Improving the chemical profiling of complex natural extracts by joint 13C NMR and LC-HRMS2 analysis and the querying of in silico generated chemical databases

Julien Cordonnier, Simon Remy, Alexis Kotland, Ritchy Leroy, Pierre Darme, Benjamin Bertaux, Charlotte Sayagh, Agathe Martinez, Nicolas Borie, Jane Hubert, et al.

ChemRxiv®

Improving the chemical profiling of complex natural extracts by joint ^{13}C NMR and LC-HRMS^2 analysis and the querying of *in silico* generated chemical databases.

Julien Cordonnier,[a,b] Simon Remy,[b*] Alexis Kotland,[d] Ritchy Leroy,[b] Pierre Darme,[a,b], Benjamin Bertaux,[b] Charlotte Sayagh,[b] Agathe Martinez,[b] Nicolas Borie,[b] Jane Hubert,[d] Dominique Aubert,[a,c] Isabelle Villena,[a,c] Jean-Marc Nuzillard,[b] Jean-Hugues Renault[b*]

[a]University of Reims Champagne Ardenne, ESCAPE EA7510, 51097 Reims, France
[b] University of Reims Champagne Ardenne, CNRS, ICMR 7312, 51097 Reims, France
[c] University of Reims Champagne Ardenne, CRB National reference Centre on Toxoplasmosis, 51097 Reims, France
[d] NatExplore, 51140 Prouilly, France
*Correspondence should be addressed to S.R. (simon.remy@univ-reims.fr)

The chemical profiling of complex natural mixtures emerges as a pivotal avenue of investigation for the discovery of new bioactive compounds. It requires a dereplication step generally based either on liquid chromatography-high resolution tandem mass spectrometry (LC-HRMS^2) or on nuclear magnetic resonance (NMR) to quickly identify known compounds. The high sensitivity of MS results in numerous but sometimes incorrect candidate compounds, whereas the greater universality of NMR leads to fewer but more accurate annotations. These two analytical techniques are rarely used in combination despite their complementarity. This study focuses on the chemical profiling of *Larix decidua* (Pinaceae) bark by joint LC-HRMS^2 and ^{13}C NMR data analysis and by querying custom *in silico*-generated chemical databases. MS-based dereplication allowed the annotation of 135 MS^2 spectra with at least two different annotation tools. Twenty-five compounds were annotated in parallel by NMR spectra analysis, including two previously undescribed myrtenic acid derivatives. Sixteen of these compounds were already reported in the Pinaceae family. Twelve compounds were jointly annotated with a high confidence level by comparing LC-HRMS^2 and ^{13}C NMR dereplication results, including compounds not reported to date in *Larix decidua*. Our results show the benefits brought by combining LC-HRMS^2 and ^{13}C NMR data and by querying custom *in silico* chemical databases to enhance the confidence level of data annotation during the chemical profiling of complex natural extracts.

Introduction

Living organisms produce low molecular weight organic natural products, called metabolites, to meet their functional needs. Primary metabolites are involved in vital biochemical pathways, while secondary or specialised metabolites mediate the interactions of living organisms with their biotic and abiotic environment. The locution "natural product" (NP) generally refers to secondary or specialised metabolites and primary metabolites, excluding macromolecules such as proteins, polysaccharides, or nucleic acids. Phytochemistry aims to chemically characterise NPs from various living organisms and to link this chemical information with biological activity profiles to address research questions mainly in chemical ecology or new active substances discovery for different economic sectors such as pharmaceutical, cosmetic, and agrochemical industries.

Three major events had a considerable impact on NP research. The first, in 1992, is linked to the Rio Convention on Biological Diversity, which laid the foundations for the Nagoya Protocol on access to genetic resources and the fair and equitable sharing of benefits arising from their use. This new regulation has been perceived as a constraint by many pharmaceutical companies and has contributed to a slowdown in research in the field. Nevertheless, an inventory of NPs introduced as drugs between 1981 and 2019 has been carried out by Newman and Cragg. Of the 1,881 molecules brought to market during these 39 years, 71 are NPs *stricto sensu*, 356 are derivatives of NPs, often obtained by semi-synthesis, and 272 are obtained by organic synthesis while having a pharmacophore inspired by NPs. This demonstrates, if proof was needed, that NPs are still an essential source of inspiration for Western medicine. They are also playing an increasingly important role in the cosmetics industry.

