extrait brut de mangoustan. Les carbones quaternaires (relaxation lente, absence d'effet nOe et de transfert de polarisation ^{13}C) ont une intensité nettement plus faible que les autres carbones, ce qui entraîne parfois une confusion de leurs signaux avec le bruit de fond, rendant impossible pour l'algorithme de les associer aux déplacements chimiques de la base de données, et impactant ainsi négativement le score des molécules associées. Le second type de problème est celui lié à la **prédiction des déplacements chimiques** qui peut, dans certains cas, être trop éloignée des données expérimentales. Le dernier type de problème vient de la façon dont le programme fonctionne puisqu'il repose sur un **algorithme qui reste glouton**, ne cherchant pas une solution optimale de façon générale, mais seulement à satisfaire des conditions de façon locale. Il arrive donc que les associations choisies entraînent l'isolement (c'est-à-dire la non-association) d'un signal qui aurait pu être associé, comme c'est le cas pour certains signaux de l'acide rosmarinique dans l'exemple de l'extrait méthanolique de romarin.

Lorsque l'on travaille avec des bases de données contenant des structures très proches (par exemple, les xanthones prénylées de la base *Garcinia*), il suffit qu'un des problèmes précédemment mentionnés empêche l'association d'un seul déplacement pour diminuer le score et faire chuter la molécule dans le bas du classement. Un exemple illustrant ce phénomène peut être celui de la fraction de *Garcinia mangostana* dans laquelle 9 xanthones prénylées possèdent un score parfait (100%) tandis que la gartanine, bien présente dans la fraction, « perdant » un carbone à cause d'un problème de prédiction, obtient un score de 96% et est reléguée en 10^{ème} position. L'avantage du programme MixONat est qu'il est cependant possible de remédier à ces différents problèmes en passant par l'analyse interactive des résultats et leur modification.

3.2.4. Perspectives d'amélioration de MixONat

Le programme pourra faire l'objet de diverses améliorations futures sur plusieurs points.

D'abord sur la lecture et le tri selon n'importe laquelle des informations présentes dans la banque de données. En effet, comme décrit précédemment, les SDF peuvent contenir autant d'information que souhaitée. On peut imaginer pour chaque molécule, des informations sur les espèces, les genres et les familles de plantes desquels elles sont issues. Des banques de données beaucoup plus larges pourrait ainsi être filtrée plus facilement selon des critères chimiotaxonomiques. Il pourrait également être possible d'indiquer certaines molécules comme étant des « biomarqueurs » de certains groupes de plantes (morphine, artémisinine, ...) ou, au contraire, des

molécules « ubiquitaires » qui sont présentes dans beaucoup de plantes différentes (β -sitostérol, kaempférol, acide chlorogénique...). D'autres informations pourraient également être stockées, notamment lors de l'utilisation de base de données expérimentales, dans lesquelles le solvant deutéré utilisé pour les analyses serait indiqué, permettant des recherches plus précises « par solvant ».

Une autre optimisation consisterait en l'ajout de la possibilité de valider une molécule une fois que sa présence a été validée dans l'extrait. Le programme se chargerait de retirer les signaux ayant déjà utilisés pour cette molécule et commencerait une nouvelle recherche avec les signaux restants. Ainsi de suite, jusqu'à ce que tous les signaux présents dans le mélange soient identifiés. Dans ce cas, le fait que les molécules soient correctement associées permettrait également de pouvoir sauvegarder les déplacements chimiques réels des molécules dans la base sous une nouvelle étiquette. Cela permettrait de progressivement enrichir la base à l'aide de données expérimentales, y compris avec des valeurs de δ_c non disponibles dans la littérature, car il n'est pas rare qu'un composé soluble en mélange (typiquement dans $CDCl_3$) ne le soit plus une fois isolé (cas de la gudraxanthone chez *G. mangostana*)

Une « chromatographie virtuelle » des mélanges pourrait également être envisagée. En effet, si l'utilisateur possède plusieurs fractions du même extrait, le logiciel se chargerait de regarder quels sets de δ_c sont toujours retrouvés ensemble parmi ces fractions. En théorie, un set de δ_c correspondra à seule molécule, ou au moins un mélange de molécules plus simple à traiter que la fraction dans son intégralité. Le programme pourra donc séparer les signaux par groupes simplifiés et effectuer sa recherche ainsi.

4. Application de la méthode

Notre méthode de déréplication par RMN du carbone 13 étant maintenant opérationnelle et validée par plusieurs exemples, elle peut être appliquée en routine au laboratoire pour des analyses déréplicatives « en aveugle ». Dans cette partie, l'application qui en sera faite portera sur un des objectifs initiaux de ce travail, à savoir l'utilisation de déréplication afin de rapidement isoler les molécules qui nous intéressent : les acylphloroglucinols polycycliques polyprénylés (ou PPAPs).

4.1. Article 4 : Polyprénylated polycyclic acylphloroglucinols identification from *Garcinia bancana* bark using ^{13}C -NMR dereplication program MixONat

4.1.1. Résumé de l'article 4

L'article suivant reprend les point-clés de ce travail.

Le but était d'isoler des PPAPs structuralement diverses, pour ensuite pouvoir tester si elles possédaient une activité anti-inflammatoire et/ou immunomodulatrice comme la guttiferone J en avait exhibé [3]. Les Clusiaceae, Calophyllaceae et Hypericaceae sont des familles de plantes connues pour contenir des PPAPs. Elles ont donc constitué notre source potentielle majoritaire pour l'isolation de PPAPs.

Une collaboration avec la Malaisie nous a permis d'obtenir 124 extraits dichlorméthaniques et méthanoliques d'écorce, de feuille et parfois de fruit de 17 espèces *Garcinia* (30 lots en tout). Les profils LC-UV-MS² des différents extraits ont permis de sélectionner les extraits contenant un grand nombre de PPAPs. Parmi eux, se trouvait un extrait dichlorométhanique d'écorce de *Garcinia bancana*. C'est sur cette espèce que l'analyse déréplivative a été menée.

L'extrait brut, ainsi que 4 fractions issues d'une étape de chromatographie flash (F8, F9, F10 et F12) ont été analysés par RMN- ^{13}C , DEPT 135 et DEPT 90. Chaque mélange a été analysé par MixONat en utilisant d'une part une banque de données chimiotaxonomique, regroupant les molécules regroupées dans le genre *Garcinia*, et d'autre par une banque de données structure-spécifique, contenant toutes les PPAPs reportées dans la littérature [33].

L'analyse de l'extrait brut a d'emblée permis de suggérer que les PPAPs constituaient les composés majoritaires, puisqu'en utilisant la banque de données *Garcinia*, seules des structures de PPAPs étaient proposées. Avec la base spécifique PPAPs, des stéréoisomères de l'isoxanthochymol étaient fortement suggérés. En vérifiant les données de la littérature pour les 22 premières hypothèses, nous avons pu vérifier que la guttiferone F, le xanthochymol, la 30-epi-cambogine, le (-)-cycloxanthochymol, la garcicowin C et le 7-epi-isogarcinol étaient bien présents dans l'extrait.

Un travail similaire sur les fractions a permis d'identifier, en plus de la guttiferone F et du xanthochymol, la garcinialiptone A dans la F8 grâce à la suggestion par le programme de 2 nouvelles PPAPs, garcinialiptone A et garciniagifolone A, présentant une cyclisation supplémentaire par rapport aux PPAPs précédemment identifiées. La guttiferone F, le xanthochymol, les deux rotamères de la garcinialiptone A, la 30-epi-cambogine et le (-)-cycloxanthochymol ont été identifiés dans la fraction 9. La présence de la 30-epi-cambogine, du (-)-cycloxanthochymol et de la garcicowin C ont été validés dans F10. Enfin, l'analyse de F12 a permis de confirmer la présence du 7-epi-isogarcinol (**Tableau 10**).

Tableau 10: Récapitulatif des molécules isolées dans les différentes fractions. Les molécules dont les noms sont soulignés ont pu être directement identifiées depuis l'extrait brut.

Molécules	F8	F9	F10	F12
<u>Guttiferone F</u>	☑	☑		
<u>Xanthochymol</u>	☑	☑		
Garcinialiptone A (rotamer 1)	☑	☑		
Garcinialiptone A (rotamer 2)		☑		
<u>30-epi-cambogin</u>		☑	☑	
<u>(-)-Cycloxanthochymol</u>		☑	☑	
<u>Garcicowin C</u>			☑	
<u>7-epi-isogarcinol</u>				☑

4.1.2. Article 4

1 **Polyprenylated polycyclic acylphloroglucinols identification from *Garcinia***
2 ***bancana* bark using the ^{13}C -NMR dereplication program MixONat**

3
4 Antoine Bruguère^a, Séverine Derbré^{a,*}, Sow Tein Leong^b, Noor Aimi Othman^b, Robert B.
5 Grossman^c, Khalijah Awang^b, Pascal Richomme^a

6 ^a SONAS SFR QUASAV, University of Angers, France.

7 ^b Department of Chemistry, Faculty of sciences, University of Malaya, Malaysia

8 ^c Department of Chemistry, University of Kentucky, Lexington, United States

9
10 **Abstract:** Polycyclic polyprenylated acylphloroglucinols (PPAPs) are a class of natural
11 products (NPs) that have previously shown modulating activity of the human immune system.
12 The goal of our work was thus to be able to quickly focus on extracts containing PPAPs with
13 either original structures or a potentially high activity. Dereplication methods that allow the
14 rapid identification of known compounds inside mixtures seemed to be the appropriate tool to
15 reach our objective. The different PPAPs often being stereoisomers, we were limited by their
16 detection with dereplication methods relying on mass spectrometry. We thus decided to turn to
17 ^{13}C -NMR dereplication instead, using a custom-made program called MixONat. We were able
18 to quickly identify 8 PPAPs, namely guttiferone F (**1**), xanthochymol (**2**), 2 rotamers of
19 garcinialiptone A (**3**), 30-epi-cambogin (**4**), (-)-cycloxanthochymol (**5**), garcicowin C (**6**) and
20 7-epi-isogarcinol as major NPs from a *Garcinia bancana* bark extract. 6 of them were directly
21 identified from the crude extract. Additional fractionation steps allowed the identification of
22 the remaining 2 minor compounds and the isolation of the molecules for further biological
23 testing. This work confirmed that MixONat is a powerful tool that can be used to speed-up
24 extract characterization in NPs research.

25

26 **Keywords:** Dereplication, ^{13}C -NMR, *Garcinia*, PPAP

27

28 1. INTRODUCTION

29

30 Our work focuses on identifying bioactive natural products (NPs) with potential therapeutic
31 application as new drugs, or that could allow the discovery of new druggable targets. Polycyclic
32 polyprenylated acylphloroglucinols (PPAPs) are a class of NPs built around an
33 acylphloroglucinol core substituted by several prenyl or geranyl groups. PPAPs usually differs
34 by their degree of oxidation and cyclization, but also by their stereochemistry [1]. Recently, we
35 showed that PPAPs, and namely guttiferone J, exerted a strong modulating effect on the Major
36 Histocompatibility Complex (MHC) biomolecules [2]. MHC are involved in both the
37 physiological and pathological immune response. Influencing the expression of such
38 biomolecules could present a new therapeutic opportunity for immunotherapy, particularly to
39 prevent graft-rejection and fight cancers [3, 4]. Aiming at using such bioactive NPs as tools in
40 the discovery of targets useful in immunomodulation, we embarked on the isolation of
41 structurally diverse PPAPs. PPAPs are usually isolated from the Clusiaceae, Calophyllaceae
42 and Hypericaceae families, with more than 600 compounds reported so far [1]. We thus decided
43 to focus on these families in order to find structurally diverse PPAPs. 124 dichloromethanic
44 and methanolic extracts from the bark, leaf and sometimes fruit of 17 Malaysian *Garcinia*
45 species (30 different batches in total) were obtained. LC-UV-MS² analyses helped to select the
46 extracts containing high amounts of PPAPs; among them was a *Garcinia bancana* bark extract.
47 *Garcinia bancana* Miq. is a species of trees belonging to the Clusiaceae family. It is growing
48 in the area of peninsular Malaysia, Sumatra and Borneo [5]. To our knowledge, few researches
49 have been done on this particular species [6, 7], making it interesting to study. We had two
50 objectives in mind: characterize the *Garcinia bancana* species, and if necessary, focus on

51 PPAPs purification along the way for possible further biological testing. In order to do so,
52 instead of using a traditional phytochemical analysis, we use dereplication methods that allow
53 to quickly identify known NPs inside complex mixtures, avoiding time-consuming unnecessary
54 purification steps. Several methods are available for a dereplication analysis, among which
55 high-performance liquid chromatography (HPLC) coupled with mass spectrometry (MS). In
56 this method, molecular weights and fragmentation patterns within a mixture are compared with
57 reference data [8]. While proved very efficient, and exploited to build molecular networks [9,
58 10], it was not adapted for the dereplication of mixtures rich in PPAPs since stereoisomers are
59 often present [1]. Indeed, their separation using LC and differentiation based on their molecular
60 weight or fragmentation pattern may be a hard task. A dereplication method based on the ^{13}C -
61 NMR analysis of the sample was thus used [11]. Using ^{13}C -NMR data and DEPT experiments,
62 MixONat compares chemical shifts of ^{13}C , type by type (i.e. quaternary carbons, methines,
63 methylenes, methyls) with the ones of selected NPs in a database (DB).

64

65 The crude dichloromethanic extract obtained from the bark of *Garcinia bancana* was
66 fractionated by flash chromatography. The resulting fractions were analyzed using ^{13}C -NMR
67 and additional DEPT-135 and 90 experiments. Each batch of spectra was submitted to the
68 dereplication program MixONat, using a *Garcinia* DB and a PPAPs-only DB. Based on
69 MixONat proposals, the present study focused on the crude extract and 4 of its fractions
70 supposed to contain PPAPs (F8, F9, F10 and F12).

71 **2. MATERIALS AND METHODS**

72

73 **2.1. Plant material**

74

75 The *Garcinia bancana* bark was collected in May 2003 around Berang, Terengganu. The plant
76 was identified by the botanist Mr. Teo Leong Eng and the voucher specimen (KL5033) was
77 deposited at the Herbarium of the Department of Chemistry, University of Malaya, Kuala
78 Lumpur, Malaysia.

79

80 **2.2. Extraction, fractionation and purification**

81

82 1.5 g of bark powder were successively extracted by sonication (2 h) with dichloromethane to
83 obtain, after solvent evaporation, approximatively 20 g of dry extract. Around 5.5 g of the latter
84 were fractionated using normal phase flash chromatography on silica gel (Chromabound® flash
85 RS 40 SiOH). The mobile phase was (A) cyclohexane and (B) ethyl acetate, with a flow of 48
86 mL/min and the following gradient: t = 0 min, B = 0%; t = 30 min, B = 20%; t = 120 min, B =
87 30%. The following 14 fractions were obtained: F1 (141.3 mg), F2 (162.4 mg), F3 (72.4 mg),
88 F4 (8.2 mg), F5 (133.9 mg), F6 (34.1 mg), F7 (53.3 mg), F8 (822.1 mg), F9 (172.6 mg), F10
89 (272.8 mg), F11 (28.0 mg), F12 (452.6 mg), F13 (69.8 mg), F14 (89.4 mg).

90

91 Around 100 mg of F8 were fractionated using normal flash chromatography on silica gel
92 (Chromabound® flash RS 40 SiOH). The mobile phase was (A) cyclohexane and (B) ethyl
93 acetate, with a flow of 48 mL/min and the following isocratic system over 120 min: B = 5%.
94 The following 6 fractions were obtained: F8-1 (13.2 mg), F8-2 (15.7 mg), F8-3 (17.8 mg), F8-
95 4 (11.9 mg), F8-5 (6.3 mg), F8-6 (12.9 mg).

96

97 11.9 mg of F8-4 were purified using semi preparative HPLC (Agilent HP 1100 Series, Agilent
98 Technologies, Les Ulis, France) on a reverse phase column (Hypersil Gold PFP, 150 x 10 mm
99 5µm), using a 50 mg/mL concentration for the injection (100 µL), with a 73% methanol / 27%
100 water + 0.1% formic acid system over 50 min (flow: 4.7 mL/min). Fractions were collected
101 using the Agilent Technologies 1260 Infinity G1364C fraction collector and the ChemStation
102 for LC 3D software for automatic UV peak detection (diode array detector G13115A). It yielded
103 1.5 mg of guttiferone F (**1**) and 2.8 mg of xanthochymol (**2**).

104

105 125.0 mg of F9 were purified using the same semi preparative HPLC with the same reverse
106 phase column, using a 50 mg/mL concentration for the injection (100 µL), with a 50%
107 acetonitrile / 50% water + 0.1% formic acid system over 40 min (flow: 4.7 mL/min). The
108 fraction collector, DAD and software were the same as previously mentioned. It yielded 6.0 mg
109 of one of the rotamers of garcinialiptone A (**3**).

110

111 50.0 mg of F10 were purified using the same semi preparative HPLC with the same reverse
112 phase column, using a 50 mg/mL concentration for the injection (100 µL), with a 70% methanol
113 / 30% water + 0.1% formic acid system over 60 min (flow: 4.7 mL/min). The fraction collector,
114 DAD and software were the same as previously mentioned. It yielded 2.3 mg of 30-epi-
115 cambogin (**4**), 4.1 mg of (-)-cycloxanthochymol (**5**) and 0.9 mg of garcicowin C (**6**).

2.3. NMR analyses

^{13}C -NMR experiments of the fractions and the pure compounds were recorded in either deuterated methanol + 0.1% trifluoroacetic acid and/or pyridine- d_5 using the JEOL 400 MHz YH spectrometer (Jeol Europe). Parameters were 30 000 scans for the ^{13}C spectra and 15 000 scans for the DEPT 135 experiments for about 30 mg of sample.

2.4. Database

Both databases that were used for this dereplication work are predictive ones made with the C + H NMR Predictor software from ACD/Labs [12]. The *Garcinia* DB was built by gathering 718 natural products reported in the *Garcinia* genus on the Dictionary of Natural Products [13]. The database was already made available in a previous work [14]. The stereochemistry information on the structures is not always present in the *Garcinia* DB, meaning that it was not taken in account for the prediction. Thus, when a structure is proposed by the algorithm, data on the stereoisomers also need to be checked. The PPAPs DB was built using the table of the 652 naturally occurring PPAPs exported as a SDF [1, 15]. As recommended by the authors of this table, it is important to note that some of the structures do not have their full stereochemistry drawn. Some of them also have been represented as the opposite stereoisomer.

2.4. MixONat program

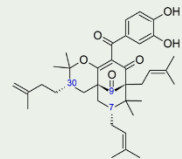
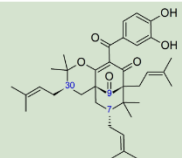
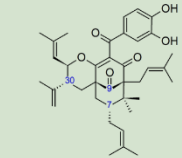
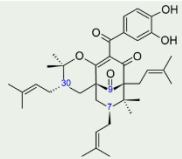
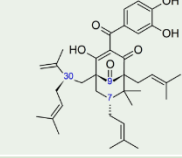
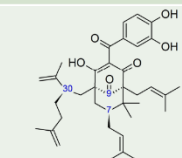
The program was used with the recommended parameters: tolerance of 1.3 ppm, tolerance incrementation, DEPT alignment of 0.02 ppm, and equivalent carbons allowed. The computing time for each analysis was around 30 seconds. The program is freely available [11].

141 **3. RESULTS AND DISCUSSION**

142

143 ^{13}C -NMR dereplication of the crude DCM extract from *Garcinia bancana* bark using MixONat
144 and the Garcinia DB suggested PPAPs as the major NPs with a high score (number of matched
145 chemical shifts out of the total number of carbons in the molecule) (**Table S1**). Supposed to
146 contain PPAPs not yet referenced in our chemical library, the extract was selected for further
147 investigations, i.e. extensive phytochemical analyses. Using the same spectral data (**Figure S1**)
148 but an appropriate DB containing all the naturally occurring PPAPs previously described [15],
149 it appeared that the extract should contain isoxanthochymol or its stereoisomers, e.g.
150 cycloxanthochymol (**5**), 30-epicambogin (**4**), 7-epi-isogarcinol (**Figure 1**, **Figure S2 and S3**).

Polyprenylated polycyclic acylphloroglucinols identification from *Garcinia bancana* bark using the ^{13}C -NMR dereplication program MixONat

Rank	Name	Score	Structure
1	(+)-Cycloxanthochymol	0.97	
2	(-)-Cycloxanthochymol	0.97	
3	Cambogin	0.95	
4	Isoxanthochymol	0.95	
5	30-epi-cambogin	0.95	
6	Nujiangefolin C	0.95	
7	Garcicowin C	0.95	
8	Garcinialiptone B	0.95	
9	Coccinone B	0.95	
10	7-epi-isogarcinol	0.92	
11	Garcicowin D	0.92	
12	Guttiferone F	0.92	
13	Garcinol	0.92	
14	Guttiferone E	0.92	
15	7-epi-coccinone B	0.92	
16	Coccinone C	0.89	
17	Epunctanone	0.89	
18	14-deoxyisogarcinol	0.89	
19	14-deoxygarcinol	0.89	
20	7-epi-garcinol	0.89	
21	Coccinone G	0.89	
22	Xanthochymol	0.89	

151

Figure 1: Results of the dereplication process with MixONat using ¹³C-NMR (pyridine-d₅), additional DEPT-135 and 90 data and PPAPs DB. Equivalent carbons were allowed. In bold, PPAPs that were later confirmed inside of the extract.

To validate such hypotheses, a manual comparison of chemical shifts with literature data (i.e. references) is required for NPs suggested in the first ranks [11]. However, this work is not always easy regarding PPAPs as NMR spectra are described either in methanol-d₄ + TFA 0.1% [16, 17] or pyridine-d₅ [16, 18]. NMR spectra on crude extracts were thus registered both in methanol-d₄ and pyridine-d₅. The careful examination of the first suggestions made by MixONat allowed the identification of guttiferone F (**1**) [17], xanthochymol (**2**) [19], 30-epi-cambogin (**4**) [17], (-)-cycloxanthochymol (**5**) [16], garcicowin C (**6**) [20] and 7-epi-isogarcinol [18] (**Tables S2, S3, S4, S6, S7 and S8**) Due to the low sensitivity of ¹³C-NMR, to refine the composition of the extract and identify minor NPs, a fractionation step is required. The crude extract was thus fractionated and dereplicated using the same method.

Dereplication analysis of F8's carbon (**Figure S4**) and DEPT 135 spectra using MixONat suggested mostly PPAPs structures with a high score with the *Garcinia* DB (**Table 1, Figure S5 and S6**).

Polyprenylated polycyclic acylphloroglucinols identification from *Garcinia bancana* bark using the ¹³C-NMR dereplication program MixONat

Table 1: Results for F8 with MixONat displaying the rank, score and name of each structure.

Molecule names in bold are the ones that were confirmed inside the fraction.

Garcinia DB (718 NPs)	Rank	Name	Score	Rank	Name	Score	PPAPs DB (652 PPAPs)
	1	(+)-Garcinialiptone A	0.92	1	(+)-Garcinialiptone A	0.92	
	2	Garciniagifolone A	0.92	2	(-)-Garcinialiptone A	0.92	
	3	Guttiferone K	0.89	3	Garciniagifolone A	0.92	
	4	(+)-Guttiferone G	0.88	4	Coccinone B	0.89	
	5	Garcinielliptone O	0.88	5	7-epi-coccinone B	0.89	
	6	Guttiferone E	0.87	6	Guttiferone D	0.88	
	7	Garcicowin A	0.86	7	Schomburgkianone D	0.87	
	8	Aristophenone A	0.85	8	Guttiferone F	0.87	
	9	13-O-Methylgarcinol	0.85	9	Garcinol	0.87	
	10	Oblongifolin B	0.84	10	Guttiferone E	0.87	
	11	Xanthochymol	0.84	11	Coccinone G	0.87	
	12	Guttiferone P	0.84	12	Xanthochymol	0.87	

A new type of PPAP skeleton, different from **1-2**, **4-6**, presenting an additional cyclization, *i.e.* garcinialiptone A and garciniagifolone A, was suggested by both DBs (**Figure 2**).

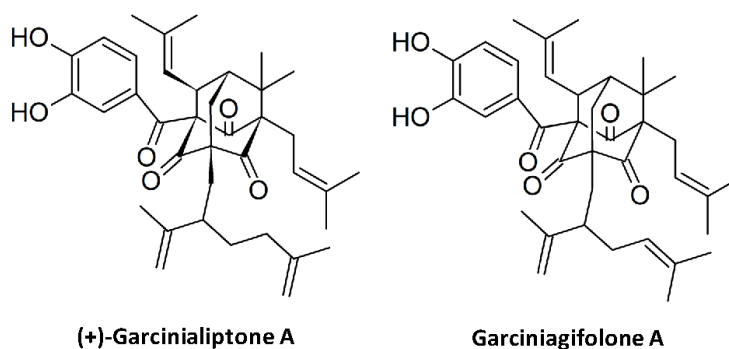


Figure 2: Structures of (+)-garcinialiptone A (**3**) and garciniagifolone A

Manual comparison of experimental data for the first hypotheses quickly allowed the identification of guttiferone F (**1**) [17] (stereoisomer of the suggested guttiferone E by the

Garcinia DB), xanthochymol (**2**) [19] and one of the rotamers of garcinialiptone A (**3**) [16] (Table S5).