The second event concerns the advent of the bioeconomy and biobased chemistry concepts, which aims to add value to biomasses, mainly agricultural, algal or forestry, without competing with food resources. Therefore, while biodiversity hotspots have historically attracted the most significant research interest, the chemical profiling of dedicated biomasses or co-products from existing industries has become an essential challenge for achieving carbon neutrality by 2050 *via* the circular economy concept. This aspect of NP research is reinforced by recent results demonstrating unambiguously that the planetary boundary related to the introduction of new chemical entities (*i.e.,* derived from organic synthesis) has widely been exceeded.

The last event concerns the increasing number of phytochemists who recently have actively engaged in addressing the principles of open science and promoting the production of FAIR data. Significant initiatives have been launched in this context, enabling our research activities to benefit from essential open-access tools in spectral data processing and databases.

Although the potential for NPs to reach high added value sectors and the interest in studying them for fundamental research questions are significant, the phytochemistry work often remains complex, time-consuming, tedious, and costly. Together with the entry into force of the Nagoya Protocol, this is one of the reasons for which industrial research programs on natural products have slowed down or even stopped since the beginning of the 2000s. Nevertheless, the massive arrival over the last 15 years of new dereplication approaches enabling known compounds to be chemically characterised as far upstream as possible in the process (1) has given new impetus to phytochemical research, both at academic and industrial levels.(2) Nevertheless, identifying unambiguous structures by describing their unequivocal molecular topology and their geometry, among mixtures, remains challenging. Hence, "annotation" is commonly used when structural information is incomplete.

The two main spectroscopic methods used to study complex mixtures of NPs are nuclear magnetic resonance (NMR) and mass spectrometry (MS), often hyphenated with HPLC for the latter.(3) Their sensitivity and universality performance affects the number and the relevance of annotated compounds.

As stated by Sumner *et al.,* exploiting the advantages of at least two unrelated techniques (*e.g.,* NMR and MS) should improve the annotation reliability.(4) Various strategies (online *versus* offline combinations) were reported by Letertre *et al.*(3) and Marshall *et al.*(5) and are shown in Fig. S1.

When examined under the same conditions, a compound originating from two distinct biological sources is expected to display identical spectral features. Consequently, rapidly identifying well-known compounds, achieved by matching experimental data with established reference spectra, is generally straightforward, assuming the necessary spectral data is accessible. However, this process becomes considerably more intricate when dealing with complex mixtures.

Presently, while structures within natural product structural databases are relatively easy to retrieve (*e.g.,* LOTUS, COCONUT-DB, NP-MRD, DNP, NPAtlas, ChEBI, KNApSAcK, *etc.*), the main challenge lies in locating databases where each structure is associated with its corresponding spectral data (*e.g.*, GNPS, MarinLit). Access to comprehensive and easily searchable experimental spectral data remains crucial to ensure a high confidence level in annotations. (6) Nevertheless, natural product databases encounter limitations, such as incompleteness, inconsistency, availability loss, and/or restricted query flexibility. Additionally, data from NMR and MS are scattered throughout scientific publications in formats that do not facilitate automated data extraction for constructing databases. As a result, even when such data exist in the literature, they are not uniformly incorporated into databases.

Due to these limitations, querying a database for known natural products is hindered by barriers that significantly slow down the overall data mining process. As a solution, predicting spectral data has emerged as a way to improve current spectral databases. In this context, existing spectral databases serve a broader purpose than expediting dereplication (*i.e.,* identifying known compounds). They also serve as the cornerstone for developing algorithms capable of predicting spectral data (*e.g*., MetFrag, CFM-ID, QCxMS). These tools can then enrich current experimental spectral databases by incorporating simulated *in silico* spectral data (*e.g.,* UNPD-ISDB, FooDB, AntiBase, Mona, *etc.*), subsequently streamlining the dereplication process (*e.g.,* Network Annotation propagation) or aiding to develop robust systems for automatic *de novo* elucidation (*e.g.,* SIRIUS, Sherlock (7)). However, the DB size and content impact annotation performance, as a more extensive database increases the probability of returning irrelevant candidates, whereas a database that is too restricted *a priori* may lead to the omission of annotations.(8) DB content should be *a priori* restricted by taxonomy, either on biological and/or chemical grounds, to reduce the number of irrelevant matching spectra.(9)

The build of such DB is nowadays an obstacle course. Large chemical libraries must be first surveyed for compounds, according to criteria defined by the purposes of the study. The molecular structures of the returned compounds must then be gathered into a structural DB. This implies all the issues concerning the stereochemistry description and the drawing of the molecular structures. Finally, these structural DBs should be able to be used as inputs for spectral prediction software. Not all software is user-friendly, and the general workflow, including all the required unit steps, is not automated and available under a single software package.