The results for F9 using ¹³C (Figure S7) and DEPT 135 analysis gave again a strong hint on the composition of the fraction (Table 2, Figure S8 and S9). Checking for the experimental chemical shifts of the first suggested NPs led to the identification of guttiferone F (**1**) [17], xanthochymol (**2**) [19], the two rotamers of garcinialiptone A (**3**) [16], 30-epi-cambogin (**4**) [17] (stereoisomer of isoxanthochymol) and (-)-cycloxanthochymol (**5**) [20]. We can observe that the PPAPs DB ranks guttiferone F and xanthochymol further away than the Garcinia DB does. This is due to the fact that the PPAPs DB is of course much more exhaustive in terms of PPAPs structures and has thus more structures that will fit the matching requirements, like other stereoisomers of the correct molecules. The important thing to focus on here is the score which is still very high (almost 90%), even at rank 25.

Table 2: Results for F9 with MixONat displaying the rank, score and name of each structure. Molecule names in bold are the ones that were confirmed inside the fraction.

Garcinia DB (718 NPs)	Rank	Name	Score	Rank	Name	Score	PPAPs DB (652 PPAPs)
	1	Cycloxanthochymol	0.97	1	(+)-Cycloxanthochymol	0.97	
	2	(+)-Garcinialiptone A	0.97	2	(-)-Cycloxanthochymol	0.97	
	3	Garciniagifolone A	0.89	3	(+)-Garcinialiptone A	0.97	
	4	Guttiferone K	0.95	4	(-)-Garcinialiptone A	0.97	
	5	(+)-Guttiferone G	0.93	5	Garciniagifolone A	0.97	
	6	13-O-Methylgarcinol	0.92	6	Cambogin	0.95	
	7	Guttiferone E	0.92	7	Isoxanthochymol	0.95	
	8	Xanthochymol	0.92	8	Coccinone G	0.95	
	9	Garciyunnanin A	0.92	
	10	Garcicoxin B	0.90	12	30-epi-cambogin	0.92	
	11	Guttiferone M	0.89	22	Guttiferone F	0.89	
	12	Isoxanthochymol	0.89	25	Xanthochymol	0.89	

198

199 For F10, based on the ^{13}C (**Figure S10**) and DEPT 135 spectra, the 12 first hypotheses made
200 by MixONat (**Table 3, Figure S11 and S12**) were manually compared to the literature to
201 confirm the presence of 30-epi-cambogin (**4**) [17], (-)-cycloxanthochymol (**5**) [20] and
202 garcicowin C (**6**) [20] (stereoisomer or homomer of garcinialiptone B [1]) in the fraction.
203 Garcicowin C and garcinialiptone B are ranked 49 and 50 respectively out of the 652 molecules
204 in the PPAPs. Just like the previous example, this is due to the high number of stereoisomers
205 of the first molecules (cycloxanthochymol, isoxanthochymol) that can be matched using the
206 same chemical shifts and thus are ranked before garcicowin C and garcinialiptone B. Even if
207 the score is still good for those molecules (almost 80%), it can be noted that the score for the
208 garcinialiptone B is not the same in both DBs. This is due to the fact that the information on the
209 stereochemistry was not present in the *Garcinia* DB, but was in the PPAPs DB, thus resulting
210 in 2 different predictions.

211

212 **Table 3:** Results for F10 with MixONat displaying the rank, score and name of each structure.

213 Molecule names in bold are the ones that were confirmed inside the fraction.

Garcinia DB (718 NPs)	Rank	Name	Score	Rank	Name	Score	PPAPs DB (652 PPAPs)
	1	Cycloxanthochymol	0.97	1	(+)-Cycloxanthochymol	0.97	
	2	Isoxanthochymol	0.87	2	(-)-Cycloxanthochymol	0.97	
	3	Guttiferone E	0.87	3	Cambogin	0.92	
	4	(+)-Guttiferone K	0.84	4	Isoxanthochymol	0.92	
	5	Garcinielliptin oxide	0.83	5	Coccinone	0.89	
	6	Subellione	0.83	6	Garcimultiflorone H	0.88	
	7	(+)-Guttiferone G	0.83	7	30-epi-cambogin	0.87	
	8	Subelliptenone B	0.83	8	Guttiferone I	0.87	
	9	13-O-Methylgarcinol	0.82	9	Oblogifolin T	0.87	
	10	Aristophenone A	0.82	
	11	Xanthochymol	0.82	49	Garcicowin C	0.79	
	12	Garcinialiptone B	0.82	50	Garcinialiptone B	0.79	

Polyprenylated polycyclic acylphloroglucinols identification from *Garcinia bancana* bark using the ^{13}C -NMR dereplication program MixONat

Finally, analysis of ^{13}C (**Figure S13**) and DEPT 135 data from F12 led to the quick identification of 7-epi-isogarcinol [18] (stereoisomer of isoxanthochymol) suggested at the first rank as the main constituent (**Table 4, Figure S14 and S15**).

Table 4: Results for F12 with MixONat displaying the rank, score and name of each structure.

Molecule names in bold are the ones that were confirmed inside the fraction.

Garcinia DB (718 NPs)	Rank	Name	Score	Rank	Name	Score	PPAPs DB (652 PPAPs)
	1	Isoxanthochymol	0.89	1	7-epi-isogarcinol	0.92	
	2	Cycloxanthochymol	0.87	2	14-deoxy-7-epi-isogarcinol	0.89	
	3	13,14-Dideoxyisogarcinol	0.84	3	30-epi-cambogin	0.89	
	4	Garsubellin A	0.8	4	Epunctanone	0.87	
	5	Guttiferone T	0.79	5	Cambogin	0.87	
	6	Oblongifoliagarcinin D	0.78	6	Isoxanthochymol	0.87	
	7	Garcinielliptone N	0.77	7	(+)-Cycloxanthochymol	0.87	
	8	Garcicowin A	0.86	8	(-)-Cycloxanthochymol	0.87	
	9	Garcinialiptone B	0.76	9	13,14-didehydroxy-7-epi-isogarcinol	0.84	
	10	13-O-Methylgarcinol	0.74	10	Coccinone C	0.84	
	11	Garsubellin B	0.74	11	Symphonone B	0.84	
	12	(+)-Guttiferone K	0.74	12	Coccinone D	0.82	

The program MixONat successfully helped us to quickly identify the complex structures inside the crude extract and the different fractions (**Table 5**). The correct molecules were always ranked among the first results with a good score (more than 80%), validating once more the quality of the prediction [14]. The given hypotheses always made the structural class of the molecules obvious. It was also easy to recognize when a new type of PPAP was present in one of the fractions as the propositions were orienting to a different skeleton [e.g. garcinialiptone A (3)]. We were thus able to characterize a part of the composition of the extract solely based on

229 the crude extract's spectrum. A fractionation step also allowed the identification of the minor
230 compounds (rotamers of garcinialiptone A) which signals were not visible in the crude extract's
231 spectrum. Two different DBs were used in this work, one base on chemotaxonomy and another
232 dedicated to a type of NPs. This work allowed to highlight their complementarity. Using a
233 chemotaxonomic database (Garcinia DB), generally orients towards one stereoisomer of the
234 correct compound, depending on which structures are inside. The chemical shifts of all
235 stereoisomers of each hypothesis must thus be checked, but still allowing their differentiation
236 by reference comparison. With a more specific database (PPAPs DB), the correct stereoisomer
237 is usually directly suggested, as such DBs tend to reach exhaustivity. The downside is that the
238 correct structure can also be "drowned" among all the other possible stereoisomers present in
239 the database, lowering the rank of the molecule and might thus not be considered as a valid
240 candidate (*e.g.* garcicowin C in F12, ranked 49/652, but still with a high score of 80%).
241 The main problem encountered was the limited literature data on PPAPs. Indeed, ^{13}C -NMR
242 chemical shifts are often reported in only one deuterated solvent for each molecule, but not
243 always the same for the whole class. If analyzed in another one, it is impossible to
244 unambiguously identify the compound by comparison, PPAPs being structurally very close to
245 one another. It is very likely to find in an extract, several PPAPs, each one reported once, and
246 in a different solvent, forcing the researcher to redo multiple time the ^{13}C -NMR analysis in
247 different solvents in order to compare. Conversely, it underlines the interest of DB of predicted
248 chemical shifts for such preliminary approaches.
249

Table 5: Molecules identified in the different fractions. Underlined molecules names are the ones that were already identified from the crude extract spectrum (**Figure S16**).

Molecules	F8	F9	F10	F12
<u>Guttiferone F</u>	☑	☑		
<u>Xanthochymol</u>	☑	☑		
Garcinialiptone A (rotamer 1)	☑	☑		
Garcinialiptone A (rotamer 2)		☑		
<u>30-epi-cambogin</u>		☑	☑	
<u>(-)-Cycloxanthochymol</u>		☑	☑	
<u>Garcicowin C</u>			☑	
<u>7-epi-isogarcinol</u>				☑

4. CONCLUSION

The dereplication method using MixONat allowed the identification of 8 PPAPs, namely guttiferone F (1), xanthochymol (2), 2 rotamers of garcinialiptone A (3), 30-epi-cambogin (4), (-)-cycloxanthochymol (5), garcicowin C (6) and 7-epi-isogarcinol. Sample characterization using this method is possible very early, directly on the crude extract, or with only one chromatographical step for some of the minor compounds. ^{13}C -NMR also allowed the differentiation of stereoisomers inside the mixtures, that would not have been possible using techniques like LC-MSⁿ. It only requires compound purification when the molecule is new or valuable for further biological investigation. As shown in this work, using the appropriate database with it, the program would be able to quickly identify know compounds in different type of plant extracts, hence reducing the amount of time necessary for usual phytochemical work.

268 **SUPPORTING INFORMATION**

269 ^{13}C -NMR tables (**Table S2-S8**) for the isolated molecules can be found in the SI.

270 REFERENCES

- 271 1. Yang, X.W., R.B. Grossman, and G. Xu, *Research progress of polycyclic polyprenylated*
272 *acylphloroglucinols*. Chem Rev, 2018. **118**(7): p. 3508-3558.
- 273 2. Rouger, C., et al., *Prenylated polyphenols from Clusiaceae and Calophyllaceae with*
274 *immunomodulatory activity on endothelial cells*. PLoS One, 2016. **11**(12): p. e0167361.
- 275 3. Sharma, P. and J.P. Allison, *The future of immune checkpoint therapy*. Science, 2015.
276 **348**(6230): p. 56-61.
- 277 4. Topalian, S.L., et al., *Mechanism-driven biomarkers to guide immune checkpoint blockade in*
278 *cancer therapy*. Nat Rev Cancer, 2016. **16**(5): p. 275-287.
- 279 5. Agriculture, U.S.D.o. *Garcinia bancana* Miq. [cited 2019; Available from:
280 <https://npgsweb.ars-grin.gov/gringlobal/taxonomydetail.aspx?id=70991>].
- 281 6. Rukachaisirikul, V., et al., *An antibacterial biphenyl derivative from Garcinia bancana* MIQ.
282 Chem Pharm Bull (Tokyo), 2005. **53**(3): p. 342-343.
- 283 7. Jantan, I. and S.H. Goh, *Flavonoids, xanthenes and triterpenes of the leaves and heartwood of*
284 *Garcinia bancana* Miq. J Sains Malaysiana, 1995. **24**: p. 23-30.
- 285 8. Chervin, J., et al., *Targeted dereplication of microbial natural products by high-resolution MS*
286 *and predicted LC retention time*. J Nat Prod, 2017. **80**(5): p. 1370-1377.
- 287 9. Yang, J.Y., et al., *Molecular networking as a dereplication strategy*. J Nat Prod, 2013. **76**(9): p.
288 1686-1699.
- 289 10. Fox Ramos, A.E., et al., *Revisiting previously investigated plants: a molecular networking-*
290 *based study of Geissospermum laeve*. J Nat Prod, 2017. **80**(4): p. 1007-1014.
- 291 11. Bruguère, A., et al., *MixONat software for mixture dereplication base on 13C NMR and DEPT*
292 *experiments*. Anal Chem, Submitted.
- 293 12. *ACD/NMR Predictors*. [cited 2019; Available from:
294 https://www.acdlabs.com/products/adh/nmr/nmr_pred/].
- 295 13. *Dictionary of Natural Products*. [cited 2016; Available from: dnpc.chemnetbase.com/].
- 296 14. Bruguère, A., et al., *(13)C-NMR dereplication of Garcinia extracts: Predicted chemical shifts*
297 *as reliable databases*. Fitoterapia, 2018. **131**: p. 59-64.
- 298 15. Yang, X.W., R.B. Grossman, and G. Xu. *Table of naturally occurring PPAPs*. [cited 2019;
299 Available from: <http://www.uky.edu/~rbgros1/PPAPs/allPPAPs.html>].
- 300 16. Zhang, L.J., et al., *Cytotoxic polyisoprenyl benzophenonoids from Garcinia subelliptica*. J Nat
301 Prod, 2010. **73**(4): p. 557-562.
- 302 17. Fuller, R.W., et al., *Guttiferone F, the first prenylated benzophenone from Allanblackia*
303 *stuhlmannii*. J Nat Prod, 1999. **62**(1): p. 130-132.
- 304 18. Marti, G., et al., *Antiplasmodial benzophenones from the trunk latex of Moronobea coccinea*
305 *(Clusiaceae)*. Phytochemistry, 2009. **70**(1): p. 75-85.
- 306 19. Iinuma, M., et al., *Antibacterial activity of some Garcinia benzophenone derivatives against*
307 *methicillin-resistant Staphylococcus aureus*. Biol Pharm Bull, 1996. **19**(2): p. 311-314.
- 308 20. Xu, G., et al., *Cytotoxic acylphloroglucinol derivatives from the twigs of Garcinia cowa*. J Nat
309 Prod, 2010. **73**(2): p. 104-108.

310

Supporting information

Polyprenylated polycyclic acylphloroglucinols identification from *Garcinia bancana* bark using the ^{13}C -NMR dereplication program MixONat

Antoine Bruguère^a, Séverine Derbré^{a,*}, Sow Tein Leong^b, Noor Aimi Othman^b, Robert B. Grossman^c, Khalijah Awang^b, Pascal Richomme^a

^a *SONAS SFR QUASAV, University of Angers, France.*

^b *Department of Chemistry, Faculty of sciences, University of Malaya, Malaysia*

^c *Department of Chemistry, University of Kentucky, Lexington, United States*

Table S1: Results for the crude extract of *Garcinia bancana* using MixONat and the Garcinia DB. Molecules in bold are the molecules (or stereoisomers of the molecules) which signals were later confirmed inside of the extract. MixONat parameters: tolerance = 1.3 ppm, equivalent carbons allowed.

Rank	Name	Score	Rank	Name	Score	Rank	Name	Score
1	Cycloxanthochymol	0.97	11	Garcinielliptone N	0.91	21	Guttiferone O	0.89
2	Guttiferone K	0.97	12	Guttiferone P	0.91	22	Hyperscabrin	0.89
3	(+)-Guttiferone G	0.95	13	83725-45-5	0.90	23	Gardicowin A	0.89
4	Garciyunnanin A	0.95	14	Garcinielliptin oxide	0.90	24	Oblongifolin D	0.88
5	Isoxanthochymol	0.95	15	13-O-Methylgarcinol	0.90	25	Garciyunnanin B	0.88
6	Garcicowin B	0.93	16	Oblongifolin E	0.89	26	Garcinielliptone J	0.87
7	Guttiferone M	0.92	17	Guttiferone E	0.89	27	Oblongifolin B	0.87
8	Guttiferone N	0.92	18	Guttiferone T	0.89	28	Garcinialiptone B	0.87
9	900501-08-8	0.92	19	1126644-31-2	0.89	29	Xanthochymol	0.87
10	Garcimultiflorone D	0.92	20	Garcinialiptone D	0.89	30	Oxyguttiferone K	0.87

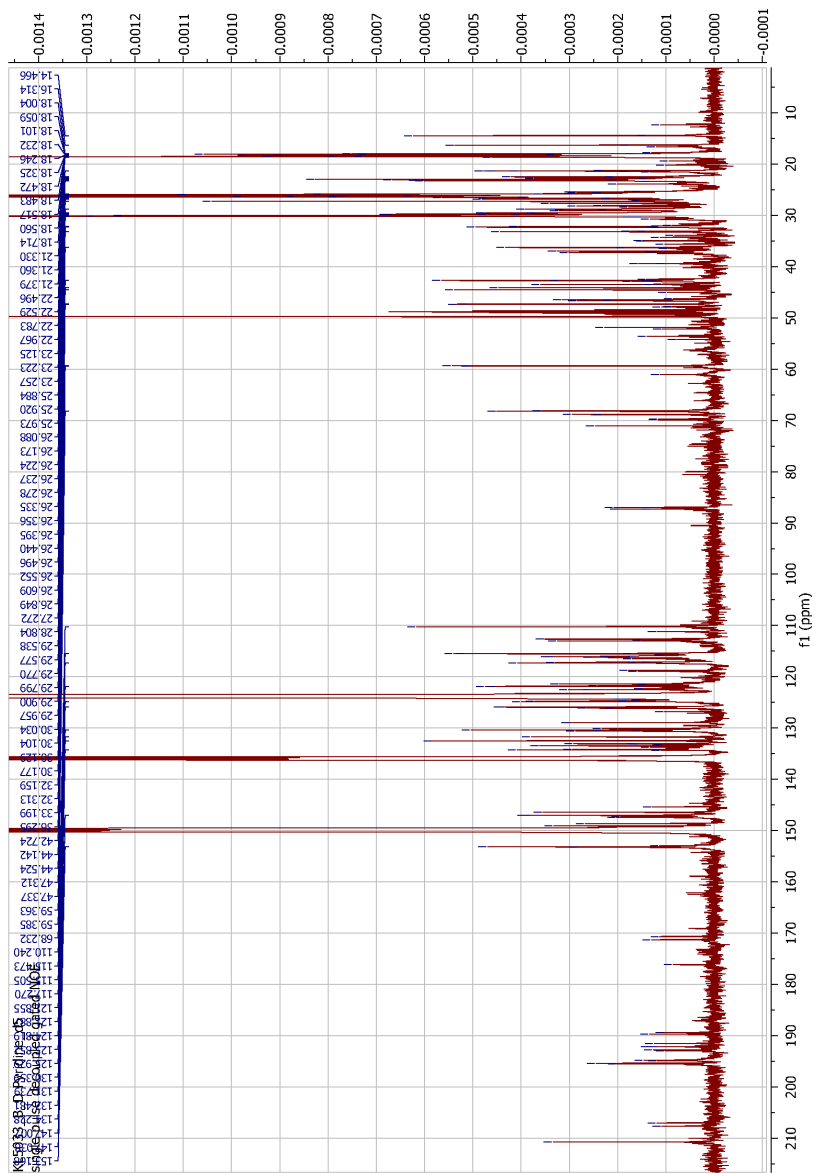


Figure S1. ^{13}C -NMR spectrum of the *Garcinia bancana* crude bark extract (pyridine- d_5)

Supporting information

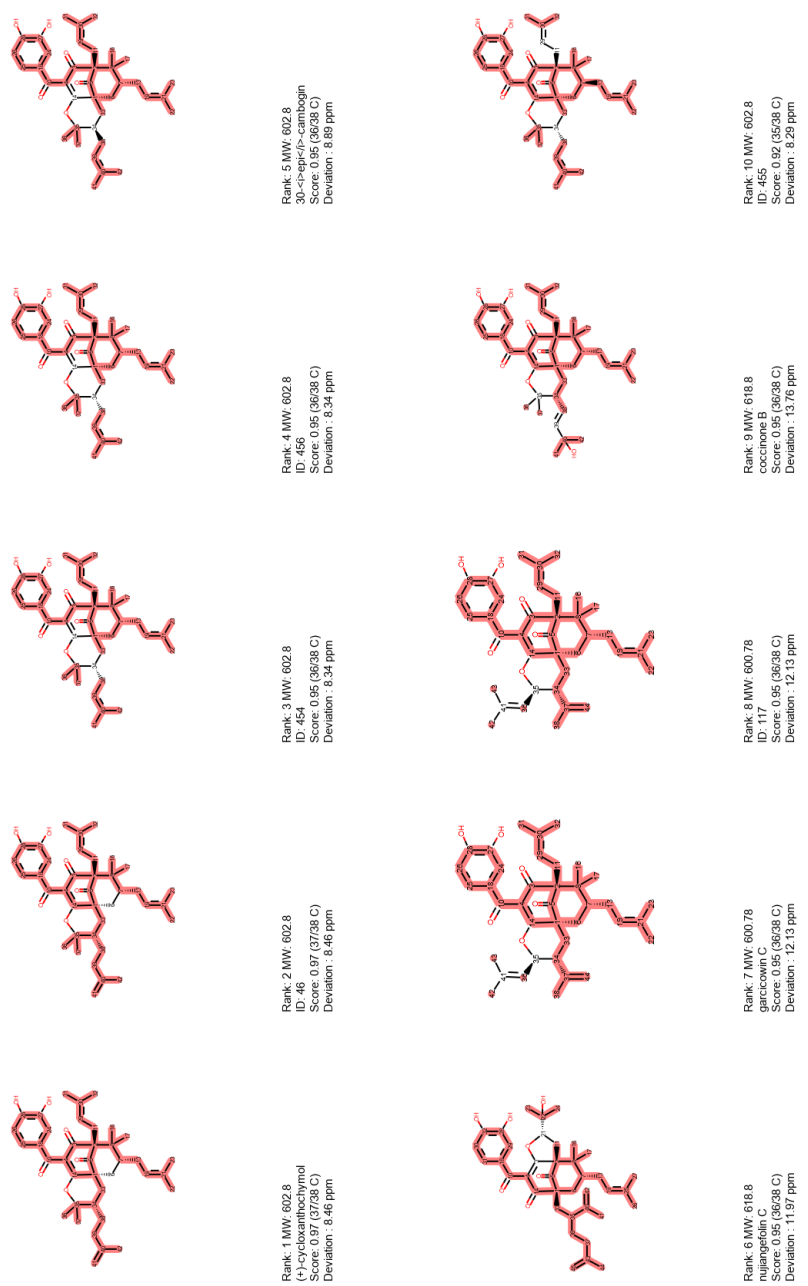


Figure S2: Results (ranks 1-10) for the *Garcinia bancana* extract (pyridine-d5) using the PPAps database. MixONat parameters: tolerance = 1.3

ppm, equivalent carbons allowed.

Supporting information



Figure S3: Results (ranks 11-20) for the *Garcinia bancana* extract (pyridine-d5) using the PPAPs database. MixONat parameters: tolerance = 1.3 ppm, equivalent carbons allowed.

S5

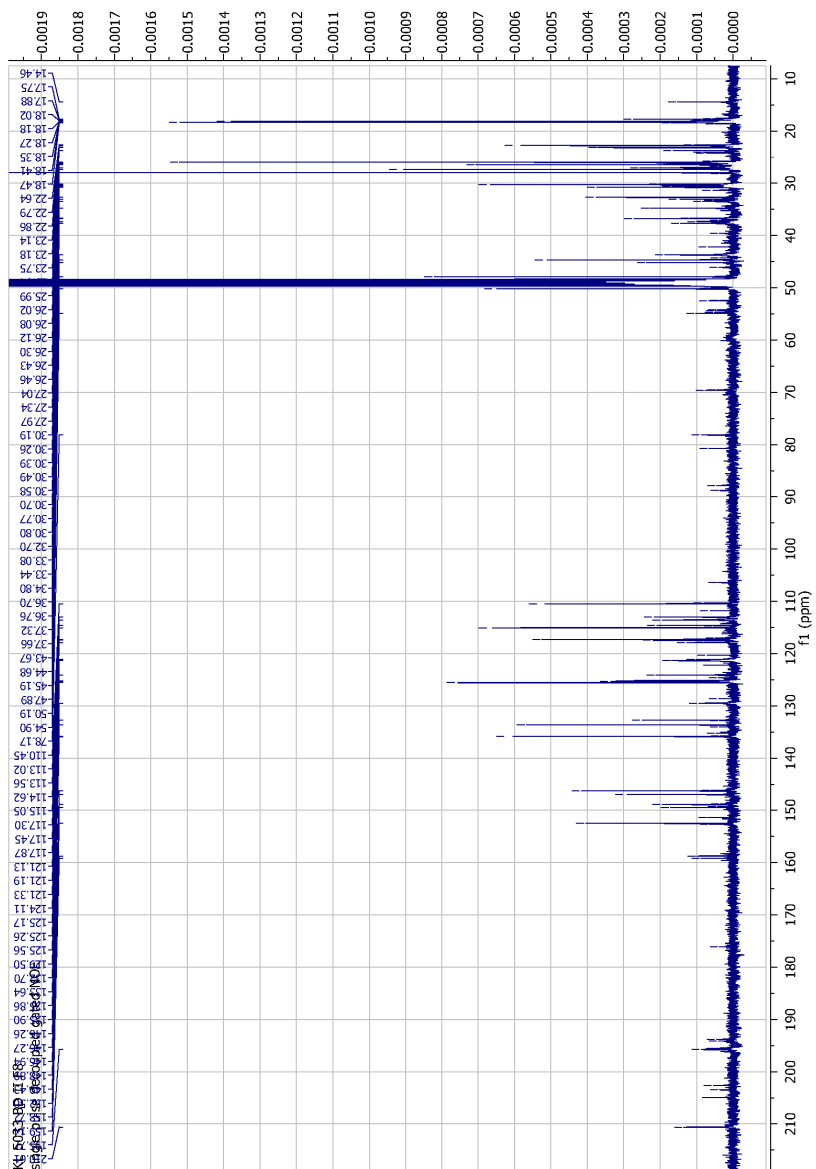


Figure S4: ^{13}C -NMR spectrum of the *Garcinia bancana* F8 (methanol- d_4)

Supporting information

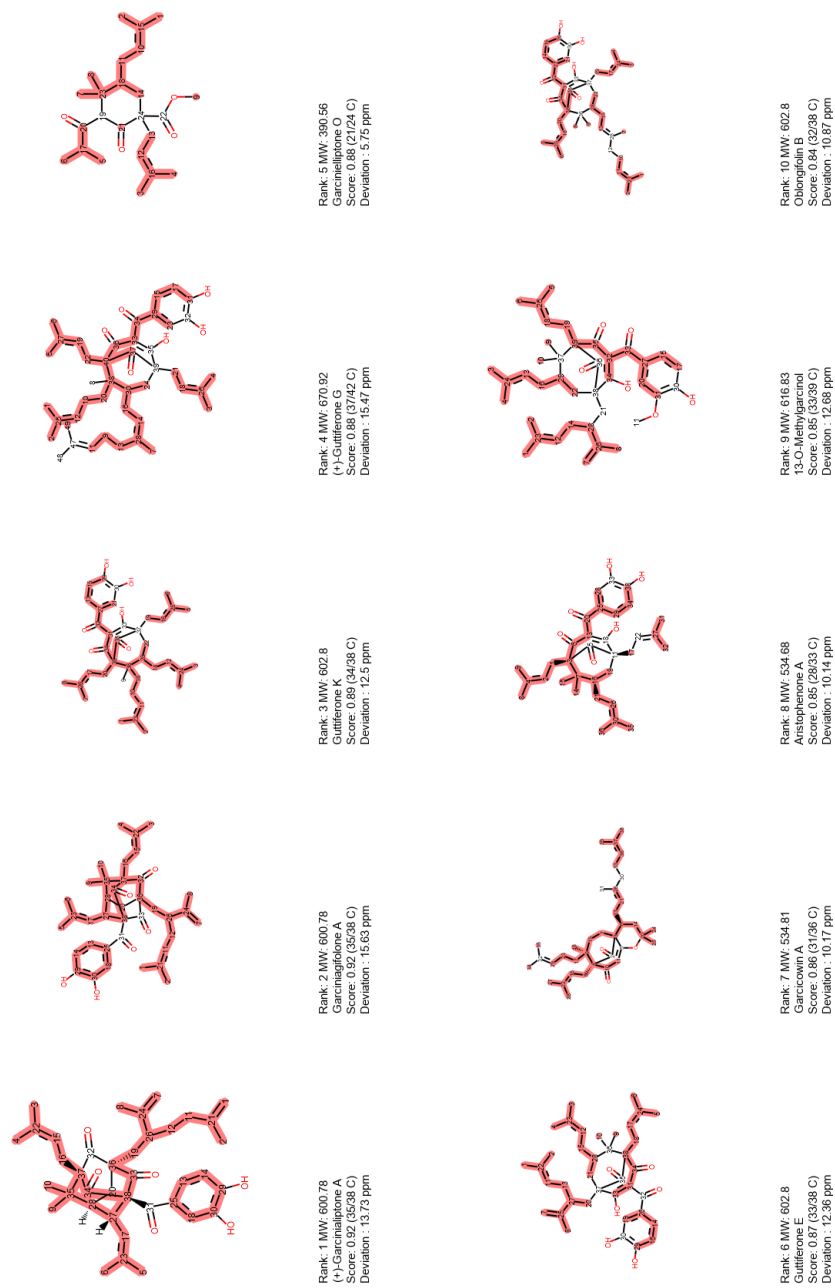


Figure S5: Results (ranks 1–10) for the *Garcinia bancana* F8 (methanol-d4) using the Garcinia database. MixONat parameters: tolerance = 1.3

ppm, equivalent carbons allowed.

S7

Supporting information

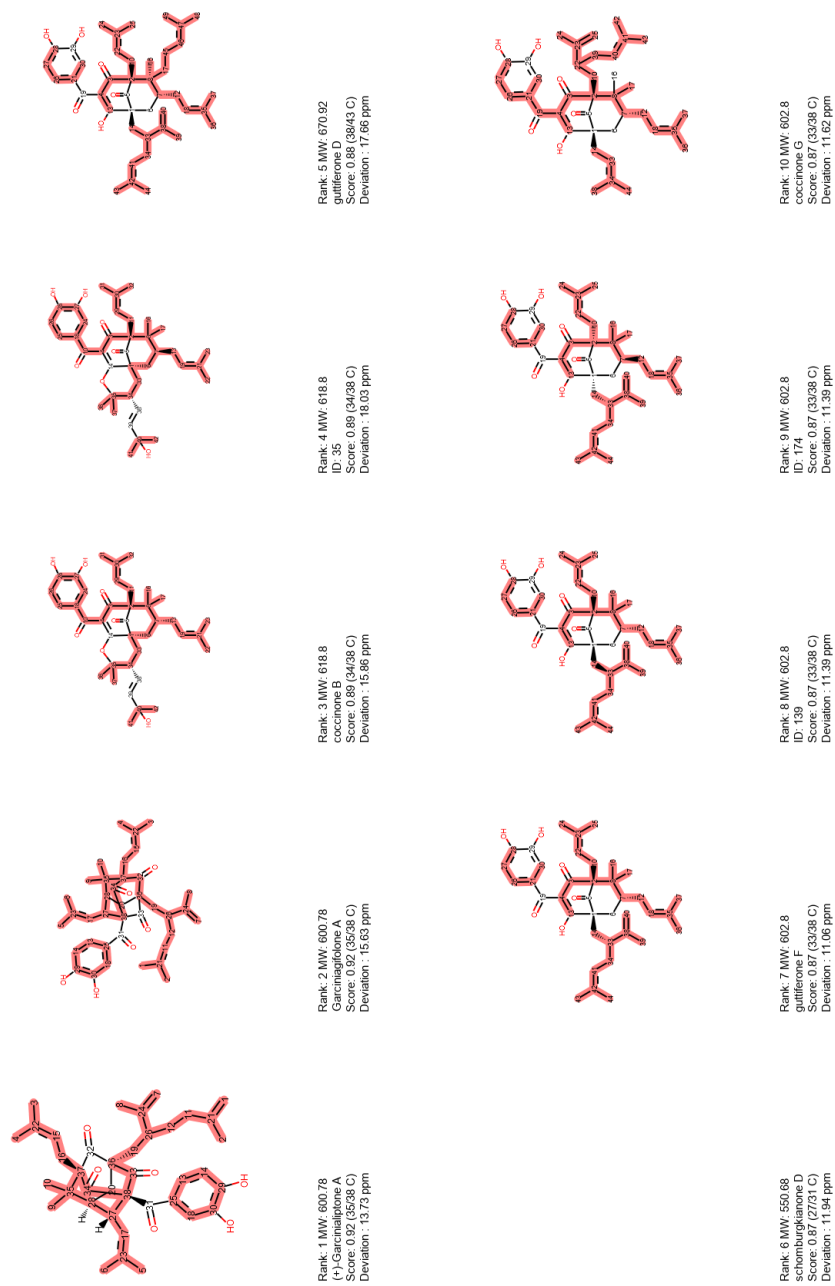


Figure S6: Results (ranks 1-10) for the *Garcinia bancana* F8 (methanol-d4) using the PPAPs database. MixONat parameters; tolerance = 1.3

ppm, equivalent carbons allowed. A bug is preventing the structure of schomburgkianone D to be displayed properly.

S8

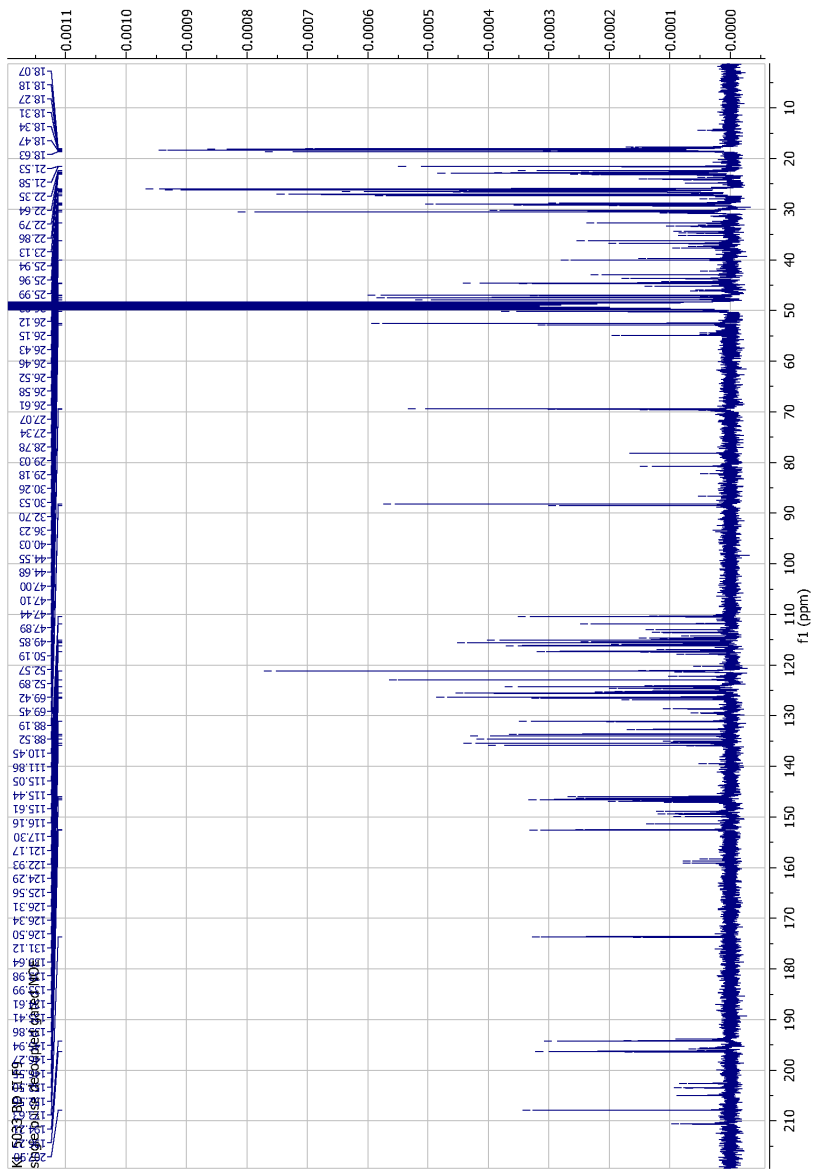


Figure S7: ^{13}C -NMR spectrum of the *Garcinia bancana* F9 (methanol-d4)

Supporting information

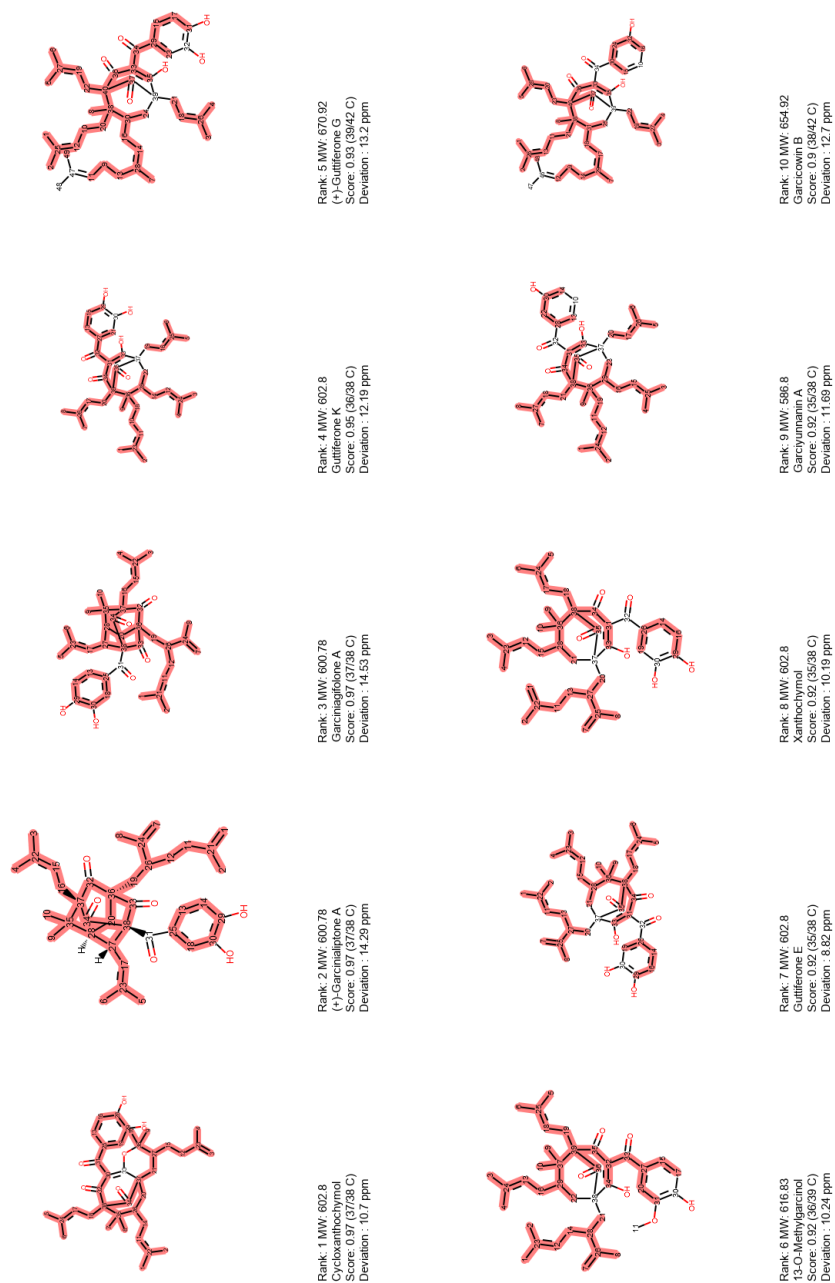


Figure S8: Results (ranks 1–10) for the *Garcinia bancana* F9 (methanol-d4) using the Garcinia database. MixONat parameters: tolerance = 1.3

ppm, equivalent carbons allowed.

S10

Supporting information

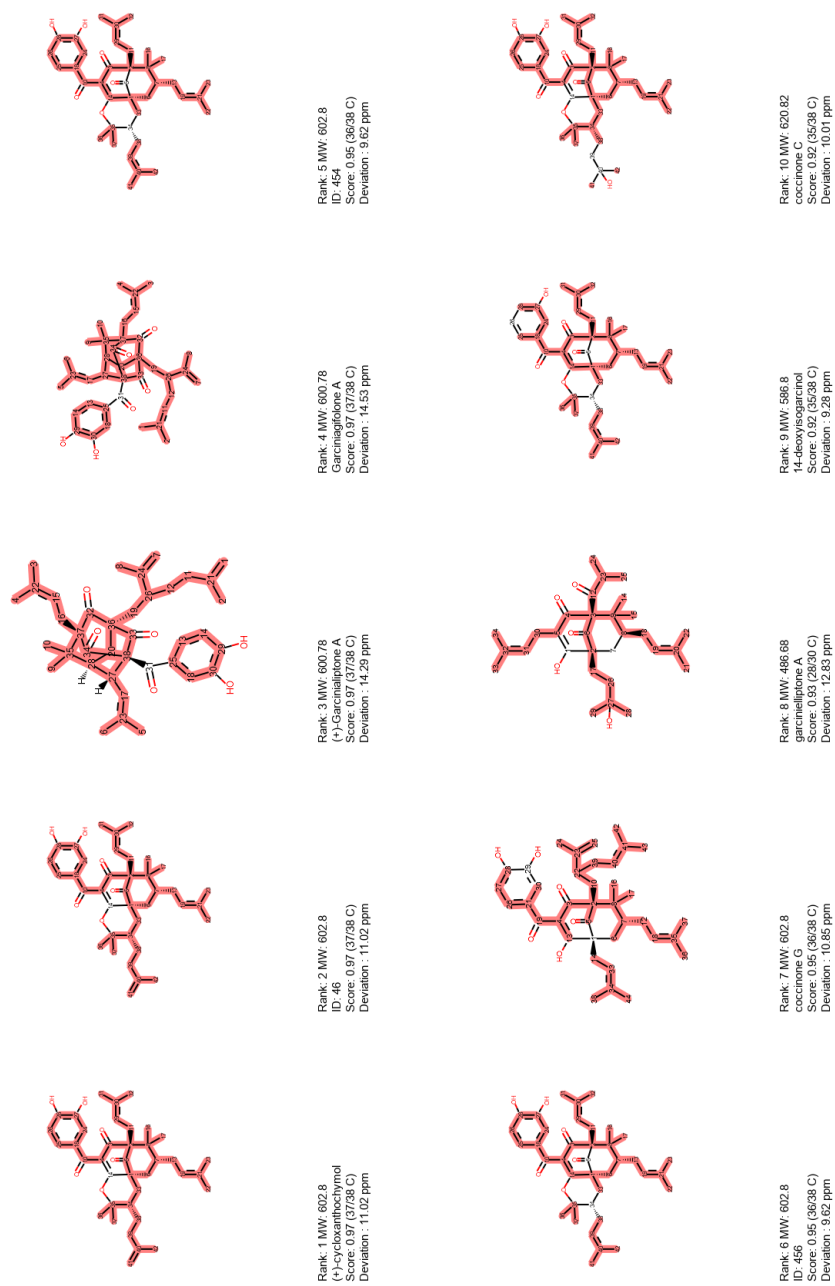


Figure S9: Results (ranks 1-10) for the *Garcinia bancana* F8 (methanol-d4) using the PPAPs database. MixONat parameters: tolerance = 1.3

ppm, equivalent carbons allowed.

S11

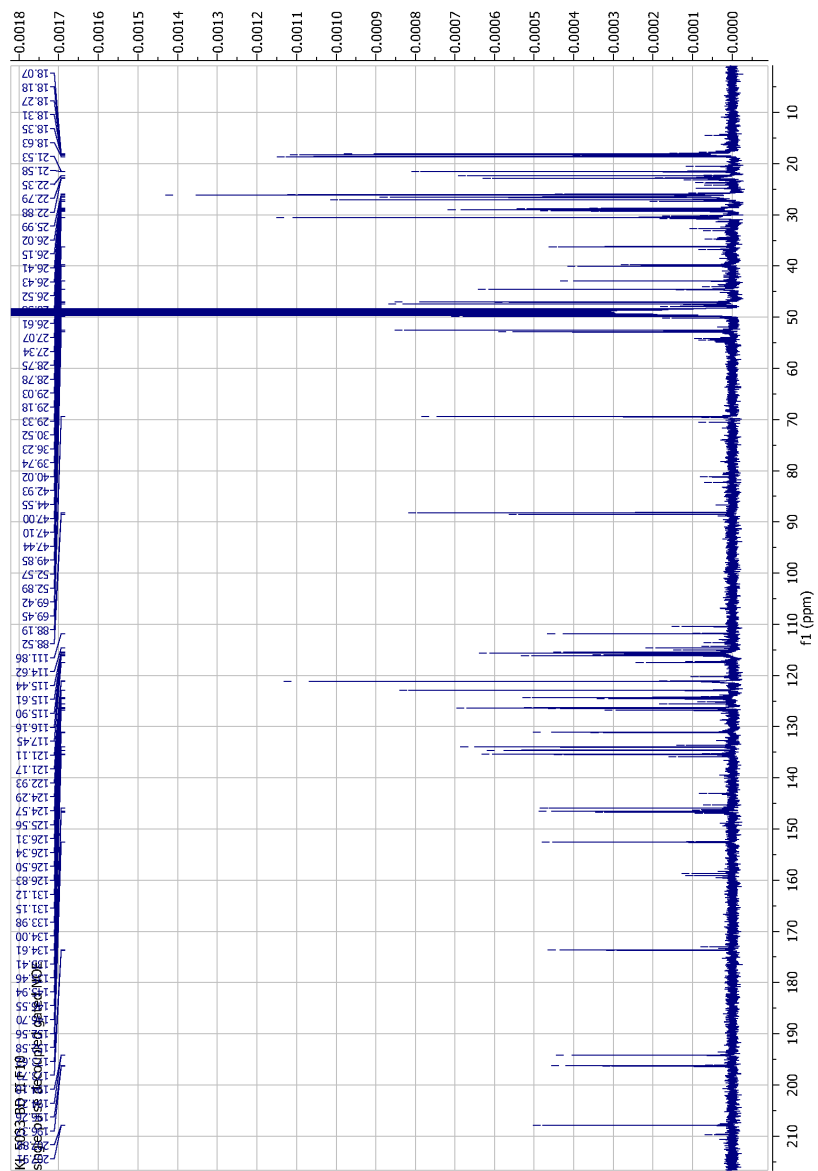


Figure S10: ^{13}C -NMR spectrum of the *Garcinia bancana* F10 (methanol- d_4)

S12

Supporting information

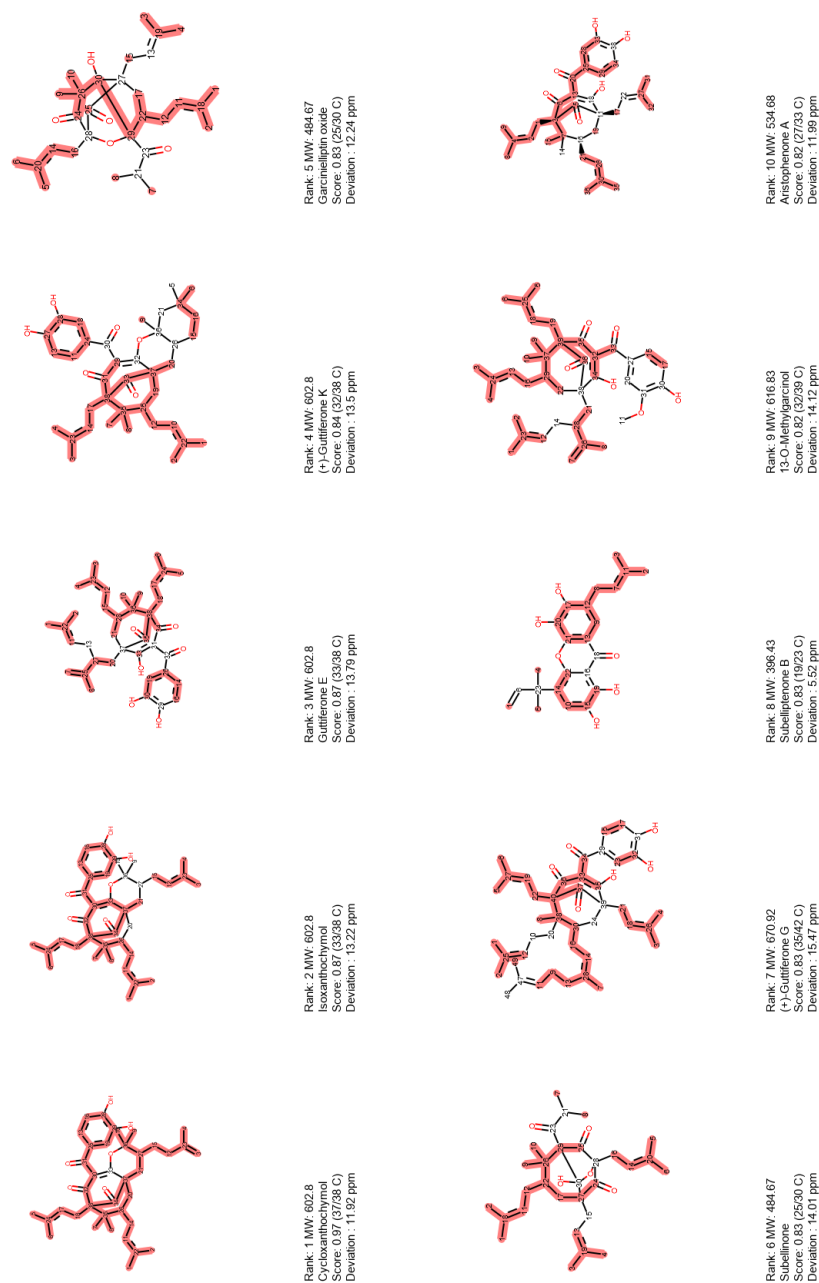


Figure S11: Results (ranks 1–10) for the *Garcinia bancana* F10 (methanol-d4) using the Garcinia database. MixONat parameters: tolerance = 1.3

ppm, equivalent carbons allowed.