This work aims to present a comprehensive workflow encompassing the creation of a relevant structural/spectral database and the combined exploitation of MS^2 and ^{13}C NMR data for the chemical profiling of a bio-based extract, with a specific focus on an ethyl acetate solid-liquid extract from *Larix decidua* bark.

Results

The dereplication workflow described hereafter and applied to the chemical profiling of *Larix decidua* (Pinaceae) bark ethyl acetate crude extract is based on the combination of LC-$HRMS^2$ and ^{13}C NMR data to annotate as many NPs as possible with the highest confidence level. First, barks were ground to powder and then macerated with EtOAc to obtain a crude extract with a mass yield of 6.2 %. A 3.81 g portion of this extract was submitted to centrifugal partition chromatography (CPC) fractionation, leading to 12 chemically simplified fractions.

These fractions were submitted to LC-$HRMS^2$ and to ^{13}C NMR analysis. The resulting data were subjected to processing and visualisation procedures. The $HRMS^2$ data were presented as molecular networks, using Ion Identity Molecular Network (IMN), Feature-Based Molecular Network (FBMN), and Propagated Annotation Network (NAP) workflows.(10–12) ^{13}C

NMR data were analysed by hierarchical cluster analysis after chemical shift alignment, and the resulting chemical shift clusters were visualised as a heatmap.(13) The HRMS^2 and NMR data were then annotated using spectral comparison with structures contained in a custom DB created by selecting structures from the LOTUS DB according to taxonomic criteria and with the help of *in silico*, experimental and *de novo* elucidation tools.(14) The results from both workflows were compiled, then ranked and finally a confidence score was assigned to each annotation using a script developed for this purpose.

Structural and *in silico* spectral database of compounds from Pinaceae

Creating a database containing NPs associated with their spectral properties requires collecting the corresponding chemical structures. At first, in the framework of this study, a DB containing the NPs isolated from plants belonging to the Pinaceae family was created. The corresponding chemical structures were downloaded from the LOTUS database using the VersaDB GUI. VersaDB is a recently developed Python-based graphical user interface that integrates the open-source database and webserver LOTUS, CFM-ID, and nmrshiftdb2 in a unique dereplication workflow.(14) It has been designed to create local custom databases of NPs selected according to biological and chemical taxonomic criteria from the LOTUS DB. The resulting structural DB and the corresponding MS and ^{13}C NMR predicted spectral DBs can be further used to perform dereplication of complex mixtures through dedicated platforms and software such as SIRIUS, NAP (MetFrag algorithm will use the structural database to predict and compare spectra to experimental data), GNPS, MetGem and ACD/Labs DB.(14)

Structure inquiry was initiated targeting the Pinaceae plant family. As a result, 2,790 non-redundant structures were obtained. The 5 most represented chemical superclasses were distributed as follow: diterpenoids (469, 16.8%), triterpenoids (316, 11.3%), sesquiterpenoids (306, 11.0%), flavonoids (287, 10.3%), and lignans (173, 6.2%). Structures which could not be classified by NPClassifier were labelled as “nan”. (Fig. S2B)

The structural DB was then used to build the MS^2 and the ^{13}C NMR spectral DBs using VersaDB program. The default prediction settings were applied for both prediction tools within the VersaDB application (CFM-ID4.2.6 and NMRShiftDB), and MS^2 spectra were predicted for three collision energies in $[M+H]^+$ mode. The prediction was operated in 1 h 15 min and 26 sec with a PRECISION 3650 computer, equipped with an Intel(R) Core(TM) i9-10900K CPU @ 3.70GHz 3.70 GHz processor and with 64.0 Go RAM.

The predicted ^{13}C NMR chemical shift values were then imported to the database management software ACD/NMR Workbook to check and correct the predicted chemical shift of each structure. Among inspected structures, 258 contained suspicious chemical shifts and thus were corrected using the ACD/NMR Workbook Suite. This database was then used for spectral matching with each chemical shift clusters resulting from the HCA of experimental spectra.