S13

Supporting information

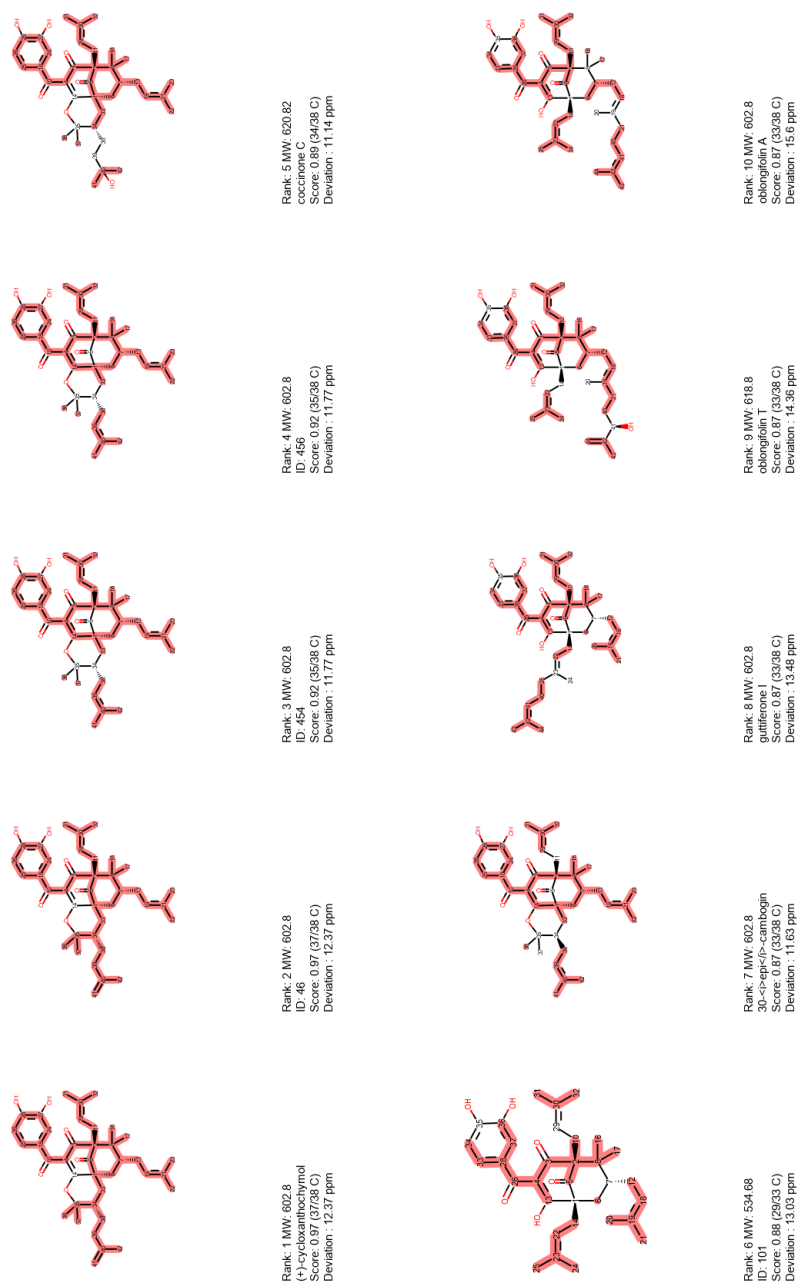
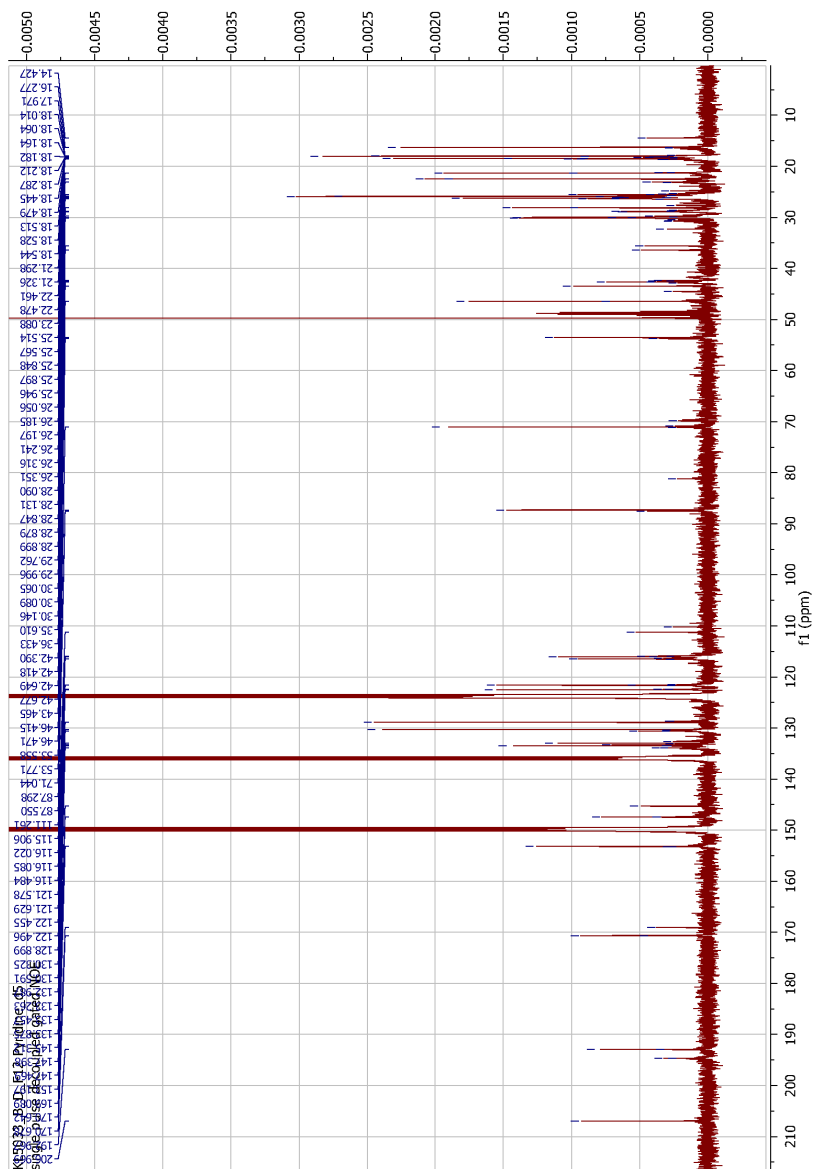


Figure S12: Results (ranks 1-10) for the *Garcinia bancana* F10 (methanol-d4) using the PPAps database. MixONat parameters: tolerance = 1.3

ppm, equivalent carbons allowed.

S14

Figure S13: ^{13}C -NMR spectrum of the *Garcinia bancana* F12 (pyridine- d_5)

Supporting information

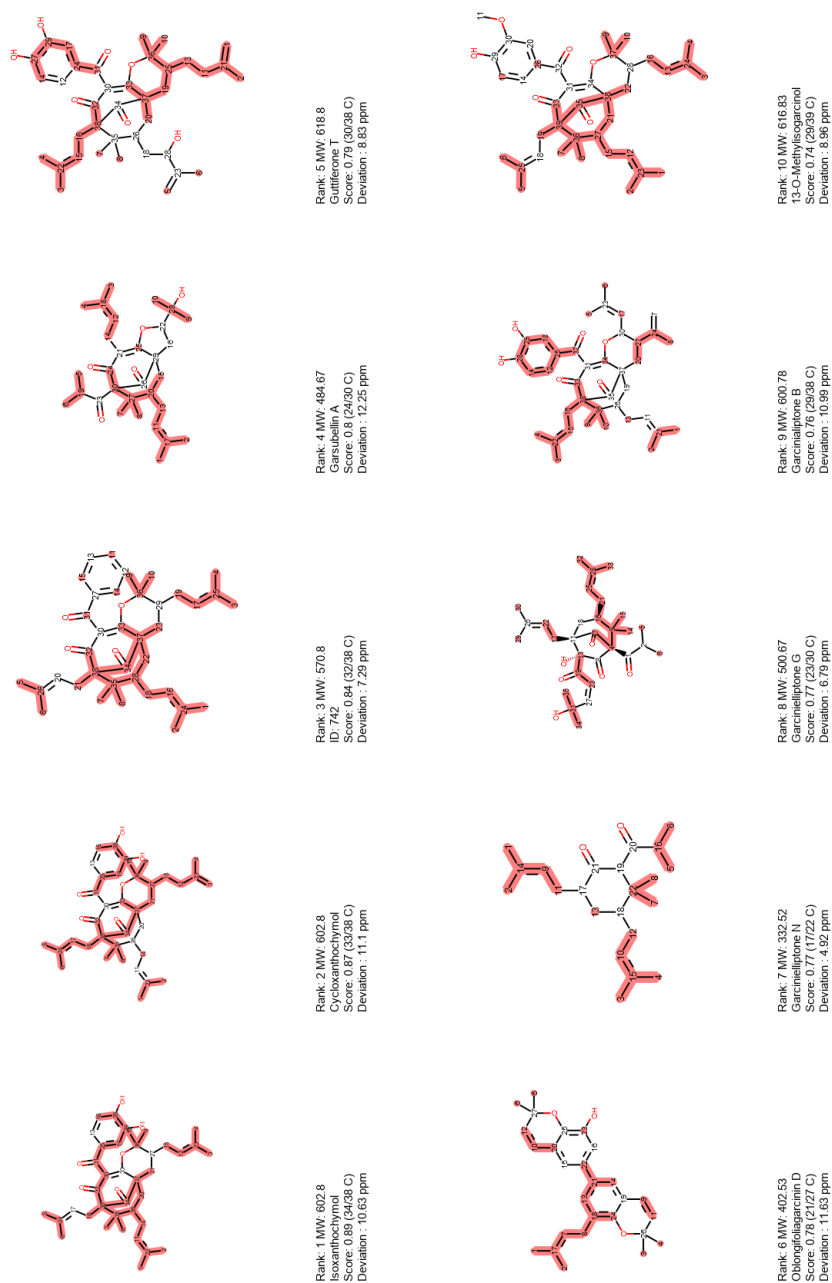


Figure S14: Results (ranks 1-10) for the *Garcinia bancana* F12 (pyridine-d5) using the Garcinia database. MixONat parameters: tolerance = 1.3 ppm, equivalent carbons allowed.

S16

Supporting information

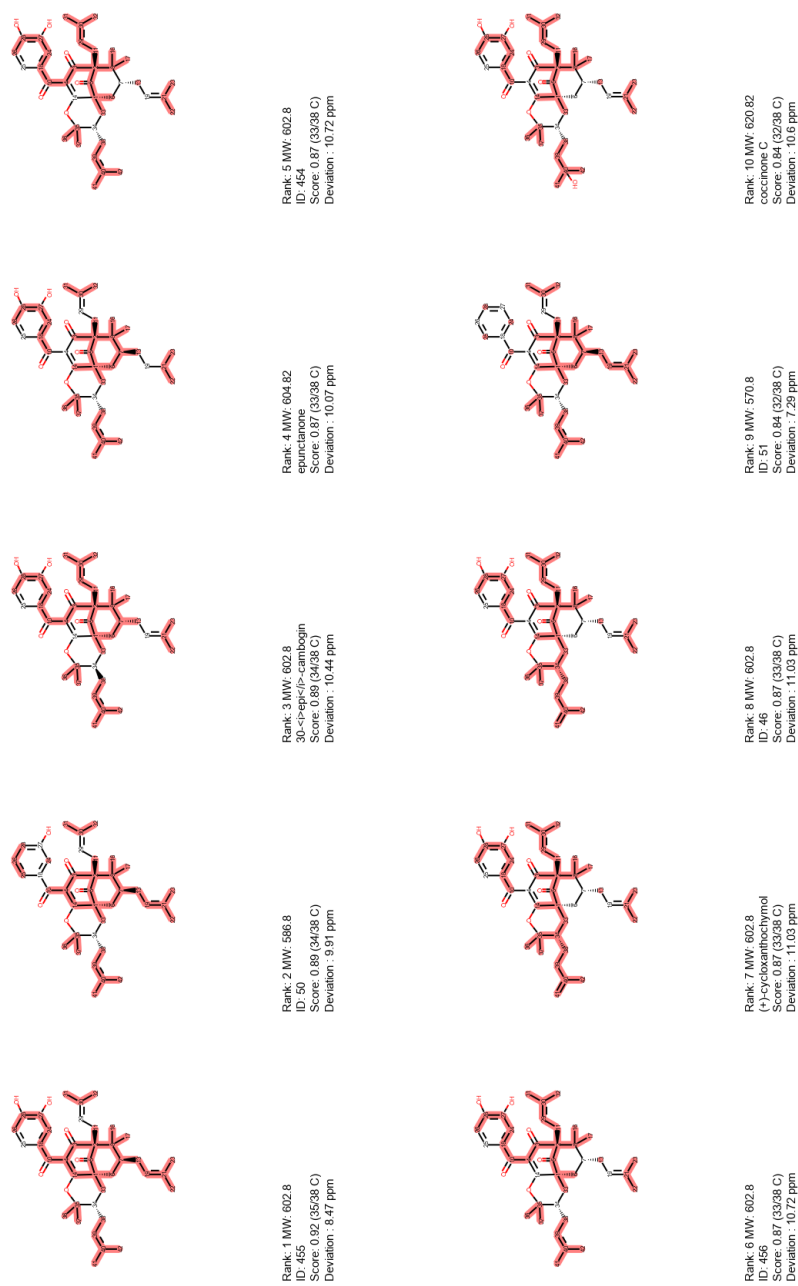


Figure S15: Results (ranks 1-10) for the *Garcinia bancana* F12 (pyridine-d5) using the PPAPs database. MixONat parameters: tolerance = 1.3

ppm, equivalent carbons allowed.

S17

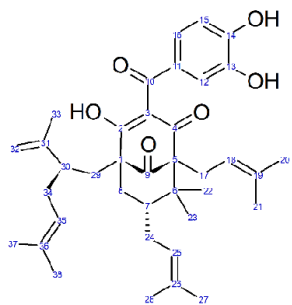


Table S2: ¹³C-NMR chemical shifts (methanol-d4) of guttiferone F.

Guttiferone F (methanol-d4)							
1	/	11	129.5	21	18.3	31	149.4
2	196.1	12	117.3	22	23.2	32	113.0
3	117.9	13	146.3	23	27.3	33	18.2
4	193.8	14	152.6	24	30.3	34	33.5
5	69.5	15	115.1	25	125.6	35	124.1
6	50.2	16	125.3	26	133.6	36	132.7
7	47.9	17	27.0	27	25.9	37	26.0
8	43.7	18	121.3	28	18.2	38	18.2
9	210.6	19	135.9	29	37.3		
10	195.5	20	26.4	30	45.2		

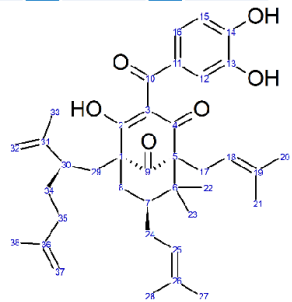


Table S3: ¹³C-NMR chemical shifts (methanol-d4) of xanthochymol.

Xanthochymol (methanol-d4)							
1	/	11	129.5	21	18.3	31	148.9
2	196.1	12	117.5	22	23.3	32	113.6
3	117.8	13	146.9	23	27.3	33	17.8
4	194.1	14	152.5	24	30.3	34	32.7
5	69.6	15	115.1	25	125.6	35	36.8
6	50.1	16	125.2	26	133.6	36	149.5
7	49.0	17	27.1	27	26.0	37	110.5
8	43.7	18	121.2	28	18.2	38	22.8
9	210.6	19	135.9	29	37.7		
10	195.6	20	26.4	30	44.7		

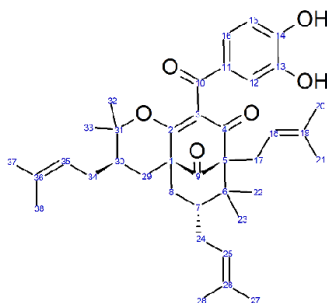


Table S4: ^{13}C -NMR chemical shifts (methanol- d_4) of 30-epi-cambogin.

30-epi-cambogin (methanol- d_4)							
1	52.6	11	131.1	21	18.3	31	88.2
2	173.7	12	116.2	22	22.9	32	29.0
3	126.5	13	146.6	23	27.1	33	21.6
4	196.3	14	152.6	24	30.5	34	30.5
5	69.4	15	115.6	25	126.3	35	122.9
6	47.0	16	124.3	26	134.0	36	134.6
7	47.5	17	26.6	27	26.0	37	26.1
8	40.0	18	121.2	28	18.6	38	18.1
9	207.9	19	135.4	29	29.2		
10	194.2	20	26.5	30	44.6		

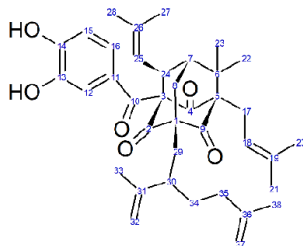
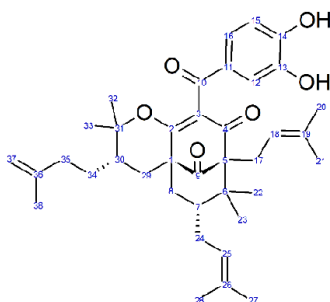


Table S5: ^{13}C -NMR chemical shifts (pyridine- d_5) of garcinialiptone A.

Garcinialiptone A (pyridine- d_5)							
1	68.6	11	127.6	21	25.8	31	148.2
2	202.1	12	117.0	22	22.0	32	113.5
3	79.6	13	146.9	23	22.6	33	17.6
4	202.4	14	152.2	24	51.5	34	32.2
5	77.1	15	115.0	25	121.9	35	35.7
6	53.8	16	123.9	26	133.6	36	146.0
7	47.6	17	23.4	27	25.6	37	109.8
8	44.4	18	120.8	28	18.0	38	22.6
9	204.0	19	133.5	29	33.7		
10	192.2	20	18.1	30	42.6		

**Table S6:** ^{13}C -NMR chemical shifts (pyridine- d_5) of (-)-cycloxanthochymol.

(-)-Cycloxanthochymol (pyridine- d_5)							
1	52.4	11	130.8	21	18.8	31	87.5
2	171.6	12	116.4	22	23.1	32	21.6
3	127.4	13	147.9	23	27.1	33	29.1
4	195.2	14	153.8	24	30.4	34	29.2
5	69.2	15	116.4	25	126.4	35	35.9
6	47.0	16	/	26	133.4	36	145.7
7	46.9	17	26.7	27	26.5	37	111.4
8	39.5	18	121.7	28	19.0	38	22.9
9	208.0	19	134.6	29	28.7		
10	193.1	20	26.8	30	42.8		

Table S7: ^{13}C -NMR chemical shifts (methanol- d_4) of (-)-cycloxanthochymol.

(-)-Cycloxanthochymol (methanol- d_4)							
1	52.9	11	131.1	21	18.3	31	88.6
2	173.8	12	115.9	22	22.9	32	21.5
3	126.8	13	146.7	23	27.0	33	28.8
4	196.4	14	152.6	24	30.5	34	29.3
5	69.5	15	115.4	25	126.3	35	36.2
6	47.4	16	124.6	26	134.0	36	146.0
7	47.1	17	26.6	27	26.1	37	11.9
8	39.7	18	121.2	28	18.6	38	22.3
9	207.9	19	135.5	29	28.7		
10	194.2	20	26.6	30	42.9		

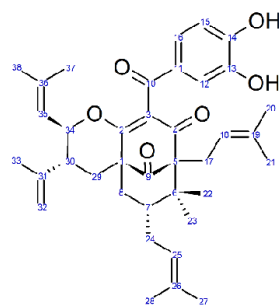


Table S8: ¹³C-NMR chemical shifts (methanol-d4) of garcicowin C

Garcicowin C (methanol-d4)							
1	/	11	131.2	21	18.3	31	81.2
2	173.1	12	116.2	22	22.7	32	122.8
3	124.3	13	146.5	23	27.1	33	143.1
4	196.1	14	152.6	24	30.6	34	18.6
5	70.5	15	115.8	25	126.4	35	25.8
6	47.6	16	124.1	26	134.0	36	145.3
7	47.5	17	26.3	27	26.1	37	114.3
8	38.5	18	121.2	28	18.0	38	20.5
9	209.7	19	135.3	29	34.5		
10	193.6	20	26.4	30	44.1		

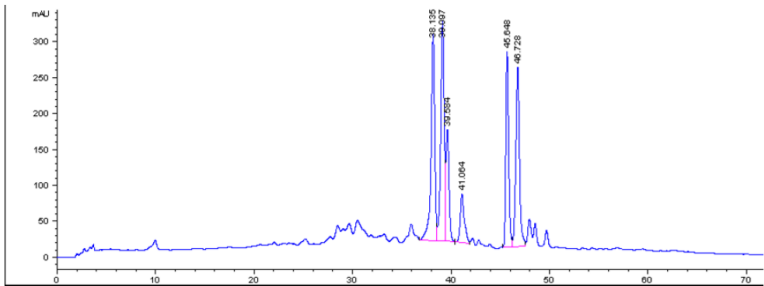


Figure S16: HPLC-UV (280 nm) chromatogram of the *Garcinia bancana* bark extract.

Hypersil Gold PFP (150 mm x 4.6 mm), from methanol 55% / 45% water + 0.1% formic acid
to methanol 90% / 10% water + 0.1 formic acid in 60 minutes with a flow of 1 mL/min.

4.1.3. Éléments de discussion

MixONat propose des structures correctes, aussi bien à partir de l'extrait brut que des différentes fractions chromatographiques. Les molécules réellement contenues dans l'extrait apparaissent avec des score élevés (au moins 80%), ce qui permet encore une fois de valider la qualité de la prédiction. Les hypothèses formulées par le programme permettent non seulement d'identifier rapidement une classe structurale mais également la présence de squelettes particuliers, comme c'est le cas ici avec la garcinialiptone A.

On peut noter que dans cet exemple, une étape de fractionnement a été nécessaire afin d'identifier les rotamères de la garcinialiptone A, composés minoritaires, puisque leurs signaux n'étaient pas directement visibles dans l'extrait brut.

L'utilisation de deux types distincts de banques de données a permis de mettre en valeur leur complémentarité. Avec des bases construites sur des critères chimiotaxonomiques (banque de données *Garcinia*), l'utilisateur est orienté vers des stéréoisomères de la molécule correcte, ou la molécule elle-même, en fonction de la richesse de la base. Ceci demande, par sécurité, de vérifier pour chaque molécule proposée, les déplacements de ses stéréoisomères, qui pourront être différenciés grâce aux données de la littérature. En utilisant des bases plus spécifiques (e.g. banque de données PPAPs), la molécule présentant la bonne stéréochimie est souvent directement proposée puisque ce genre de base tend à l'exhaustivité. L'inconvénient de ce type de base est que la structure peut se retrouver « noyée » parmi tous les autres stéréoisomères possibles présents dans la base, comme c'est le cas avec la garcicowin C, classée 49/652 mais présentant quand même un score de 80%.

Conclusion générale et perspectives

L'étude des méthodes de déréplication publiées dans la littérature nous ont permis de choisir la méthode RMN- ^{13}C comme la plus appropriée à notre problématique des PPAPs. Cependant, plusieurs problèmes ont été soulevés.

D'abord, il a fallu déterminer si la construction des bases de données utilisées comme références pouvaient contenir uniquement des données prédites, à défaut de pouvoir construire des bases de données expérimentales. Même si des différences notoires entre différents logiciels de prédiction ont pu être observées, nous avons pu conclure que les logiciels de prédiction, et donc l'utilisation des bases de données construites à partir de ces derniers, étaient suffisamment précis pour permettre des études déréplicatives de qualité.

Ensuite, lors de l'utilisation des différents algorithmes de déréplication disponibles dans la littérature, commercialement ou sur internet, nous avons supposé qu'il était possible d'améliorer leur caractère discriminatoire en complétant l'analyse ^{13}C classique par des expériences additionnelles de DEPT 135 et 90. Celles-ci permettent en effet de différencier les différents types de carbone (quaternaires, tertiaires, secondaires et primaires), et donc de rechercher simultanément deux informations lors de la déréplication : un type de carbone associé à un déplacement chimique particulier. Nous avons donc codé un programme, appelé MixONat, qui permet d'intégrer les informations DEPT lors de la recherche. En plus de ceci, nous avons choisi d'associer prioritairement les déplacements chimiques les plus proches des valeurs présentes dans la base référente, et d'optimiser les résultats de manière locale, après l'étape d'association des déplacements chimiques, afin de réduire le côté « glouton » de l'algorithme. Une interface graphique interactive a également été implantée, afin que les utilisateurs du programme puissent « avoir le contrôle » sur les résultats finaux, et les moduler en fonction de différents éléments. Ceci aussi bien en considération de paramètres expérimentaux (δ_{C} de carbones quaternaires absents, prédictions insuffisamment précises, intensités des signaux incohérentes) que de critères chimiotaxonomiques.