Concerning MS^2 spectra, CFM-ID failed to predict spectra from 3 compounds since these were molecules for which the SMILES retrieved from LOTUS contained more than one structure. Consequently, MS^2 predicted spectral DB gathered 8,361 MS^2 spectra (3 collision energies x 2,787 compounds).

To benefit from the NAP workflow as a complement to the forthcoming spectral matching with experimental databases, the structural database has been enriched, using the http://dorresteinappshub.ucsd.edu/, with the following information: "MonoisotopicMass InChI SMILES Identifier InChIKey2 InChIKey1 MolecularFormula kingdom_name superclass_name class_name subclass_name”. (15)

Compound identification from ^{13}C NMR data by the CaraMel workflow

The CaraMel chemical profiling workflow (13) is based on the unsupervised clustering of experimental ^{13}C NMR chemical shift values using the peak intensity profiles across centrifugal partition chromatography (CPC) fractions as discriminating features, as explained in the Method section (^{13}C NMR dereplication: CaraMel workflow). The resulting heatmap highlighted 25 chemical shift clusters (Fig. 1) from the 403 chemical shift values picked in the spectra of 12 CPC fractions of the EtOAc *Larix decidua* bark extract.

The chemical shift values in each cluster were compared to those from the custom Pinaceae ^{13}C NMR predicted database (2,790 structures associated with their predicted ^{13}C NMR chemical shifts).

As a result, 12 compounds were directly annotated, as shown in Fig. 1, by submitting the ^{13}C chemical shifts of the corresponding cluster from HCA to the ^{13}C NMR Pinaceae database, using the structure search engine included in the ACD/NMR Workbook database management tool. If relevant, the annotations of these compounds contained in the EtOAc bark extract and their relative configurations were confirmed by the complementary analysis of ^{1}H and 2D NMR data (COSY, HSQC and HMBC spectra). These compounds, marked by a blue star symbol in Fig. 1, are catechin (cluster 1), epicatechin (cluster 2), quercetin-3-rhamnoside (cluster 3), trans-astringin (cluster 6), lavandoside (cluster 9), larixyl acetate (cluster 11), ferulic acid (cluster 13), larixinol isomers (cluster 14), 13-epimanool (cluster 20), isopimaric acid (cluster 21), dehydroabietic acid (cluster 22) and 7-oxo-dehydroabietic acid (cluster 23).

Thirteen other structures were annotated (Fig.1) based on the initial proposals returned by ACD/NMR workbook (although erroneous, they often provide very useful hints for "manual" annotation) and from 2D NMR data of chemically simplified fractions: acetic acid (cluster 4), piceatannol-3’-o-glucoside(*trans*) (cluster 5), glucosyl frambinone (cluster 7), glycerol monoacetate (cluster 8), glucosyl-*trans-para*-coumaric acid (cluster 10), 2-[2,4-dihydroxybenzoyl)oxyphenyl]acetic acid (cluster 12), dianthoside (cluster 15), oleic acid (cluster 18), linoleic acid (cluster 19), butanol (cluster 1’), tannins (catechic units) (cluster 2’). Three of these NPs were not directly annotated even though they were present in the VersaDB Pinaceae database generated for this study because many of their chemical shifts have been clustered with those of other compounds (*e.g., trans*-astringin (cluster 6) and *trans*-piceatannol-3'-o-glucoside (cluster 5), or lavandoside (cluster 9) and glucosyl *trans-para*-coumaric acid (cluster 10). Moreover, the structures reported in Table S8 of two previously non-reported compounds: rhamnosyl-(1->6)-glucosyl-myrtenic acid and arabinosyl-(1->6)-glucosyl-myrtenic acid (respectively cluster 17 and 16), were elucidated based on 1D and 2D NMR data of CPC fraction 3. HMBC correlations were observed between protons H-1' and carbon C1, between proton H-1'' and carbon C-6', and between protons H-6' and carbon C-1', thus confirming the structure of arabinosyl-(1→6)-glucosyl-myrtenic acid. In addition, the correlation between carbons C-4'' and C-5'' and proton H-6'' confirmed the rhamnosyl-(1→6)-glucosyl-myrtenic acid structure. These two compounds were also detected while analysing CPC fraction 3 by mass spectrometry, and the corresponding features have been annotated with their molecular formula (respectively $C_{21}H_{32}O_{11}$: feature 248, 483.18 m/z, $[M+Na]^+$, feature 294, 499.16 m/z,$[M+K]^+$; and C22H34O11: feature 306, 497.20 m/z, $[M+Na]^+$). Their structures are close to that of sacranoside A, but with myrtenic acid as aglycon moiety instead of myrtenol. Interestingly, the monoterpene myrtenol and oxidised derivatives have also been previously identified from the bark of *Picea abies* (Pinaceae).(16)