Le programme a d'abord été testé sur plusieurs mélanges de produits naturels de compositions connues, afin d'évaluer son efficacité. Il a été capable de correctement classer les molécules présentes dans chacun des mélanges, avec des scores élevés.

Ces résultats encourageant nous ont permis de valider la méthode, et de l'utiliser pour répondre à notre problématique de départ : l'isolation de nouvelles PPAPs. Grâce à MixONat, les structures des PPAPs présents dans un extrait de *Garcinia bancana* ont pu être identifiées. Les produits majoritaires ont été directement caractérisés à partir de l'extrait brut, tandis qu'une étape de fractionnement a permis d'identifier les composés minoritaires.

Ce travail de doctorat a ainsi conduit au développement de MixONat, un programme de déréplication utilisant la RMN- ^{13}C , capable de rapidement identifier les molécules connues dans des mélanges complexes.

La qualité des hypothèses présentées par le logiciel varie selon plusieurs facteurs. D'abord, la qualité des données initiales, c'est-à-dire celle des spectres RMN. Le programme n'est évidemment pas capable d'associer des pics n'apparaissant pas sur le spectre (produits minoritaires), n'ayant pas été sélectionnés ou ayant été mal référencés. L'autre limite vient de la qualité des bases de données : il n'est pas possible pour le programme de trouver le composé correct si la molécule ne s'y trouve pas, ou si la prédiction des δ_c est mauvaise. Enfin, le caractère « glouton » de l'algorithme qui cherche seulement des solutions optimales à un niveau local, et non général, peut impacter négativement les résultats.

Afin d'améliorer la qualité des hypothèses du programme, plusieurs améliorations pourront y être apportées comme l'ajout de fonctions permettant de filtrer les bases de données (critères chimiotaxonomiques, solvants deutérés...), de pouvoir enrichir les bases de données prédites avec des données expérimentales, ou encore la possibilité de réaliser des « chromatographies virtuelles ».

Annexes

1. Filtre d'intensité des algorithmes de recherche

Il semble important de prendre en compte l'intensité des différents signaux associés aux déplacements chimiques matchés car elle permettrait théoriquement de différencier des signaux appartenant à des molécules différentes. En effet, l'intensité des signaux observée en RMN dépend de la concentration respective des composés présents dans un mélange. Dans le cas où ces composés ont des proportions suffisamment différentes, il est possible de différencier les déplacements chimiques d'une molécule majoritaire de ceux d'une molécule minoritaire. Cependant, l'intensité est également liée à la liaison que font les carbones avec les hydrogènes : on obtient donc des signaux généralement intenses pour les CH₃, liés à 3 protons, et pour des carbones quaternaires, des signaux pouvant être nettement plus faibles que ceux des CH₃. Il est donc possible, en présence d'un mélange d'une molécule majoritaire et d'une minoritaire, que les signaux des carbones quaternaires du composé majoritaire aient une intensité proche des signaux des carbones non quaternaires de la molécule minoritaire, rendant plus difficile la distinction des deux molécules en se basant seulement sur l'intensité des signaux. Tout ceci, sans prendre en compte les phénomènes d'équivalences inter ou intra moléculaires, qui peuvent donner lieu à des déplacements chimiques x fois plus intenses que le reste des signaux, alors qu'ils appartiennent bien à la même molécule (x étant le nombre de carbones partageant la même valeur de déplacement chimique).

Un filtre mathématique, permettant de systématiquement détecter des signaux possédant des intensités « anormales » (donc n'appartenant pas à la même molécule) semble compliqué à mettre en place. Cependant, l'équipe de Reims en propose un dans le script Python pour déréplication qu'ils ont publié [6]. Dans un premier temps, nous avons cherché à comprendre le fonctionnement de ce filtre et essayé d'évaluer son efficacité. Nous avons ensuite tenté à notre tour diverses méthodes pour appliquer un filtre d'intensité robuste lors d'analyses déréplicatives.

1.1. Le filtre d'intensité de DerepCrude

L'algorithme proposé par l'équipe de Reims propose à l'utilisateur l'application d'un filtre d'intensité, qui peut être ajusté à l'aide d'un facteur appelé « **e** ». Le filtre fonctionne ainsi : lors du processus de *matching*, lorsque 2 déplacements chimiques ont déjà été associés pour une molécule, un nouveau déplacement chimique, répondant aux critères sélectionnés par l'utilisateur (c'est-à-dire, rentrant dans la marge autorisée ou *looseness factor*), peut uniquement être associé si son intensité est comprise entre la moyenne des intensités des déplacements chimiques précédemment associés \pm **e** écart(s)-type (cf. 2.3. Fonctionnement de l'algorithme DerepCrude). Les concepteurs conseillent de choisir **e** = 2. Mais il faut rappeler que l'algorithme, avant tout processus de *matching*, trie par ordre croissant les déplacements chimiques. Donc, lors d'une boucle de *matching*, il y a de fortes chances que les 2 premiers déplacements associés soient des déplacements chimiques ayant une valeur faible en ppm correspondant, la plupart du temps, à des CH₃ ou CH₂, donc à des déplacements chimiques associés à des intensités importantes. La moyenne et l'écart-type seront calculés à partir de ces valeurs d'intensité très importantes. Généralement, plus le déplacement chimique du carbone augmente (ppm), plus l'intensité du signal qui lui est associé diminue, car les carbones les plus substitués (d'intensité plus faible par « déficit » de nOe hétéronucléaire) sont aussi les plus déblindés. La moyenne d'intensité étant calculée à partir des valeurs les plus

élevés, lorsque l'algorithme rencontre un carbone quaternaire, il est possible que son intensité associée sorte de la marge des 2 écarts-type recommandés, et soit donc considéré comme appartenant à une molécule différente. Un second exemple théorique de dysfonctionnement que l'on peut imaginer est le cas où les deux premiers déplacements chimiques associés lors du *matching* sont deux déplacements appartenant à 2 molécules différentes, l'une majoritaire et l'autre minoritaire. La moyenne des intensités qui sera calculée sera alors basée sur un signal intense, celui du composé majoritaire, et un signal faible, correspondant au composé minoritaire. La moyenne et l'écart-type associés à ces valeurs représentera donc une gamme de valeurs très large, dans laquelle toutes les autres intensités seront retrouvées. Le filtre ne sera donc plus du tout discriminant.

Afin de vérifier ces théories, et de pouvoir baser l'évaluation de la pertinence du filtre d'intensité sur un exemple concret, nous avons repris les résultats présentés par Bakiri *et al.* dans leur publication [6], obtenu avec un totum alcaloïdique de *Peumus boldus* avec un filtre d'intensité paramétré de façon standard ($e = 2$), et comparé ces résultats avec ceux obtenus « sans filtre d'intensité » ($e = 1000$). La **Tableau 12** présente les résultats de cette expérience; avec **(A)** en bleu le tableau regroupant les 30 premiers résultats avec le filtre d'intensité standard ($e = 2$) et **(B)** en vert le tableau regroupant les 30 premiers résultats sans filtre d'intensité ($e = 1000$). Les différences entre les deux tableaux sont marquées en jaune. On remarque alors que si les 30 premiers résultats sont similaires, 2 nouveaux monoterpènes apparaissent dans « top five » du classement : le limonène (rang 2) et le terpinolène (rang 5). Ces deux molécules sont également décrites dans l'huile essentielle de *Peumus boldus* [34-36]. Après comparaison des signaux de l'extrait avec ceux de la littérature pour ces 2 composés (**Tableau 11**), leur présence est définitivement confirmée. Cet exemple conforte ainsi l'hypothèse selon laquelle la façon dont le filtre d'intensité est programmé peut altérer négativement le processus de déréplication.

Tableau 11: Comparaison des signaux du limonène et du terpinolène avec ceux de l'extrait de boldo.

Limonène	Extrait boldo	Terpinolène	Extrait boldo
150,1	150,1	134,0	134,0
133,7	133,6	128,0	128,1
120,8	120,8	121,6	121,7
108,5	108,6	121,0	121,1
41,2	41,2	31,7	31,7
30,9	31,0	29,7	29,6
30,7	30,6	26,9	26,9
28,1	28,1	26,9	26,9
23,6	23,7	23,4	23,2
20,8	20,7	19,7	19,7

Tableau 12: 30 premiers résultats donnés par l'algorithme DerepCrude lors de la déréplication d'un extrait brut de *Peumus bolus* (**A**) avec filtre d'intensité ou (**B**) sans. En jaune sont surlignées les différences majeures.

A.								
	Molécule	Score		Molécule	Score		Molécule	Score
1	Linalol	1.00	11	3-carene	1.00	21	Alpha-terpinene	1.00
2	Alpha-terpineol	1.00	12	Thymol	1.00	22	2-carene	1.00
3	Alpha-pinene	1.00	13	Isocorydine	1.00	23	Alpha-thujene	1.00
4	Beta-ocimene	1.00	14	Norisocorydine	1.00	24	Sabinene	1.00
5	4-terpineol	1.00	15	Rogersine	1.00	25	Dehydro-1,8-cineole	1.00
6	p-cimene	1.00	16	Boldine	1.00	26	Sabinene hydrate	1.00
7	Caryophyllene	1.00	17	Reticuline	1.00	27	Trans-p-meth-2-en-ol	1.00
8	1,8-cineole	1.00	18	N-methylcoclaurine	1.00	28	Ascaridole	1.00
9	Beta-pinene	1.00	19	Coclaurine	1.00	29	Beta-elemene	1.00
10	Phellandrene	1.00	20	Laurotetanine	1.00	30	Spathulenol	1.00

B.								
	Molécule	Score		Molécule	Score		Molécule	Score
1	Linalol	1.00	11	Beta-pinene	1.00	21	Alpha-terpinene	1.00
2	Limonene	1.00	12	Phellandrene	1.00	22	2-carene	1.00
3	Alpha-terpineol	1.00	13	3-carene	1.00	23	Alpha-thujene	1.00
4	Alpha-pinene	1.00	14	Thymol	1.00	24	Sabinene	1.00
5	Terpinolene	1.00	15	Isocorydine	1.00	25	Dehydro-1,8-cineole	1.00
6	Beta-ocimene	1.00	16	Norisocorydine	1.00	26	Sabinene hydrate	1.00
7	4-terpineol	1.00	17	Rogersine	1.00	27	Trans-p-meth-2-en-ol	1.00
8	p-cimene	1.00	18	Boldine	1.00	28	Ascaridole	1.00
9	Caryophyllene	1.00	19	Reticuline	1.00	29	Beta-elemene	1.00
10	1,8-cineole	1.00	20	N-methylcoclaurine	1.00	30	Spathulenol	1.00

Pour renforcer cette théorie, d'autres expériences ont été faites, sans passer par la fonction recherche du script, mais simplement en vérifiant si, sur des mélanges de concentration(s) moléculaire(s) connue(s), un filtre dans l'esprit de celui codé par Reims, était capable de distinguer correctement plusieurs groupes d'intensités. Les exemples de mélange choisis sont (**A**) une molécule pure, la 30-epi-cambogine, avec un signal de solvant résiduel, (**B**) un mélange de quercétine (65%) et d'harpagoside (35%) et (**C**) un mélange de boldine (75%) et d'harpagoside (25%).

Dans cet essai, les déplacements chimiques du mélange sont triés par intensité décroissante. Puis, la moyenne des intensités des 2 premiers points est calculée et est faite figurer sur le graphique (ligne rouge épaisse). Deux autres lignes rouges, correspondant à + 2 écart-types et - 2 écart-types sont également tracées. Si le point suivant rentre dans cet intervalle, alors la moyenne et l'écart-type sont recalculés en intégrant la nouvelle valeur du point ajouté. Et ainsi de suite jusqu'à ce qu'un point sorte de l'intervalle, et que l'on considère donc le premier groupe d'intensités ainsi formé. En théorie, chaque groupe devrait correspondre à une molécule différente, les différences d'intensités au sein de ces mélanges étant importantes. On se référera à ce filtre en parlant de type « moyenne \pm écart-type incrémental ». Le processus pour le mélange (**A**) est illustré à la **Figure 34**.

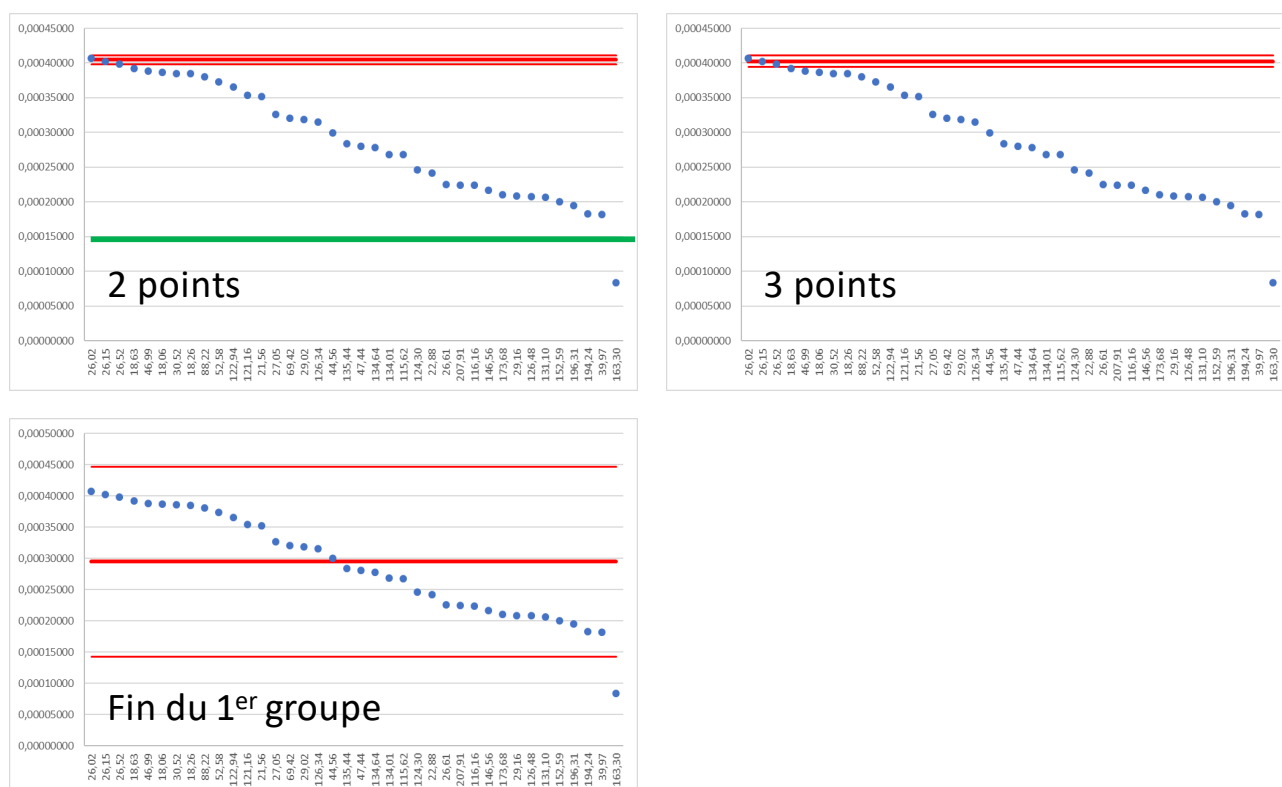


Figure 34: Test de filtre d'intensité de type « moyenne \pm écart-type incrémental » sur le mélange (A).

Sur le premier graphique, moyenne et écart-type ont été calculés avec les 2 premiers points. Le 3^{ème} point rentrant dans cet intervalle, il est intégré au nouveau calcul de la moyenne et écart-type représenté sur le second graphique. L'incrémentation est interrompue lorsqu'un point ne rentre plus dans l'intervalle, créant ainsi le premier groupe, comme représenté sur le dernier graphique. Ici, le premier groupe correspond bien à tous les déplacements chimiques de la molécule pure (30-epi-cambogine). Le signal restant non intégré est celui du solvant. La séparation des groupes est celle attendue (ligne verte). Le filtre type « moyenne \pm écart-type incrémental » fonctionne donc bien sur cet exemple.

Le même exercice sur le mélange (B) ne donne malheureusement pas d'aussi bons résultats (Figure 35). Un premier groupe est formé dès le 3^{ème} point intégré (15 points normalement attendus). Si l'on considère que ce n'est qu'un artefact et que l'on continue à intégrer les points suivants, un groupe se forme dès le 8^{ème} point intégré.

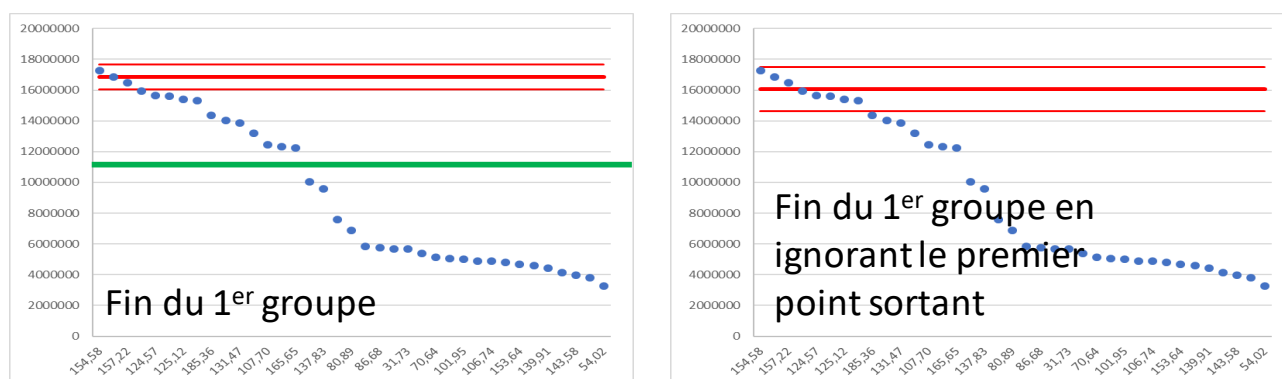


Figure 35: Test de filtre d'intensité de type « moyenne \pm écart-type incrémental » sur le mélange (B).

Les résultats obtenus ne sont pas meilleurs avec le dernier mélange (C) (Figure 36), dans lequel un groupe se forme dès les 2 premiers points (19 attendus). Le même type de calcul mais avec des paramètres différents a été essayé dans ces mêmes exemples, comme appliquer le filtre sur des valeurs triées dans un ordre croissant des intensités (Figure 36). Les autres essais (non présentés ici), y compris ceux faisant varier le nombre d'écart-types autorisé, n'ont pas donné les résultats escomptés : il n'a donc pas été possible d'obtenir une formule appropriée aux 3 exemples de mélange.

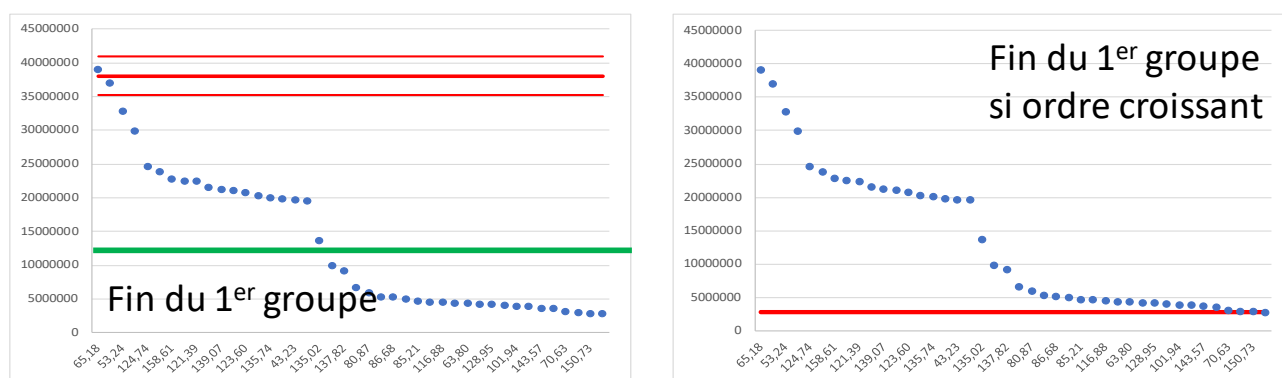


Figure 36: Test de filtre d'intensité de type « moyenne \pm écart-type incrémental » sur le mélange (C).

1.2. Méthodes de clustering

Le clustering présentait une alternative statistique intéressante à la création de groupes d'intensité. Nous nous sommes intéressés à l'opérer un clustering via Python, le langage informatique utilisé pour notre programme de déréplication.

La Figure 37 représente les différents scripts de clustering disponibles et la façon dont ils traitent les jeux de données. Chaque ligne correspond à un jeu fictif de données, avec une répartition des valeurs différente. Chaque colonne représente un script de clustering différent. Dans les cases, une couleur représente un cluster formé par le script correspondant, et en bas à droite se trouve le temps de calcul nécessaire pour distinguer ces clusters. Cela permet d'avoir une représentation visuelle de la façon dont travaille chacun des scripts.

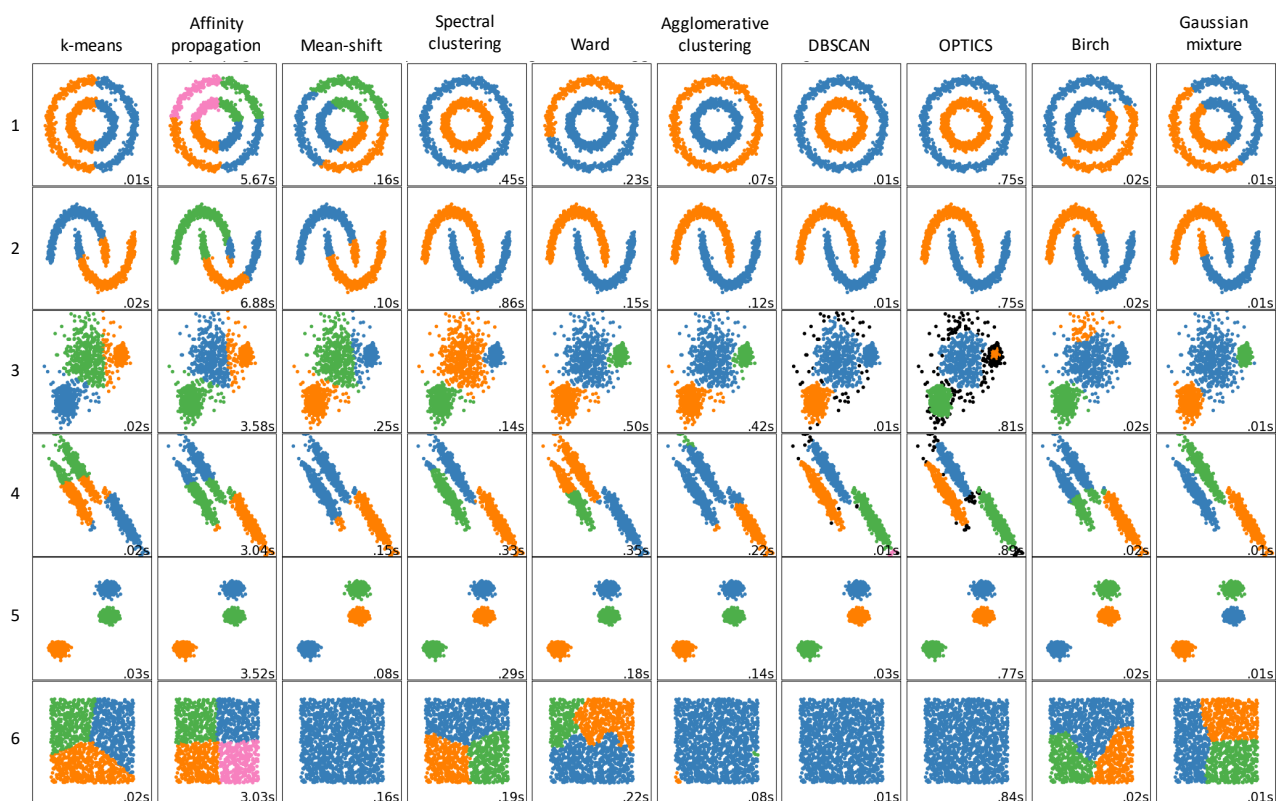


Figure 37: Représentation visuelle du résultat donné par chaque script de clustering sur différents types de jeux de données [37].

Le **Tableau 13** récapitulatif permet aussi d’apprécier les paramètres nécessaires à chaque script, leur usage recommandé, le type d’échantillon sur lequel ils fonctionnent, etc...

Tableau 13: Propriétés des différents scripts de clustering. Les couleurs représentent les critères d’exclusion : jaune = nombre de cluster prérequis, vert = taille de cluster homogène, orange = nombreux paramètres, gris = représentation graphique non appropriée [37].

Method name	Parameters	Scalability	Use case	Geometry (metric used)
k-means	Number of clusters	Very large n samples, medium n clusters	General-purpose, even cluster size, flat geometry, not too many clusters	Distance between points
Affinity propagation	Damping, sample preference	Not scalable with n samples	Many clusters, uneven cluster size, non-flat geometry	Graph distance (e.g. nearest-neighbor graph)
Mean-shift	Bandwidth	Not scalable with n samples	Many clusters, uneven cluster size, non-flat geometry	Distances between points
Spectral clustering	Number of clusters	Medium n samples, small n clusters	Few clusters, even cluster size, non-flat geometry	Graph distance (e.g. nearest-neighbor graph)
Ward hierarchical clustering	Number of clusters	Large n samples and n clusters	Many clusters, possibly connectivity constraints	Distances between points
Agglomerative clustering	Number of clusters, linkage type, distance	Large n samples and n clusters	Many clusters, possibly connectivity constraints, non-Euclidean distances	Any pairwise distance
DBSCAN	Neighborhood size	Very large n samples, medium n clusters	Non-flat geometry, uneven cluster sizes	Distances between nearest points
Gaussian mixtures	Many	Not scalable	Flat geometry, good for density estimation	Mahalanobis distances to centers
Birch	Branching factor, threshold, optional global clusterer	Large n clusters and n samples	Large dataset, outlier removal, data reduction	Euclidean distance between points

Afin de choisir le script adapté, nous avons procédé par élimination. On ne peut pas se servir d'un script nécessitant de préciser un nombre de cluster, puisque c'est ce qu'on cherche à déterminer. Ce type de script cherche à créer absolument un nombre de cluster défini, même si une séparation n'est visiblement pas nécessaire. C'est par exemple le cas de la méthode k-means qui sépare le jeu de données 6 en 3 clusters, alors qu'il s'agit visiblement d'un bloc. Dans l'application que l'on veut en faire, il est possible d'avoir seulement un ou plusieurs groupes d'intensités au sein d'un mélange, le nombre de cluster est donc variable. Les autres types de scripts à éliminer sont ceux qui cherchent regrouper les points en clusters de taille homogène. En effet, dans notre cas, il est possible que les groupes d'intensités regroupent des nombres de points identiques (2 molécules à 20 carbones en mélange par exemple), mais il est également possible que seulement quelques signaux « parasites » se retrouvent en mélange avec une seule grosse molécule. Les groupes peuvent donc être de taille très différente. Il est également difficile de gérer des scripts nécessitant de multiples paramètres, comme celui des *Gaussian mixtures*. La mise en place de multiples paramètres satisfaisant de manière optimale et universelle chacun des différents types de mélanges que l'on peut retrouver en déréplication semble illusoire. Enfin, l'aspect des différents résultats présentés en **Figure 37**, et le temps de calcul qui leur est associé, a représenté notre dernier critère de choix. Concernant le temps de calcul, les jeux de données qui seront concernés seront relativement petits (nombre de carbones dans la molécule, donc généralement moins de 60 pour les plus gros métabolites secondaires communs). En revanche, le nombre de molécules sera celui de la banque de données. Il pourra donc être très important en fonction de la banque utilisée (plusieurs dizaines de milliers de molécules pour les plus grandes bases de métabolites secondaires de type CH-NMR-NP [17]) Cela nécessite donc un script ayant un temps de calcul rapide. Concernant l'aspect général des résultats, les scripts présentant des clusters formés de façon trop « artificielle », c'est-à-dire des clusters ne semblant pas suivre les formes des ensembles de points, ont été évités. Les scripts donnant les résultats les plus intéressants visuellement sont DBSCAN et OPTICS. Ils présentent également un autre avantage : on repère, en plus des clusters formés en bleu, orange et vert, des points noirs qui représentent des points non clusterisés. Cette fonctionnalité de considérer certains signaux isolés comme « parasites », et non d'autres clusters, est très appropriée à ce que l'on souhaite obtenir. En effet, on peut imaginer un mélange de deux molécules avec 1 signal de solvant résiduel par exemple.

Après avoir pris en considération tous ces différents paramètres, le script DBSCAN a semblé le plus adapté à notre objectif.

1.3. Fonctionnement de l'algorithme DBSCAN

DBSCAN (*density-based spatial clustering of applications with noise*) est défini ainsi [37]:

« L'algorithme DBSCAN voit les clusters comme des aires de forte densité séparées par des aires de faible densité. A cause de cette vision assez générique, les clusters formés par DBSCAN peuvent prendre n'importe quelle forme, contrairement à la méthode des *k-means* qui suppose que chaque cluster a une forme convexe. La composante principale de DBSCAN est le concept de *core samples*, qui sont les échantillons de forte densité. Un cluster est donc un ensemble de *core samples*, chacun proche de l'autre (la proximité étant mesurée en distance) et un ensemble de *non-core samples* qui ne sont pas proches d'un *core sample* (mais qui ne sont eux même pas des *core samples*). Il y a deux paramètres dans l'algorithme, *min_samples* et *eps*, qui définissent ce que signifie

« dense ». Plus grand est *min_samples* ou plus petit est *eps*, plus haute est la densité nécessaire pour former un cluster.

De manière plus stricte, un *core sample* est défini comme étant un échantillon du jeu de données tel qu'il existe *min_sample* autres échantillons dans une distance d'*eps*, qui sont définis comme *voisins* du *core sample*. Cela nous dit que le *core sample* est une aire dense de l'espace vectoriel. Un cluster est un ensemble de *core samples* qui peuvent être construit en prenant de façon récurrente un *core sample*, trouvant tous ces voisins étant des *core samples*, trouvant tous *leurs* voisins étant des *core samples*, etc... Un cluster a aussi un ensemble de *non-core samples* qui sont des échantillons voisins d'un *core sample* dans le cluster, mais ne sont eux-mêmes pas *core samples*. De façon intuitive, ces échantillons sont en marge d'un cluster.

Par définition, chaque *core sample* fait partie d'un cluster. Chaque échantillon qui n'est pas un *core sample*, et situé à plus d'*eps* d'un *core sample*, est considéré comme étant une donnée aberrante.

Bien que le paramètre *min_samples* contrôle principalement la tolérance de l'algorithme vis-à-vis du bruit, le paramètre *eps* est crucial pour le jeu de données et la distance autorisé et ne peut généralement pas être laissé à la valeur par défaut. Ce paramètre contrôle le voisinage local des points. Si choisi trop petit, la plupart des données ne sera pas clusterisé (et étiquetée comme bruit). Si choisi trop grand, il entraîne la fusion de cluster trop proche les uns des autres en un super cluster et, au final, l'intégralité du jeu de données se considéré comme un seul cluster. Des approches heuristiques ont été discutées dans la littérature afin de choisir ce paramètre, basées par exemple sur l'observation graphique des distances au plus proche voisin. »

La **Figure 38** représente des résultats issus de DBSCAN. Les larges cercles représentent les *core samples*, les petit cercles les *non-core samples*, et les points noirs sont les données aberrantes.

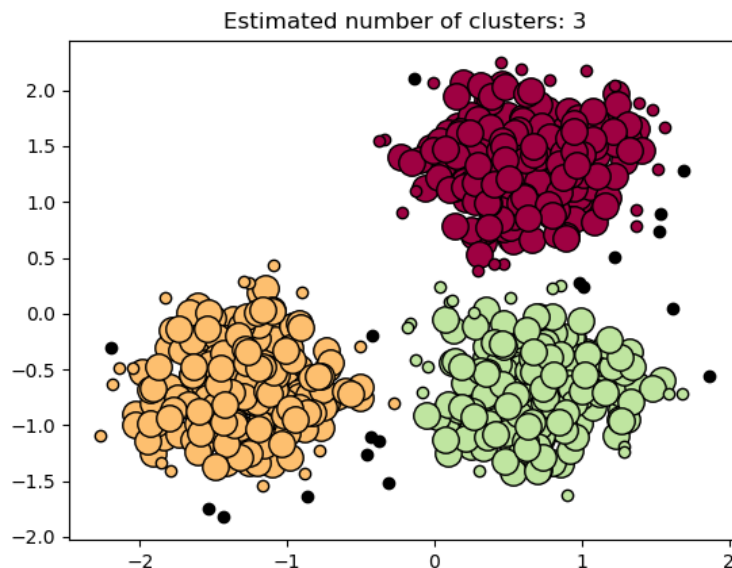


Figure 38: Exemple de clustering avec DB scan. Chaque cluster a une couleur différente. Les larges cercles sont les *core samples*, les petits cercles les *non-core samples* et les points noirs représentent les données aberrantes [37].

Le processus de clusterisation est également représenté en **Figure 39**. Les points rouges sont les *core-samples*. Depuis ces points, et avec un rayon d'*eps* ou ϵ , il est possible d'atteindre les points B et C en jaune. Ces points, placés en périphérie du cluster, seront des *non-core samples*. Enfin le point bleu N est une donnée aberrante, puisqu'il n'y a pas au moins *min_samples* situé dans un rayon d' ϵ .

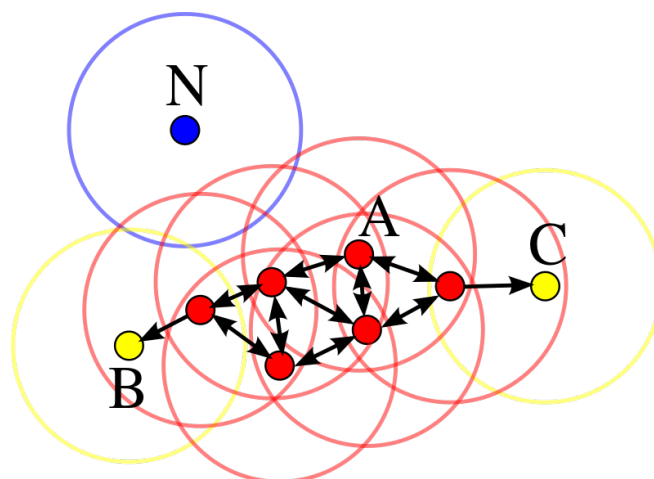


Figure 39: Processus de clustering avec DBSCAN. Les points rouges sont les *core samples*, les points jaunes les *non-core samples* et le point bleu est une donnée aberrante. Les rayons figurés sont d'une distance d' ϵ [38].

1.4. Essais de clustering avec DBSCAN

Le challenge principal était ici de définir le seul paramètre de l'algorithme, ϵ , de façon à ce qu'il satisfasse n'importe quel mélange. L'idée n'est évidemment pas de trouver une valeur brute d'épsilon, mais une relation entre ce dernier et une des caractéristiques « chiffrables » du mélange.

La première piste suivie est celle recommandée dans la description même du script, c'est-à-dire la méthode de l'épaulement (*knee* en anglais) observé sur le graphique des distances au voisin le plus proche [39]. Pour ce faire et pour chaque point, la distance au voisin le plus proche est calculée, puis les points sont triés par distance croissante. C'est à partir de la courbe en résultant que l'on peut déterminer la valeur idéale d' ϵ , celle-ci se trouvant dans l'épaulement de la courbe, dont il suffira de lire la valeur en ordonnée (Figure 40).

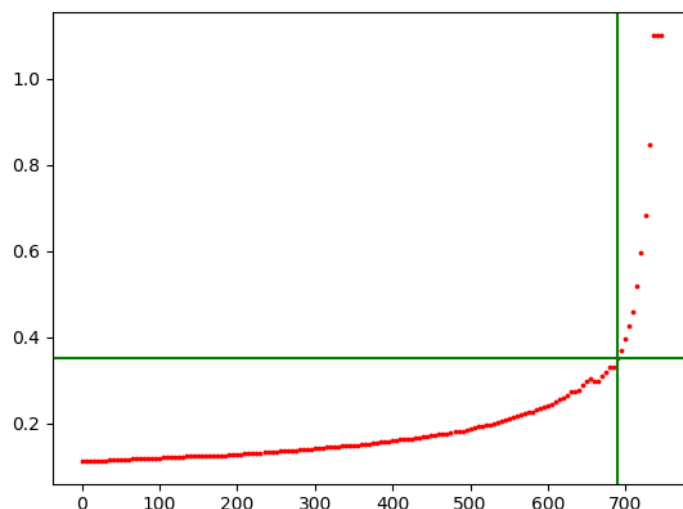


Figure 40: Exemple de la méthode de l'épaulement. L' ϵ approprié pour ce jeu de données sera d'environ 0,35, soit la valeur de l'ordonnée de l'épaulement.

Les exemples de mélanges choisis sont les mêmes que précédemment, soit **(A)** une molécule pure, la 30-epi-cambogine, mais comportant un signal résiduel de solvant, **(B)** un mélange de quercétine (65%) et d'harpagoside (35%) et **(C)** un mélange de boldine (75%) et d'harpagoside (25%).

La méthode de l'épaulement a été testée sur les échantillons en combinaison avec l'algorithme, c'est-à-dire qu'une fois la valeur d' ϵ déterminée grâce à la méthode de l'épaulement, elle était directement utilisée comme paramètre pour créer des clusters avec DBSCAN. Il convient de noter que les valeurs d'intensité de déplacement chimique entrées dans DBSCAN sont normalisées avant toute autre opération. Deux problèmes se sont présentés. Le premier est que certains échantillons ne permettaient pas d'observer d'épaulement : le graphique tenant plus de la droite que de la courbe... Il était donc impossible de déterminer un ϵ adéquat. Cela peut s'expliquer par le fait que le jeu de données soit relativement petit par rapport aux larges *datasets* qui peuvent être utilisés dans les exemples de cette méthode. L'autre explication est que la répartition des données dans ce jeu n'est pas si hétérogène que cela. Il y a donc peu de points situés à des distances extrêmes de leurs voisins. Ces points extrêmes sont cependant nécessaires pour pouvoir tracer la partie « droite » de la courbe, c'est-à-dire celle comportant les distances au voisin le plus proche les plus élevées. Le second problème est que, pour les exemples pour lesquels la méthode fonctionne, le ϵ trouvé ne forme pas les clusters attendus quand on l'utilise avec DBSCAN.

Il est important de préciser que l'on cherche non pas à obtenir une clusterisation parfaite, c'est-à-dire qu'un cluster corresponde aux intensités des signaux d'une seule et même molécule. On souhaite que chaque cluster formé contienne un pourcentage élevé de signaux appartenant à la même molécule. En effet, l'avantage de cette

méthode par rapport à celle de la moyenne \pm écart-type incrémenté est qu'elle ne s'arrête pas lorsqu'une valeur aberrante est rencontrée. L'ensemble du jeu de données est donc traité et, même si les clusters ne sont pas parfaits, toutes les valeurs y figureront.

Une autre approche a été tentée pour trouver un ϵ adéquat. Sur chaque exemple, la valeur du ϵ augmentée de façon incrémentale de 0,05. A chaque incrémentation, le nombre de clusters formé est noté. On considérera que, sur un intervalle d' ϵ donné tel que le nombre de fois où 1 seul cluster est formé ne dépasse pas 10 occurrences (signal d'un ϵ probablement trop large), le nombre de cluster qui ressort le plus souvent est le correct. Cette méthode ne fonctionne pas, le nombre de cluster le plus fréquent n'étant pas toujours le bon.

Nous avons alors tenté une approche différente en notant, pour chaque exemple, les valeurs limites de ϵ pour lesquelles le nombre de cluster formé est correct (**Tableau 14**), puis en essayant de relier cette valeur avec une autre propriété du mélange. Malheureusement aucun lien n'a pu être établi entre les valeurs correctes d'épsilon et les différents tests de paramètres des mélanges : nombres de points dans le mélange, distance cumulée entre les points, distance cumulée par nombre de point, distance maximale entre deux points, distance minimale entre deux points. Des exemples de clusters considérés comme corrects sont présentés en **Figure 41**.

Tableau 14: Pour les trois mélanges **A**, **B** et **C**, valeurs minimales et maximales de ϵ pour que les clusters formés soient corrects. La troisième colonne représente la marge entre la valeur minimale et maximale.

Mélange	ϵ min	ϵ max	Fenêtre ϵ
A	0,065	0,240	0,175
B	0,065	0,125	0,060
C	0,135	0,150	0,015

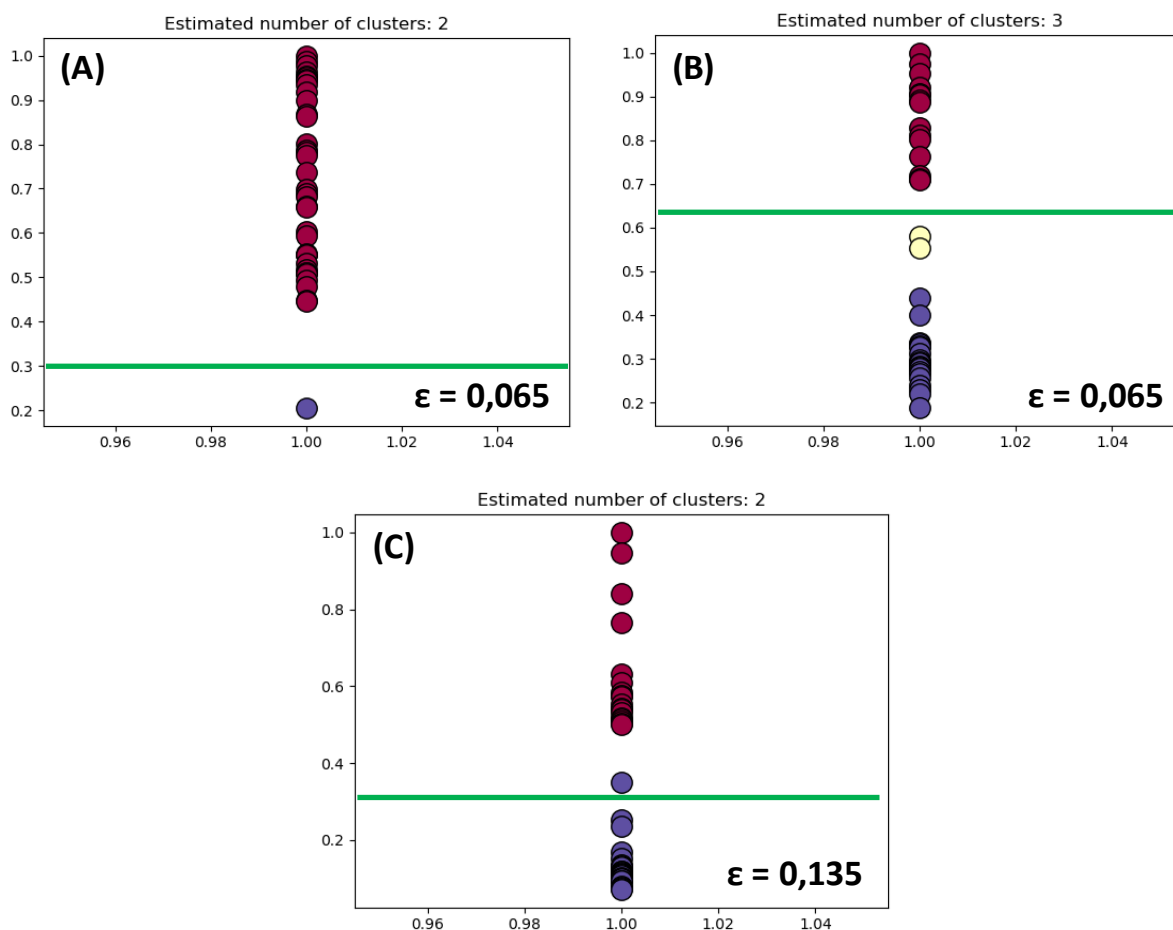


Figure 41: Exemple de clusters, considérés comme corrects, formés par DBSCAN sur les 3 mélanges **A**, **B** et **C**. La séparation réelle des valeurs est symbolisée par la ligne verte.

La dernière approche testée est la suivante : Sur des valeurs normalisées, si la différence entre deux valeurs consécutives est supérieure à une valeur δ , on considère alors qu'un nouveau cluster est formé. Sinon, ces points appartiennent au même cluster. 3 valeurs de δ ont été choisies : 0,20, 0,15 et 0,10. Pour chaque valeur, le nombre de clusters formé a été noté et comparé au nombre réel de clusters attendus (**Figure 42**).

Mélange	Clusters 0,20	Clusters 0,15	Clusters 0,10	Clusters réels
A	2	2	2	2
B	3	1	1	2
C	4	2	1	2

Figure 42: Nombre de clusters formés pour les 3 mélanges en fonction du δ autorisé.

En plus des 3 mélanges présentés ici, d'autres exemples ont été testés au travers de toutes les approches citées précédemment. Ces exemples comprennent des molécules pures, des mélanges de 2 molécules, des molécules pures et quelques signaux « parasites ». Ceux-ci ne sont pas présentés ici car les conclusions sont similaires à celles-ci-dessus décrites.

La conclusion générale de ces essais est que la complexité des différents types de mélanges que l'on peut retrouver lors d'analyses phytochimiques, et surtout la répartition des intensités de leur signaux, est trop élevée pour pouvoir trouver une méthode universelle automatisée qui permettrait de correctement différencier 2 molécules différentes en se basant sur l'intensité des signaux de RMN- ^{13}C . *In fine*, ce qui a pu ressortir de ces tentatives est que l'interprétation visuelle de ces jeux de données est plus facile pour un utilisateur humain. C'est pourquoi il a été décidé qu'aucun filtre d'intensité automatisé ne serait appliqué dans le processus de *matching*, mais que l'intensité des signaux associés seraient cependant graphiquement présentée à l'utilisateur qui pourrait utiliser ses compétences pour repérer des signaux d'intensité aberrante (**Figure 43**) et éventuellement les éliminer.

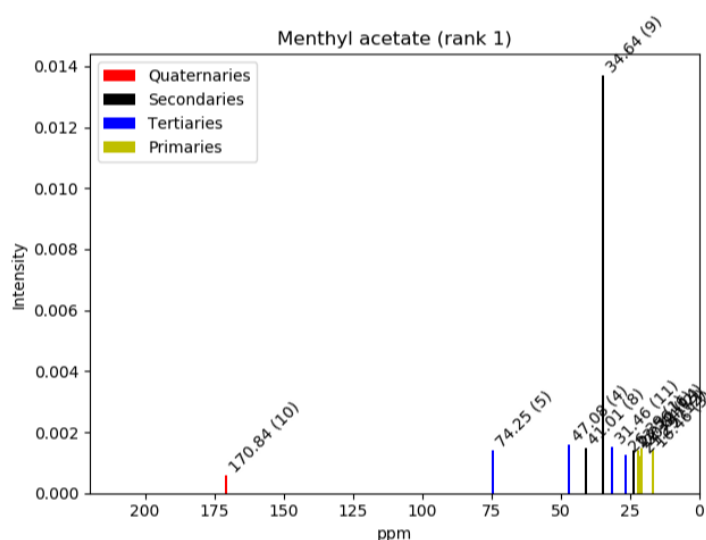


Figure 43: Exemple de l'interface graphique du logiciel MixONat qui permet de repérer rapidement des signaux d'intensité aberrante (ici, le signal à 34,64 ppm).

2. Valorisation biologique

L'objectif des tests biologiques détaillés ci-dessous est d'évaluer si l'effet de la guttiferone J sur les molécules du Complexe Majeur d'Histocompatibilité (CMH), observé par Caroline Rouger lors de sa thèse [40], est partagé par d'autres composés de la même classe structurale des acylphloroglucinols polycycliques polyprénylés (PPAPs). Si c'est le cas, les molécules actives permettront de pouvoir déterminer leur cible et préciser leur mécanisme d'action. Les études biologiques ont été réalisées avec l'aide de Chloé Coste, étudiante en master 2 (de 01/2017 à 06/2017).

2.1. Type cellulaire

Les tests ont été menés sur des cultures primaires de cellules endothéliales humaines. Les cellules endothéliales sont les cellules constituant la couche la plus interne des vaisseaux sanguins. Se situant à l'interface entre le sang et les tissus, elles participent à la réponse immunitaire (Figure 44 et Figure 45). Le choix de ce type cellulaire lors des tests s'explique par le fait que ce sont des cellules exprimant CMH de classe I et de classe II, ainsi que HLA-E et MHC class I related chain A (MICA), 4 molécules de surface impliquées dans la réponse adaptative et innée. Les cellules utilisées sont fortement activées par l'interféron gamma (IFN γ), une cytokine immunostimulatrice, ce qui permet de mieux visualiser quelque effet immunomodulateur que pourraient entraîner les PPAPs testés.