As a result, 16 of the 25 compounds identified by the NMR workflow were previously reported in the Pinaceae and are marked with the PNC symbol in Fig. 1. The chemical diversity of the 25 identified compounds was studied according to NPClassifier output, including five diterpenoids (larixyl acetate, 13-epimanool, isopimaric acid, dehydroabietic acid, 7-oxo-dehydroabietic acid), five flavonoids (catechin, epicatechin, suercetin-3-rhamnoside, larixinol isomers, tannins (catechic units)), three phenylpropanoids (lavandoside, glucosyl-*trans-para*-coumaric acid, ferulic acid), three fatty acids (oleic acid, linoleic acid, butanol), two stilbenoids (*trans*-astringin, piceatannol-3-o-glucoside), two monoterpenoids arrabinosyl-glucosyl-myrtenic acid, rhamnosyl-glucoside-myrtenic acid), one aromatic polyketide (2-[2,4-dihydroxybenzoyl)oxyphenyl]acetic acid), one cyclic polyketide (dianthoside), two undefined compounds (glucosyl-frambinone, glycerol monoacetate), and the erroneous annotation from NPClassifier for acetic acid as small peptide. Fig. S2B shows the global chemical distribution for these compounds.

Table 1 and Fig. 1 show the distribution of the identified compounds within the 12 fractions. The ^{13}C and ^{1}H chemical shifts of the annotated compounds are reported in Table S8.

LC-HRMS^2 data annotations through molecular network workflows

Dereplication applied to CPC fractions was performed using LC-HRMS^2, and experimental spectral data were subjected to annotation using specific algorithms, including IIMN (17), FBMN (18), NAP (10), and SIRIUS (19). The experimental data were also compared (18) to the simulated MS data of the custom Pinaceae structural and spectral DB using the open-source software MetGem (17). However, these results are not reported due to the low annotation rate, attributed to the poor quality of the predicted spectra as shown in Fig. S4.

The raw data processing resulted in the detection of 1,958 sets of signals corresponding to a chemical entity, also called features. MS^2 spectra similarity-based molecular network was generated using the Feature Based Molecular Network (FBMN) workflow (see GNPS platform (18)). The chromatographic peak shape correlation analysis was integrated into the molecular network by connecting and collapsing the different ion species of the same molecule via the Ion Identity Molecular Networking (IIMN), which enhances annotations within the fragmentation pattern similarity molecular network. Annotations of MS^2 spectra through spectral matching with reference spectra (FBMN) were improved through the use of an *in silico* approach known as Network Annotation Propagation (NAP), as well as employing a *de novo* strategy provided by the SIRIUS software

IIMN + FBMN

The FBMN workflow returned 83 matches with the GNPS speclibs collection, 54 unique spectra (46 unique compounds): seven candidates were validated by the IIMN, encompassing six distinct spectra associated with unique compounds: epicatechin, quercetin, and curcumin, as well as norethindrone, gestoden and a piperazine-derivative antihistamine: cyclizine. The annotation of these three latter compounds within the natural substance mixture is indeed surprising, considering that they are of synthetic origin. However, norethindrone and gestoden, two synthetic steroids, could be considered as natural product-like compounds when calculating their natural product likeness scores (NaPLeS = 1.8 and 2, respectively), in comparison to cyclizine (NaPLeS = -0.6). The more positive the score, the higher the NP-likeness.(19) The most representative annotated chemical class were steroid (29%), diterpenoïd (23%), and flavonoid (17%), as shown in Fig. S2. B.

NAP

Fifty-three of the FBMN matches were also annotated with NAP. Thirty-seven of these annotations result from at least one structure candidate of the *in silico* fragmentation search with the MetFrag algorithm (20), including 25 unique spectra. NAP allowed the annotation of 156 not previously annotated features, using only *in silico* fragmentation search with MetFrag. Hence, the *in silico* fragmentation search allowed the annotation of 39.8% of all the considered features by the NAP algorithm versus 4.2% of all data set using only experimental spectral database search. In total, NAP led to the annotation of 9.9% of all the detected features.

SIRIUS

The *de novo* strategy facilitated the annotation of 1,786 features with at least one molecular formula and corresponding adduct type. Among these, 896 features garnered multiple proposals, while 890 features were associated with a single proposal, resulting in a cumulative count of 2,847 proposals. Following the subsequent re-ranking step, 1,300 distinct molecular formulas were attributed to the 1,786 features.

Integration of the Pinaceae structural database into the annotation workflow enabled the annotation of 498 features, each associated with at least one potential structural candidate retrieved from the Pinaceae structural database, encompassing the assignment of 196 distinct structures as primary candidates. The five most frequently assigned structures the following: LTS0139452 ($C_{19}H_{28}O_2$, 22 times), LTS0056945 ($C_{20}H_{28}O_4$, 16 times), LTS0251392 ($C_{20}H_{28}O_3$, 15 times), LTS0084143 ($C_{20}H_{30}O_3$, 13 times) and LTS0241153 ($C_{20}H_{32}O_2$, 11 times). Ultimately, 1,728 features have been annotated up to the chemical class level using the webservice CANOPUS integrated in SIRIUS. These annotations are categorised across seven biosynthesis pathways: terpenoids (663 instances), fatty acids (307 instances), polyketides (234 instances), shikimates and phenylpropanoids (160 instances), alkaloids (116 instances), amino acids and peptides (106 instances), and finally carbohydrates (41 instances).

Combination of ^{13}C NMR and LC-HRMS^2 data annotations

The experimental spectra were annotated through three strategies: experimental spectral databases matching, *in silico* spectral databases matching, and *de novo* annotation. The first one relies on the use of GNPS workflows and FBMN. The *in silico* annotation workflow relies on NAP and CaraMel. The manual analysis of NMR data and the application of the SIRIUS software, both employed in this study, are integral components of the third strategy. The latter allowed for the annotation of experimental spectra without depending on spectral databases, even in cases undisclosed chemical structures were involved. A new workflow was created to combine the annotations from all the GNPS workflows (*i.e.,* FBMN, NAP), SIRIUS and the ^{13}C NMR annotation workflow called CaraMel. This new workflow named « CATHEDRAL » for « Combining THE DiffeRent Annotation tooLs. » is an in-house script written in Python. This script allows the assignment of a confidence level score using a custom confidence level system, determined based on the outcomes of the comparative analysis.

First, FBMN and NAP results are merged to retrieve the spectral matches *versus* experimental spectral libraries from GNPS and Pinaceae *in a silico* database simulated by MetFrag.

Secondly, the resulting metadata is exported as a "recap.csv" file. A PRED column is added to the table, and features are "scored" as follows: 1 for the FBMN annotated features, 2 for the features without any FBMN annotation but with at least one NAP – MetFrag candidate, and 3 for all the remaining features.

The comparison process starts with the SMILES from the FBMN workflow and the MetFragSMILES from the NAP workflow. The corresponding InChIKeys are then generated *via* functions from the RDKit Python library.

The canonical SMILES from ^{13}C NMR annotations structures, obtained from ChemDraw 20.1.1., and gathered into a file called «smile_nmr.txt», follow the same process as above.

The resulting InChiKeys for each feature and CaraMel annotations are thus compared to the corresponding SIRIUS annotations InChiKeys.