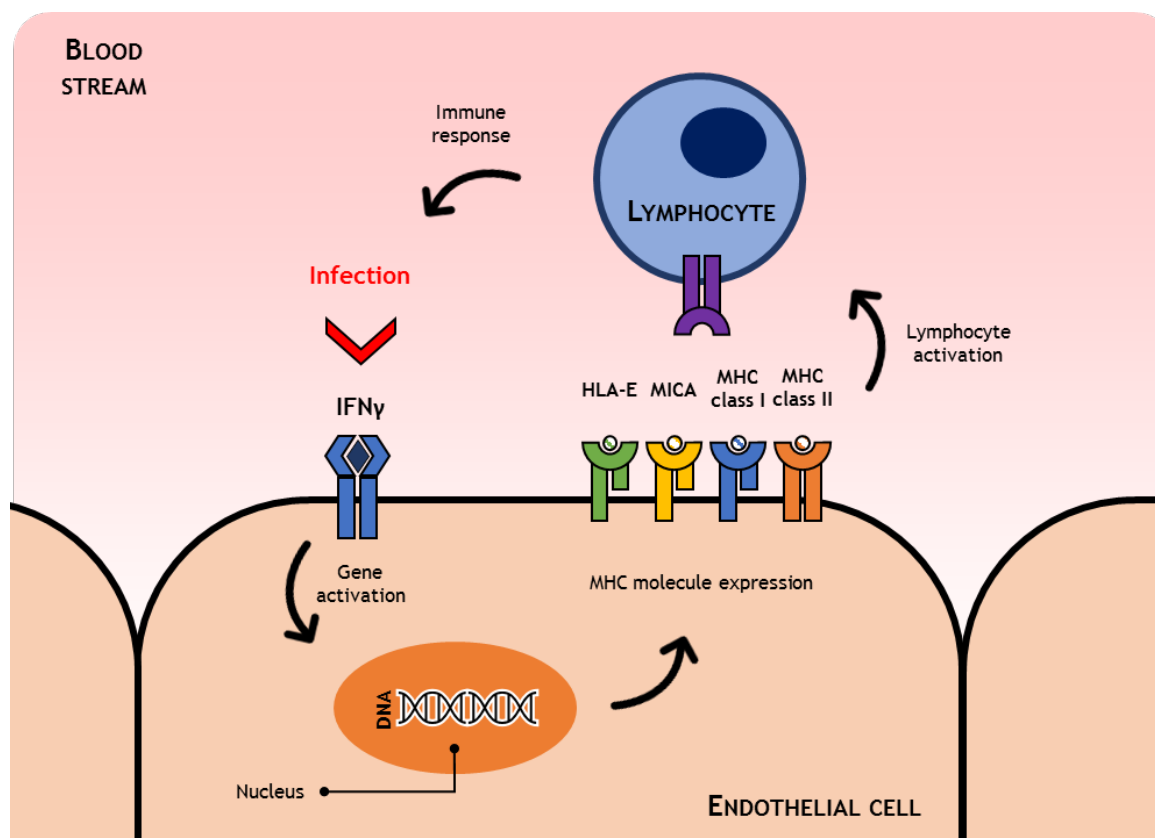


Figure 44: Cellule endothéliale et réponse immunitaire.

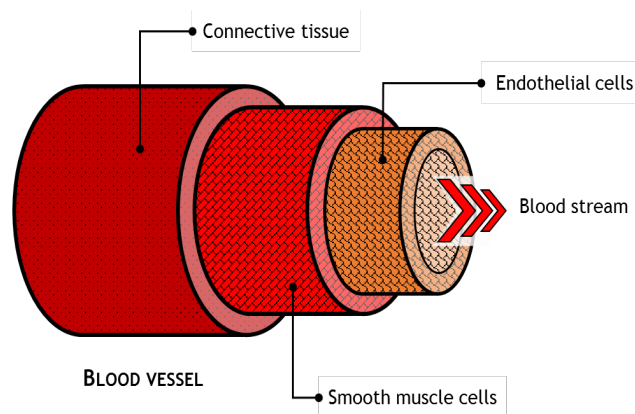


Figure 45: Cellules endothéliales au sein d'un vaisseau sanguin.

Les tests ont permis d'évaluer les effets de :

- la guttiferone J, disponible dans la chimiothèque du laboratoire.
- un mélange de guttiferone F et de xanthochymol, isolé de l'écorce de *Garcinia bancana*.
- un standard commercial de garcinol, puisque ce PPAP est décrit dans la littérature comme possédant des propriétés immunomodulatrices et des effets anti-cancéreux [41].

2.2. Expression des protéines de surface

Les cellules endothéliales sont cultivées jusqu'à confluence, c'est-à-dire jusqu'à ce qu'elles forment un tapis de cellules serrées, correspondant à leur état physiologique lorsqu'elles forment l'endothélium des vaisseaux sanguins. Les molécules à tester sont ensuite ajoutées et, après une heure, les cellules sont activées ou non par l'IFN γ . Différentes périodes d'incubation ont été testées (16, 24 et 48 heures) et des anticorps spécifiques, couplés à des fluorochromes, se lient à aux différentes protéines de surface (CMH de classe I, CMH de classe II, HLA-E et MICA). **Fluorescence Activated Cell Sorting (FACS)**, un type spécialisé de cytométrie en flux, a permis l'évaluation de l'expression de chaque marqueur à la surface des cellules (**Figure 46**).

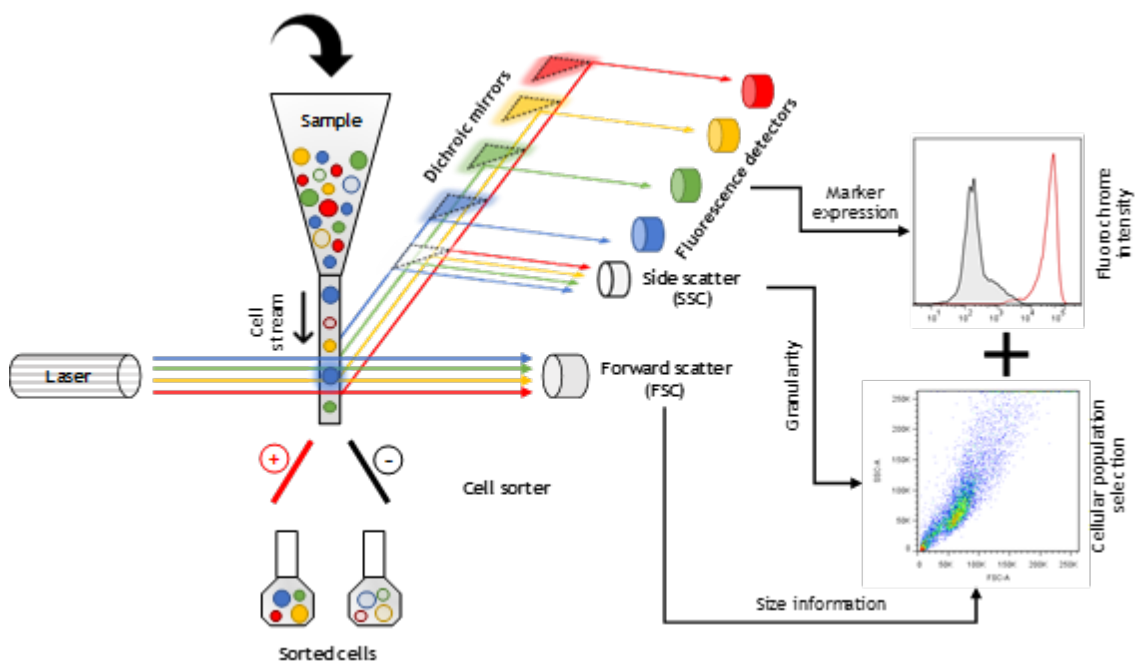


Figure 46: Principe de la cytométrie en flux.

Les effets des différentes PPAPs sur l'expression des différents marqueurs ont ensuite été comparés, à la fois au niveau basal (sans activation des cellules par l'IFN γ) et après activation par l'IFN γ . Il ressort de ces expériences que le garcinol et le mélange guttiferone F + xanthochymol diminue l'expression basale du CMH de classe I et de HLA-E, alors que la guttiferone J seulement diminue l'expression de HLA-E. Seul le garcinol semble réduire l'expression de MICA (**Figure 47**).

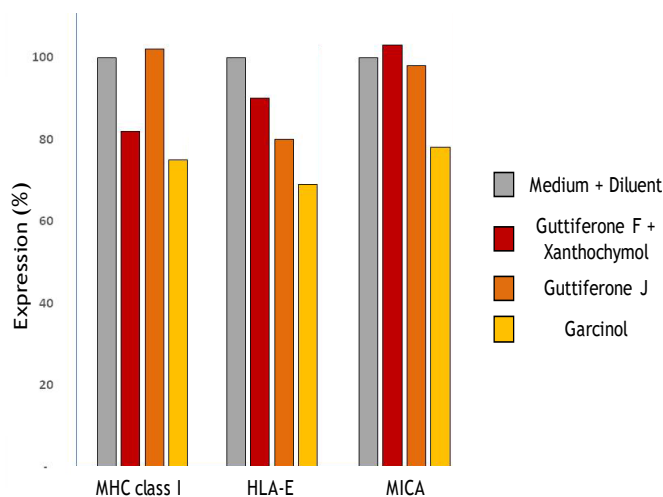


Figure 47: Effet des PPAPs sur l'expression basale des molécules du CMH.

Lorsqu'elles sont activées par l'IFN γ , les cellules endothéliales expriment fortement les molécules CMH de classe I et HLA-E, tout en diminuant légèrement l'expression de MICA. Après 48 heures, l'expression des molécules CMH de classe II est également fortement induite. Toutes les PPAPs ont montré une inhibition significative des effets de l'IFN γ , diminuant drastiquement l'expression des molécules CMH de classe II et HLA-E. Elles présentent

également un effet « léger » sur les molécules du CMH de classe I et MICA, réduisant leur expression (**Figure 48**).

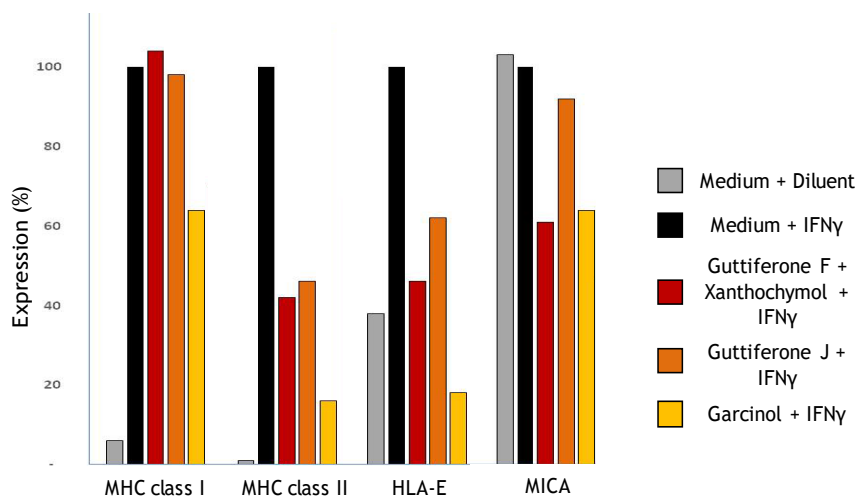


Figure 48: Effet des PPAPs sur l'expression des molécules du CMH sur des cellules activées par l'IFN γ .

Les PPAPs semblent partager des effets similaires sur l'expression des molécules du CMH à la surface des cellules, le garcinol produisant les effets les plus marqués.

2.3. Transcrits ARNm

Afin d'élucider le mécanisme d'action des PPAPs, il est important de savoir si les effets observés sur l'expression des molécules du CMH sont liés à une régulation des transcrits de leurs **acides ribonucléiques messagers (ARNm)**. Cette information permettra de savoir si la cible des PPAPs est située avant ou après l'étape de transcription de l'ARNm codant pour les marqueurs du CMH (**Figure 49**).

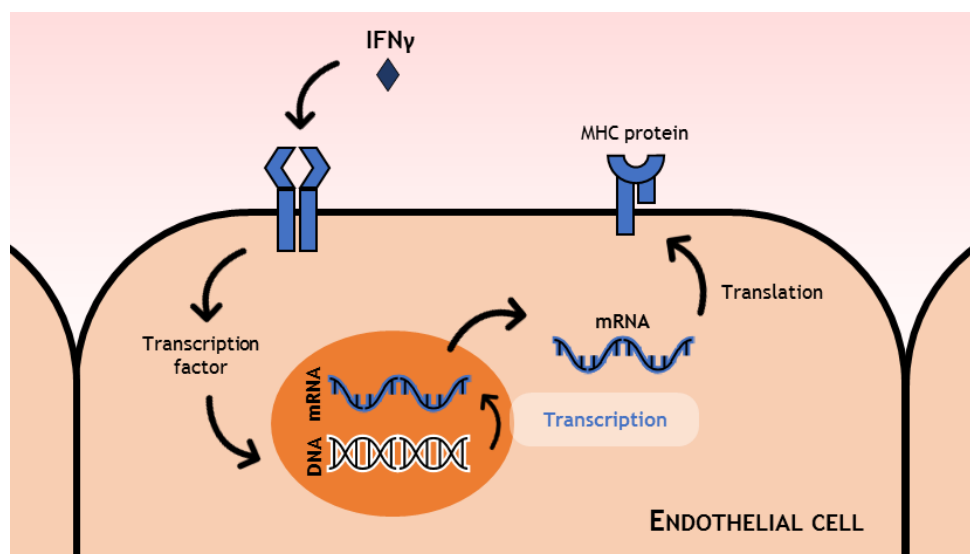


Figure 49: Processus d'activation génique menant à l'expression des molécules du CMH.

Les cellules endothéliales ont été cultivées dans les mêmes conditions que précédemment décrit; le protocole est également identique concernant l'étape d'ajout des produits à tester et de l'IFN γ . Après incubation, les cellules furent lysées et le matériel ARNm fut obtenu après des étapes de centrifugation et de précipitation; il fut amplifié grâce à la technique de **Reverse Transcription Polymerase Chain Reaction (RT-PCR)** (**Figure 50**).

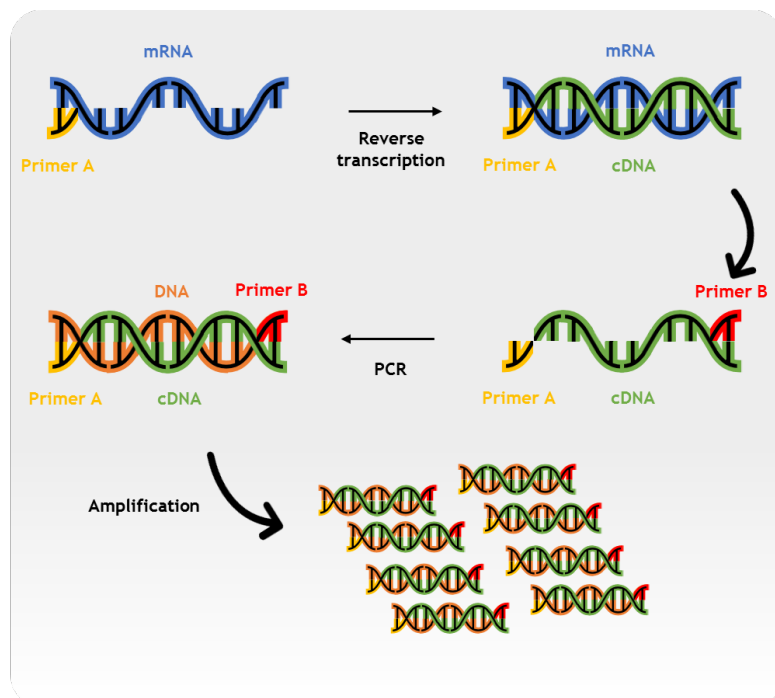


Figure 50: Principe de la RT-PCR.

Les résultats permettent de dégager la même tendance que celle observée sur l'expression des molécules de surface, mais avec un effet plus modéré (**Figure 51**). Cela peut être expliqué par un temps d'incubation plus court (16 heures au lieu de 48 heures). Cependant, cela permet quand même de suggérer que le mécanisme responsable de l'effet des PPAPs est un phénomène pré-transcriptionnel.

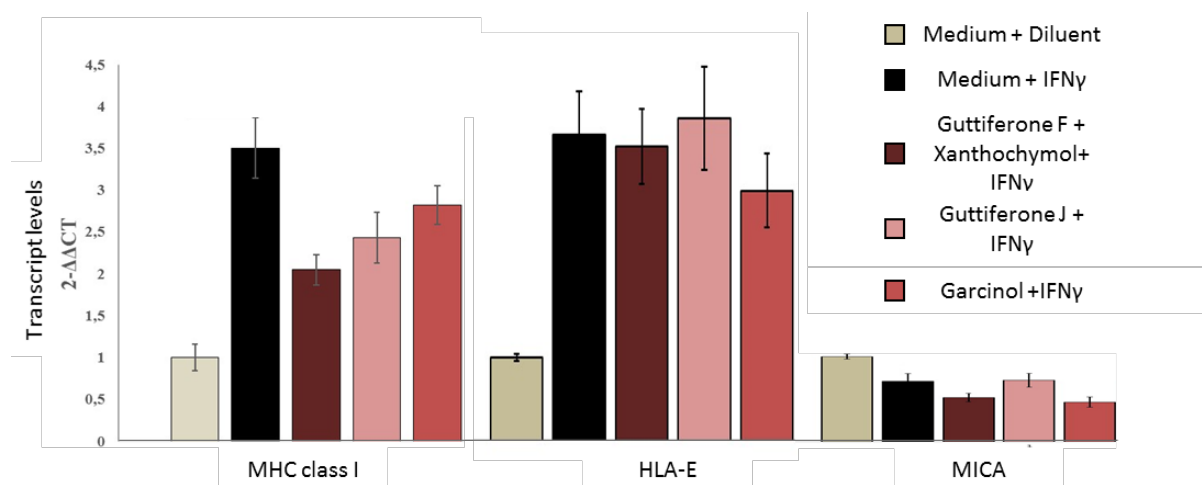


Figure 51: Niveau de transcrits ARNm après 16 heures de traitement.

2.4. Voie de l'IFN γ

L'effet des PPAPs sur l'expression des molécules du CMH étant plus intense après activation par l'IFN γ , il est important de décrire la voie de fonctionnement de ce dernier afin de pouvoir faire une hypothèse sur leur cible (**Figure 52**).

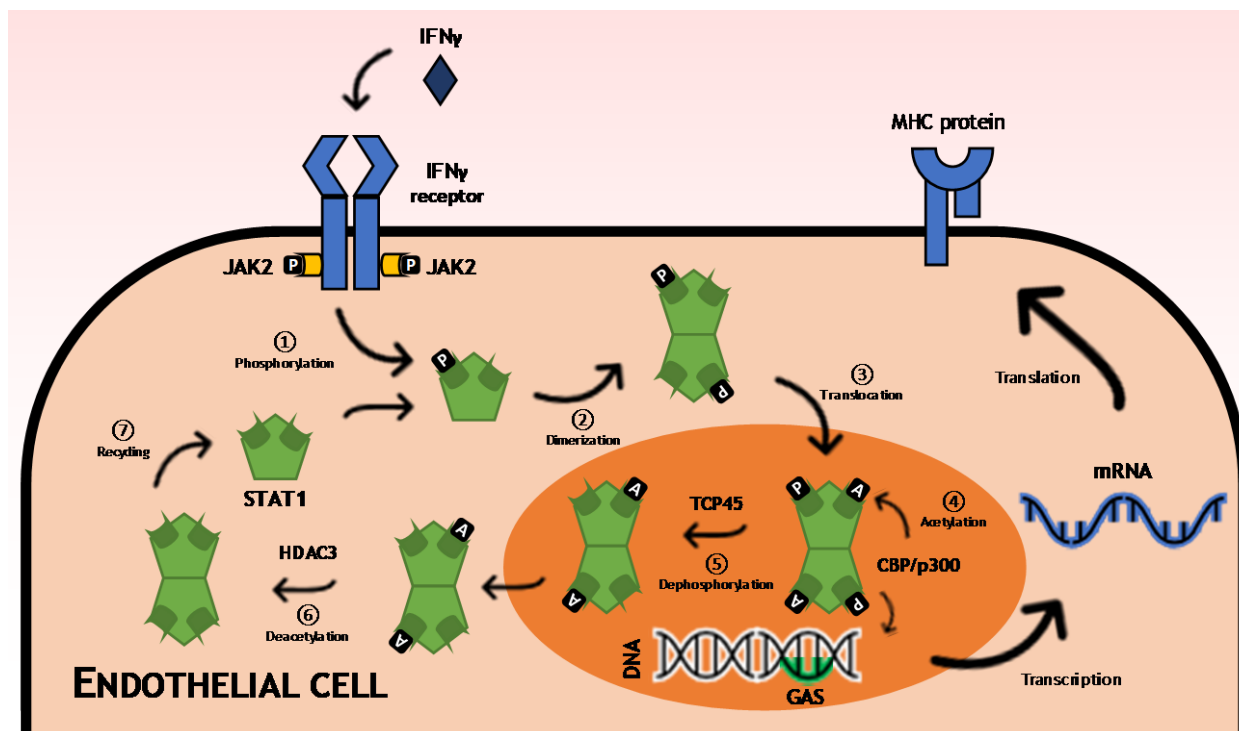


Figure 52: Voie de l'interféron gamma.

Lorsqu'il se lie à son récepteur, l'IFN γ induit l'autophosphorylation de *Janus Kinase 2* (**JAK2**). Cela permet à JAK2 d'activer *Signal Transducer and Activator of Transcription 1* (**STAT1**) en le phosphorylant (1). Les monomères activés de STAT1 forment un homodimère (2) qui se transloque dans le noyau (3) afin d'initier la transcription génique sur un site spécifique : *Gamma interferon Activated Sequence* (**GAS**). La fixation sur ce site est possible grâce à l'acétylation de STAT1 par une *Histone Acetyl Transferase* (**HAT**) appelée *CREB-Binding Protein* (**CBP**)/p300 (4). L'activation de ces gènes induit l'expression des molécules du CMH à la surface de la cellule. La déphosphorylation de STAT1, par la *T-Cell protein tyrosine Phosphatase 45* (**TCP45**), lui permettra de retourner dans le cytoplasme (5). STAT1 y sera désacétylé par l'*Histone Deacetylase 3* (**HDAC3**) (6) lui permettant ainsi de retourner à sa forme monomérique, configuration nécessaire pour être recyclée (7) [42-44].

2.5. Expression intracellulaire des protéines

Afin de valider les effets précédemment observés et pour statuer sur l'implication de STAT1, le niveau d'expression de ce facteur de transcription doit être mesuré. Le Western blot est une technique utilisée en biologie moléculaire pour évaluer le taux d'expression des protéines au sein des tissus ou cellules. Une électrophorèse sur gel permet la séparation des différentes protéines en fonction de leur taille et charge respective. Des anticorps spécifiques sont utilisés pour marquer les protéines désirées et l'échantillon est ensuite placé sur une membrane. En

combinant la spécificité des anticorps et la distance de migration, il est possible de détecter les protéines d'intérêt, tout en évaluant leur taille et leur taux d'expression (**Figure 53**).

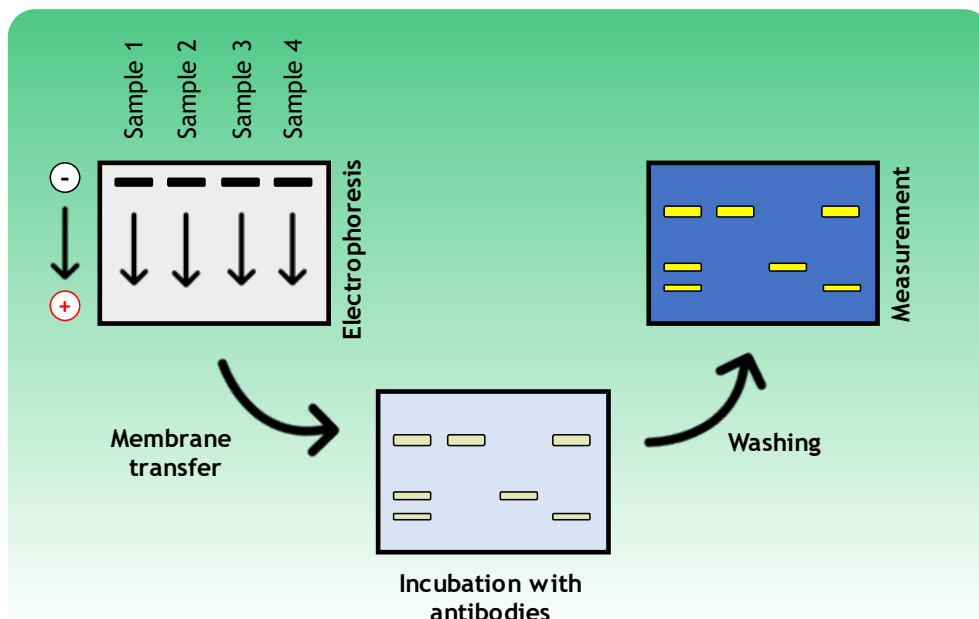


Figure 53: Principe du Western blot.

Les cellules endothéliales furent de nouveau cultivées et les PPAPs et/ou l'IFN γ furent ajoutés après différentes périodes d'incubation. Après l'électrophorèse, un transfert de membrane et l'utilisation d'anticorps spécifiques pour STAT1 total et *phosphorylated* **STAT1** (**pSTAT1**), la forme phosphorylée activée de STAT1, la membrane peut être lue (**Figure 54**).

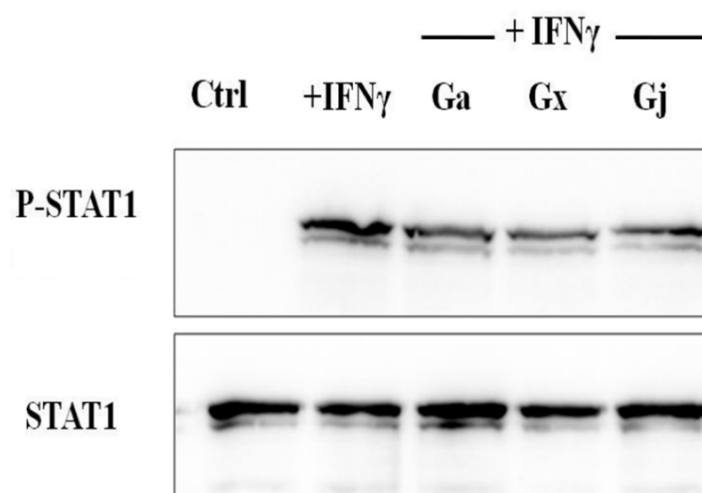


Figure 54: Expression induite de pSTAT1 par les PPAPs.

Après 6 heures de traitement par l'IFN γ , une augmentation générale de pSTAT1 a été observée. Les échantillons incluant PPAPs et l'IFN γ ont montré une augmentation plus faible de pSTAT1. Cela semble confirmer que les PPAPs régulent la phosphorylation de STAT1, et donc son activation.

2.6. Conclusion et perspectives

Ces différentes expériences ont permis de montrer que les PPAPs partagent le même effet général : une diminution de l'expression des molécules du CMH à la surface des cellules, de la transcription de leur ARNm, et une diminution de la phosphorylation de STAT1.

Le garcinol est décrit dans la littérature comme étant inhibiteur de STAT1 [41], mais aucun lien n'a encore été montré entre ses effets sur HLA-E et ceux sur STAT1. Ce lien semble être intéressant à investiguer lors de prochains travaux, dans lesquels les PPAPs pourraient servir d'outils.

Même si les PPAPs partagent des effets similaires, ces derniers peuvent être plus ou moins modérés en fonction des molécules testées. De plus, des effets toxiques peuvent apparaître avec certains composés. Les études de relations structure/activité pour la classe structurale des PPAPs, combinées à des expériences de *docking* sur STAT1 [41, 45, 46], pourront faire l'objet d'études futures.

Bibliographie

1. Gonzalez-Gallego, J., et al., *Fruit polyphenols, immunity and inflammation*. Br. J. Nutr., 2010. **104** **Suppl 3**: p. S15-27.
2. Cechinel Filho, V., C. Meyre-Silva, and R. Niero, *Chemical and pharmacological aspects of the genus Calophyllum*. Chem. Biodivers., 2009. **6**(3): p. 313-327.
3. Rouger, C., et al., *Prenylated polyphenols from Clusiaceae and Calophyllaceae with immunomodulatory activity on endothelial cells*. PLoS One, 2016. **11**(12): p. e0167361.
4. Yang, J.Y., et al., *Molecular networking as a dereplication strategy*. J. Nat. Prod., 2013. **76**(9): p. 1686-1699.
5. Yang, X.W., R.B. Grossman, and G. Xu, *Research progress of polycyclic polyprenylated acylphloroglucinols*. Chem. Rev., 2018. **118**(7): p. 3508-3558.
6. Bakiri, A., et al., *Computer-aided (13)C NMR chemical profiling of crude natural extracts without fractionation*. J. Nat. Prod., 2017. **80**(5): p. 1387-1396.
7. Bitzer, J., et al., *Accelerated dereplication of natural products, supported by reference libraries*. Chimia, 2007. **61**(6): p. 332-338.
8. Bighelli, A., F. Tomi, and J. Casanova, *Computer-aided carbon 13 NMR study of phenols contained in liquids produced by pyrolysis of biomass*. Biomass Bioenerg., 1994. **6**(6): p. 461-466.
9. Hubert, J., et al., *Identification of natural metabolites in mixture: a pattern recognition strategy based on (13)C NMR*. Anal. Chem., 2014. **86**(6): p. 2955-2962.
10. Roux, D., et al., *Structure-activity relationship of polyisoprenyl benzophenones from Garcinia pyrifera on the tubulin/microtubule system*. J. Nat. Prod., 2000. **63**(8): p. 1070-1076.
11. Marti, G., et al., *Antiplasmodial benzophenones from the trunk latex of Moronobea coccinea (Clusiaceae)*. Phytochemistry, 2009. **70**(1): p. 75-85.
12. ACD/NMR Predictors. Available from: https://www.acdlabs.com/products/adh/nmr/nmr_pred/.
13. Metrelab NMR Predict. Available from: <https://mestrelab.com/software/mnova/nmr-predict/>.
14. ChemDraw Professional. Available from: <https://www.perkinelmer.com/product/chemdraw-professional-chemdrawpro>.
15. Bremser, W., *Hose — a novel substructure code*. Anal. Chim. Acta, 1978. **103**(4): p. 355-365.
16. Bruguiera, A., et al., *(13)C-NMR dereplication of Garcinia extracts: Predicted chemical shifts as reliable databases*. Fitoterapia, 2018. **131**: p. 59-64.
17. CH-NMR-NP. Available from: <https://www.j-resonance.com/en/nmrdb/>.
18. Dictionary of Natural Products. Available from: dnpc.chemnetbase.com/.
19. SciFinder. [cited 2019; Available from: <https://scifinder.cas.org/>.]
20. Elfita, E., et al., *Antiplasmodial and other constituents from four Indonesian Garcinia spp.* Phytochemistry, 2009. **70**(7): p. 907-912.
21. Foundation, P.S. Python 2.7. Available from: <https://www.python.org>.
22. Conda. Available from: <https://docs.conda.io/projects/conda/en/latest/>.
23. RDKit. Available from: <https://www.rdkit.org/>.
24. Hunter, J.D., *Matplotlib: A 2D Graphics Environment*. Comput. Sci. Eng., 2007. **9**(3): p. 90-95.
25. Kivy. Available from: <https://kivy.org/#home>.
26. EFSA, *Use of rosemary extracts as food additive*. EFSA Journal, 2008. **721**: p. 1-29.
27. Bruguiera, A., et al., *MixONat, a software for mixtures dereplication based on 13C NMR experiments*. Anal. Chem., Submitted.
28. EMA. *Mentha piperitae aetheroleum*. [cited 2019 May]; Available from: <https://www.ema.europa.eu/en/medicines/herbal/menthae-piperitae-aetheroleum/>.
29. Agency, E.M. *Rosmarini folium*. [cited 2019 April]; Available from: <https://www.ema.europa.eu/en/medicines/herbal/rosmarini-folium>.
30. Bruguiera, A., et al., *Identifying Natural Products (NPs) as potential UPR inhibitors for crop protection*. Planta Med., 2016. **81**(S 01): p. S1-S381.
31. Li, G., et al., *Selective modulation of endoplasmic reticulum stress markers in prostate cancer cells by a standardized mangosteen fruit extract*. PLoS One, 2013. **8**(12): p. e81572.
32. Xu, X.H., et al., *Garcinone E induces apoptosis and inhibits migration and invasion in ovarian cancer cells*. Sci. Rep., 2017. **7**(1): p. 10718.
33. Yang, X.W., R.B. Grossman, and G. Xu. *Table of naturally occurring PPAPs*. [cited 2019; Available from: <http://www.uky.edu/~rbgros1/PPAPs/allPPAPs.html>].
34. Vogel, H., et al., *Studies of genetic variation of essential oil and alkaloid content in boldo (Peumus boldus)*. Planta Med., 1999. **65**(1): p. 90-91.

35. Urzua, A., et al., *Insecticidal properties of *Peumus boldus* Mol. essential oil on the house fly, *Musca domestica* L.* B. Latinoam. Caribe, 2010. **9**(6): p. 465-469.
36. Rakotonanahary-Maldonado, M., *Peumus boldus* M.: de la botanique à la thérapeutique - état des connaissances en 2012, in *Faculty of pharmacy*. 2013, Grenoble.
37. *Scikit learn: clustering*. [cited 2019; Available from: <https://scikit-learn.org/stable/modules/clustering.html>.]
38. *DBSCAN*. [cited 2019; Available from: <https://fr.wikipedia.org/wiki/DBSCAN>.]
39. Elbatta, M.T.H. and W.M. Ashour, *A dynamic method for discovering density varied clusters*. International Journal of Signal Processing, Image Processing and Pattern Recognition, 2013. **6**(1): p. 123-134.
40. Rouger, C., *Activité pharmacologique de dérivés polyphénoliques isolés de Clusiaceae et de Calophyllaceae malaisiennes: effets régulateurs sur des marqueurs endothéliaux de l'inflammation et de l'immunité*, in *Faculty of pharmacy*. 2015, Angers.
41. Masullo, M., et al., *Direct interaction of garcinol and related polyisoprenylated benzophenones of *Garcinia cambogia* fruits with the transcription factor STAT-1 as a likely mechanism of their inhibitory effect on cytokine signaling pathways*. J. Nat. Prod., 2014. **77**(3): p. 543-549.
42. Zhuang, S., *Regulation of STAT signaling by acetylation*. Cell Signal, 2013. **25**(9): p. 1924-1931.
43. Antunes, F., A. Marg, and U. Vinkemeier, *STAT1 signaling is not regulated by a phosphorylation-acetylation switch*. Mol. Cell. Biol., 2011. **31**(14): p. 3029-3037.
44. Kramer, O.H. and T. Heinzel, *Phosphorylation-acetylation switch in the regulation of STAT1 signaling*. Mol. Cell. Endocrinol., 2010. **315**(1-2): p. 40-48.
45. Zhou, X.Y., et al., *The C8 side chain is one of the key functional group of Garcinol for its anti-cancer effects*. Bioorg. Chem., 2017. **71**: p. 74-80.
46. Han, C.M., et al., *13,14-Dihydroxy groups are critical for the anti-cancer effects of garcinol*. Bioorg. Chem., 2015. **60**: p. 123-129.

Table des matières

PRODUCTIONS SCIENTIFIQUES

1. Publications
2. Communications orales
3. Posters

INTRODUCTION 1

I. ÉTAT DE L'ART 2

1. Article 1 : A highlight on ¹³C-NMR based dereplication methods 2

II. TRAVAUX PERSONNELS 44

1. Construction des bases de données 44

1.1. L'étape clé des références 44

1.2. Article 2: ¹³C-NMR dereplication of Garcinia extracts: Predicted chemical shifts as reliable databases 47

1.2.1. Résumé de l'article 2 47

1.2.2. Article 2 48

2. Algorithme de déréplication 60

2.1. Éléments de fonctionnement des algorithmes de déréplication par RMN-¹³C existants 60

2.2. Essais de déréplication 63

2.2.1. Partie expérimentale 63

2.2.2. Résultats 64

2.2.3. Discussion 67

2.3. Fonctionnement de l'algorithme DerepCrude 69

2.4. Recherche d'améliorations algorithmiques 73

2.4.1. Exploitation des données issues des expériences DEPT 73

2.4.2. Matching des déplacements chimiques des ¹³C 76

2.4.3. Utilisation de filtres 76

2.4.4. Interactivité 77

2.4.5. Intensité des signaux 77

2.5. Description de l'algorithme MixONat 77

2.5.1. Génération des bases de données : onglet CTypeGen 78

2.5.2. Matching : onglets Inputs et Parameters 83

2.6. Fonctionnement du matching de l'algorithme MixONat 86

2.7. Présentation des résultats de l'algorithme MixONat 88

3. Validation de la méthode 93

3.1. Huile essentielle de menthe poivrée 93

3.1.1. Partie expérimentale 93

3.1.2. Résultats et discussion 94

a) Composition de l'huile essentielle de menthe poivrée 94

b) Résultats de déréplication proposés par MixONat 95

3.2. Article 3 : MixONat, a software for mixtures dereplication based on ¹³C-NMR experiments. 98

3.2.1. Résumé de l'article 3 98

3.2.2. Article 3 101

3.2.3. Éléments de discussion 164

3.2.4. Perspectives d'amélioration de MixONat 164

4. Application de la méthode 166

4.1. Article 4 : Polyphenylated polycyclic acylphloroglucinols identification from *Garcinia bancana* bark using ¹³C-NMR dereplication program MixONat 166

4.1.1. Résumé de l'article 4 166

4.1.2. Article 4 168

4.1.3. Éléments de discussion 207

CONCLUSION GÉNÉRALE ET PERSPECTIVES 208

ANNEXES 210

1. Filtre d'intensité des algorithmes de recherche 210

1.1. Le filtre d'intensité de DerepCrude 210

1.2.	Méthodes de clustering.....	214
1.3.	Fonctionnement de l'algorithme DBSCAN	216
1.4.	Essais de clustering avec DBSCAN	218
2.	Valorisation biologique.....	223
2.1.	Type cellulaire	223
2.2.	Expression des protéines de surface	224
2.3.	Transcrits ARNm.....	226
2.4.	Voie de l'IFN γ	228
2.5.	Expression intracellulaire des protéines.....	228
2.6.	Conclusion et perspectives.....	230
BIBLIOGRAPHIE		231
TABLE DES FIGURES		235
TABLE DES TABLEAUX.....		237

Table des figures

Figure 1: (A) Principe général du code HOSE et (B) extrait du codex HOSE.....	45
Figure 2: HOSE code détaillé du C-1 de la molécule A	46
Figure 3: Fenêtre de recherche dans ACD Labs®.....	61
Figure 4: Fenêtre de recherche sur CH-NMR-NP.....	62
Figure 5: (A) Profil de l'extrait d' <i>Allanblackia floribunda</i> à 254 nm. (B) Type de structure identifiée dans la fraction : fukugiside.	63
Figure 6: Résultats proposés par l'algorithme ACD/Labs® avec la base de données <i>Allanblackia</i> concernant la composition de la fraction enrichie d' <i>Allanblackia floribunda</i>	64
Figure 7: Résultats proposés par l'algorithme ACD/Labs® avec la base de données <i>Garcinia</i> concernant la composition de la fraction enrichie d' <i>Allanblackia floribunda</i>	65
Figure 8: Résultats proposés par l'algorithme DerepCrude avec la base de données <i>Allanblackia</i> concernant la composition de la fraction enrichie d' <i>Allanblackia floribunda</i>	66
Figure 9: Résultats proposés par l'algorithme DerepCrude avec la base de données <i>Garcinia</i> concernant la composition de la fraction enrichie d' <i>Allanblackia floribunda</i>	67
Figure 10: Lancement de l'algorithme DerepCrude.	69
Figure 11: Fonctionnement de l'algorithme DerepCrude.	71
Figure 12: Fichier texte (.txt) de résultats générés par l'algorithme DerepCrude.	72
Figure 13: Fichier image (.png) de résultats de l'algorithme DerepCrude.	72
Figure 14: Information sur la multiplicité des atomes de carbone déduite des expériences de RMN- ¹³ C et DEPT 135 & 90.....	73
Figure 15: Onglet CTypeGen permettant de trier les déplacements chimiques de chaque molécule de la base de données SDF en fonction du type de carbone.....	78
Figure 16: Anatomie d'un SDF : le Mol Block du benzaldéhyde.....	80
Figure 17: Anatomie d'un SDF : les étiquettes du benzaldéhyde.	81
Figure 18: Extrait des étiquettes du benzaldéhyde après transformation par CTypeGen.....	82
Figure 19: Tri des bases de données par type de carbone par le premier script CTypeGen.	82
Figure 20: Onglet <i>Inputs</i> permettant de charger via un explorateur les différents fichiers d'entrée dans le programme. Base de données (SDF) et spectre ¹³ C (.csv) sont obligatoires, DEPT 135 et DEPT 90 (.csv) optionnels.	83
Figure 21: Exemple d'un fichier .csv (<i>comma-separated values</i>) d'un spectre RMN- ¹³ C généré avec Microsoft Excel. Le déplacement chimique et son intensité sont séparés par une virgule.	84
Figure 22: Onglet <i>Parameters</i> permettant de modifier les paramètres du programme. Des valeurs par défauts sont proposées à l'utilisateur qui peut choisir de les modifier en fonction des informations qu'il possède.....	85
Figure 23: Fonctionnement de l'algorithme MixONat : données d'entrée et tri par type de carbone.....	86
Figure 24: Fonctionnement de l'algorithme MixONat : processus de <i>matching</i> et paramètres.	87
Figure 25: Exemple de correction locale des associations : l'association proposée par l'algorithme en sortie de <i>matching</i> à gauche et la solution optimale corrigée par le script à droite.	88
Figure 26: Fenêtre de présentation générale des résultats interactifs dans MixONat.....	89

Figure 27: Fenêtre de présentation spécifique d'un résultat dans MixONat.	90
Figure 28: Fonctionnement de l'algorithme MixONat : données de sortie.....	90
Figure 29: Extrait du fichier texte de résultats présentant les paramètres ainsi que le premier résultat d'une analyse.....	91
Figure 30: Extrait du fichier image de résultats présentant les 5 premiers résultats de la même analyse.....	92
Figure 31: Chromatogramme GC-FID de l'huile essentielle de <i>Mentha piperita</i> et structures identifiées.	94
Figure 32: 15 premiers résultats de l'analyse déréplicative de l'huile essentielle de <i>Mentha piperita</i>	96
Figure 33: Résultats 1-10 pour la déréplication de la fraction enrichie de <i>Garcinia mangostana</i>	100
Figure 34: Test de filtre d'intensité de type « moyenne \pm écart-type incrémental » sur le mélange (A).	213
Figure 35: Test de filtre d'intensité de type « moyenne \pm écart-type incrémental » sur le mélange (B).	214
Figure 36: Test de filtre d'intensité de type « moyenne \pm écart-type incrémental » sur le mélange (C).....	214
Figure 37: Représentation visuelle du résultat donné par chaque script de clustering sur différents types de jeux de données [37].	215
Figure 38: Exemple de clustering avec DB scan. Chaque cluster a une couleur différente. Les larges cercles sont les <i>core samples</i> , les petits cercles les <i>non-core samples</i> et les points noirs représentent les données aberrantes [37].	218
Figure 39: Processus de clustering avec DBSCAN. Les points rouges sont les <i>core samples</i> , les points jaunes les <i>non-core samples</i> et le point bleu est une donnée aberrante. Les rayons figurés sont d'une distance d' ϵ [38].	218
Figure 40: Exemple de la méthode de l'épaulement. L' ϵ approprié pour ce jeu de données sera d'environ 0,35, soit la valeur de l'ordonnée de l'épaulement.	219
Figure 41: Exemple de clusters, considérés comme corrects, formés par DBSCAN sur les 3 mélanges A, B et C. La séparation réelle des valeurs est symbolisée par la ligne verte.	221
Figure 42: Nombre de clusters formés pour les 3 mélanges en fonction du δ autorisé.	221
Figure 43: Exemple de l'interface graphique du logiciel MixONat qui permet de repérer rapidement des signaux d'intensité aberrante (ici, le signal a 34,64 ppm).	222
Figure 44: Cellule endothéliale et réponse immunitaire.	223
Figure 45: Cellules endothéliales au sein d'un vaisseau sanguin.	224
Figure 46: Principe de la cytométrie en flux.	225
Figure 47: Effet des PPAPs sur l'expression basale des molécules du CMH.	225
Figure 48: Effet des PPAPs sur l'expression des molécules du CMH sur des cellules activées par l'IFN γ	226
Figure 49: Processus d'activation génique menant à l'expression des molécules du CMH.	226
Figure 50: Principe de la RT-PCR.	227
Figure 51: Niveau de transcrits ARNm après 16 heures de traitement.	227
Figure 52: Voie de l'interféron gamma.....	228
Figure 53: Principe du Western blot.	229
Figure 54: Expression induite de pSTAT1 par les PPAPs.....	229

Table des tableaux

Tableau 1: Déplacements chimiques carbone 13 en ppm correspondant au fukugiside et aux signaux restants.	68
Tableau 2: Nombre de produits naturels de la base de données <i>Garcinia</i> partageant la même combinaison de [C, CH ₂ , CH ₃ +CH].	74
Tableau 3: Nombre de fois (colonne de gauche) où une combinaison de [C, CH ₂ , CH ₃ +CH] amène à un groupe dont la taille est indiquée de la colonne de droite.	75
Tableau 4: Nombre de produits naturels de la base de données <i>Garcinia</i> partageant la même combinaison de [Cq, CH, CH ₂ , CH ₃].	75
Tableau 5: Nombre de fois (colonne de gauche) où une combinaison de [Cq, CH, CH ₂ , CH ₃] amène à un groupe dont la taille est indiquée de la colonne de droite.	76
Tableau 6: Pourcentage relatif en GC-FID des différentes molécules contenues dans l'huile essentielle de <i>Mentha piperita</i>	94
Tableau 7: Comparaison des déplacements chimiques RMN ¹³ C entre l'huile essentielle de menthe et les données de la littérature [29] pour le menthol, la menthone, le cinéole, l'acétate de menthyle, l'isomenthone et le limonène.	95
Tableau 8: Nom, pourcentage relatif dans l'extrait, rang et score des molécules correctes.	96
Tableau 9: Effet de l'incrémentation de la marge autorisée lors de la recherche.	97
Tableau 10: Récapitulatif des molécules isolées dans les différentes fractions. Les molécules dont les noms sont soulignés ont pu être directement identifiées depuis l'extrait brut.	167
Tableau 11: Comparaison des signaux du limonène et du terpinolène avec ceux de l'extrait de boldo.	211
Tableau 12: 30 premiers résultats donnés par l'algorithme DerepCrude lors de la déréplication d'un extrait brut de <i>Peumus bolus</i> (A) avec filtre d'intensité ou (B) sans. En jaune sont surlignées les différences majeures. .	212
Tableau 13: Propriétés des différents scripts de clustering. Les couleurs représentent les critères d'exclusion : jaune = nombre de cluster prérequis, vert = taille de cluster homogène, orange = nombreux paramètres, gris = représentation graphique non appropriée [37].	215
Tableau 14: Pour les trois mélanges A, B et C, valeurs minimales et maximales de ε pour que les clusters formés soient corrects. La troisième colonne représente la marge entre la valeur minimale et maximale.	220

Titre : Mise au point d'une méthode d'analyse déréplicative par RMN du carbone 13

Mots clés : déréplication, RMN-¹³C, programme, produits naturels

Résumé : L'extraction et la purification de produits naturels peut représenter un travail long et fastidieux, n'aboutissant pas forcément à l'isolement de molécules valorisables. C'est pourquoi des méthodes de déréplication ont progressivement été développées : elles permettent en effet d'identifier les molécules d'un mélange sans avoir à les séparer, en comparant leurs signaux à ceux de références regroupées en bases de données. Dans ce travail, nous avons cherché à utiliser une méthode de déréplication pouvant nous permettre de rapidement identifier des acylphloroglucinols polycycliques polyprénylés (PPAPs), molécules pouvant présenter un intérêt comme « outil thérapeutique » permettant de mieux comprendre certains mécanismes immunitaires et inflammatoires. En RMN du ¹³C, nous avons d'abord pu établir que la construction de bases de données contenant des valeurs prédites, au lieu de données expérimentales, permettait tout de même de réaliser des analyses déréplicatives de qualité. Les bases prédites ont donc été utilisées pour la suite de nos expériences.

Après avoir passé en revue les différentes méthodes de déréplication déjà décrites dans la littérature, nous avons décidé de développer notre propre programme de déréplication utilisant la RMN du ¹³C, dans l'optique de pouvoir le rendre plus discriminant que les méthodes préalablement proposées. Pour ce faire, en plus du spectre de RMN-{H}-¹³C, des informations déduites d'expériences DEPT (135 et 90) sont ajoutées, afin d'affiner les recherches par type de carbone. Une interface graphique a également été implémentée, permettant non seulement une facilité d'utilisation mais, également, une interactivité dans la validation des résultats, grâce à laquelle l'utilisateur peut optimiser les hypothèses faites par le programme. Cette nouvelle méthode a d'abord été testée avec succès sur différents mélanges de produits naturels, ce qui a permis d'en valider le concept. Enfin, la méthode a pu être appliquée sur des extraits de composition inconnue de *Garcinia bancana*, conduisant à l'identification des PPAPs recherchés. Les molécules d'intérêt ont finalement été purifiées pour de futurs tests biologiques.

Title: Developing a dereplication analysis method based on carbon 13 NMR

Keywords: dereplication, ¹³C-NMR, program, natural products

Abstract: Extraction and isolation of natural products (NPs) may be associated with a tedious and time-consuming work and can unfortunately lead to molecules of little to no interest. That is why dereplication methods have been developed: they allow the identification of molecules within a mixture, without having to separate them, by comparing their signals to those of references, gathered in databases. As far as NPs are concerned, our need for a proper dereplication strategy was focused on polycyclic polyprénylated acylphloroglucinols (PPAPs), molecules that could be used as a "therapeutic tool" in order to better understand mechanisms involved in immune and inflammatory responses. Using ¹³C-NMR, we first were able to conclude that building databases with predicted values, instead of experimental ones, gave quality results to a dereplication work.

Predicted databases were thus used for the remaining experiments. After a literature survey of the different kind of ¹³C-NMR based dereplication methods, we decided to develop our own program in order to make it even more discriminating. To do so, in addition to ¹³C data, DEPT (135 and 90) information were added, allowing to narrow the search by carbon type. A graphic user interface was also implemented, making the program easier to use, but also providing the user with the possibility to interact with the result validation. This new method was first successfully tested on diverse natural products mixtures, allowing the validation of the concept. Eventually, the method was applied to *Garcinia bancana* extracts of unknown composition. It made possible the quick identification of the PPAPs we were interested in, which were purified for further biological testing.