CATHEDRAL.py can be executed through a detailed step-by-step jupyter notebook (CATHEDRAL.ipynb) under an RDKit anaconda environment. The recap.csv file is first opened, and features corresponding to the PRED score = 1 are selected. FBMN annotations are compared to MetFrag candidates for each feature. In the second place, FBMN annotations are compared to the corresponding SIRIUS candidates. In the third place, the FBMN, NAP (MetFrag), and SIRIUS annotations are compared to each other to highlight those with the same candidate through the three annotation tools. The greater the number of tools giving the same annotation, the better the score and priority accorded to annotation tools based on their effectiveness in providing a relevant structural candidate (*c.f.* CASMI). Thus, a spectral match with a reference experimental spectrum (*i.e.,* FBMN) will be prioritised over a predicted spectrum (*i.e.* MetFrag). However, *de novo* annotations (*i.e.*, SIRIUS) will be ranked higher than the latter. Furthermore, the scores are better when a candidate is corroborated by NMR, unlike cases where the annotation solely relies on MS/MS tools. Consequently, the features with consistent annotation along the three mass annotation tools are scored with a Confidence_Level = 9. The features with consistent annotations between FBMN and SIRIUS are scored 10, the features with consistent annotations between FBMN and MetFrag are scored 12, and features with consistent annotations between MetFrag and SIRIUS are scored 11+. The corresponding annotation for each scored feature is then compared to the CaraMel candidates. If the MS^2 candidate is part of CaraMel annotations, the header « Is_NMR_Annotated_SU » = 1, and the corresponding molecular name is given to the header « Molecular_Name_SU » else « Is_NMR_Annotated_SU » = 0, where SU means Sum-Up. Consequently, features previously scored (9, 10, 11, 11+, 12), also annotated by CaraMel, are respectively re-scored 1,2, 3+ or 4. Then, only the data with a PRED score = 2 are selected. The features with consistent annotations between MetFrag and SIRIUS are scored 11. The annotation is then compared to the CaraMel annotations, as previously explained. If these features are also annotated with CaraMel, the score is raised to 3.

If an annotation is common to at least two MS^2 annotation tools, the third tool annotation is equally compared to the CaraMel annotations. If it matches, the corresponding molecular name is reported in the header « 3rd_Tool_SU » of the sum-up file.

All the other features that do not have consistent annotations between the different MS^2 tools are equally compared to the CaraMel annotations. If the annotation from FBMN matches, Confidence_Level corresponds to 5; if one annotation candidate matches for at least the NAP workflow, Confidence Level is equal to 7. Finally, if only a candidate annotated from SIRIUS matches with a CaraMel annotation, Confidence_Level will be 6. CaraMel annotations that do not correspond to any mass annotation are scored 8. All the corresponding confidence levels are reported in Table. 2.

The comparison outcomes are then written as a table in a sum-up file called « df_resume_confidence.tsv » starting with the following table column headers: « 3rd_Tool_SU, Confidence_Level, Feature_SU, GNPS_SU, Is_NMR_Annotated_SU, Molecular_Name, NAP_SU, Not_Matching_Tool_Annotation_SU, Rt_SU, SIRIUS_SU, m/z ». This sum-up file is then imported to the previously created network collection in Cytoscape software, and metadata are overlayed on the graph *via* the in-house style, as exemplified in Fig. 2.

CATHEDRAL workflow highlights up to 41 unique compounds, annotated with at least two indiscriminately MS/MS or NMR dereplication tools, including 7 compounds in Fig. 3. The outcomes are reported in Tables. S1-S7. The best confidence level annotated structures correspond to those with the same candidates for all annotation tools. They are all classified as flavonoid derivatives and FBMN / SIRIUS jointly annotated features. In addition, 21 among the 27 structures jointly annotated by NAP / SIRIUS correspond to diterpenoids (abietane and derivatives) and triterpenoids. Most features annotated jointly by at least two different tools are $[M+H]^+$ (37 among the 50 iterations of annotated compounds). The structure candidates highlighted by NAP / SIRIUS and SIRIUS / CaraMel comparisons also include other ion species and in-source modifications, respectively $[M+Na]^+$ and $[M+K]^+$ adducts or in-source modifications.

As a result, the strategy for matching experimental spectral databases appears to prioritize annotating features associated with flavonoid chemical families and $[M+H]^+$ adducts over other chemical families and ion species. Moreover, the chemical diversity is more prominent when considering the chemical classes related to annotated compounds from *in silico* and *de novo* strategies.

Finally, the *de novo* strategy for mass data annotation allowed the annotation of nine of the 25 structures found by ^{13}C NMR data annotation. Five compounds among them were neither jointly listed by experimental nor predicted mass spectral database matching.

Conclusion & Discussion

Summary

Untargeted metabolomics produces vast amounts of analytical data, yet the laborious analysis does not always result in a high annotation confidence level.

Due to the complementarity of ^{13}C NMR and LC-HRMS^2, their combined use can potentially increase the number of detected compounds and significantly enhance the confidence level in their identification. Consequently, various strategies have been developed recently, such as online (*e.g.,* LC-MS-SPE-NMR) and offline approaches (*e.g.* multiblock).(3) However, the former requires bulky and expensive equipment (*i.e.,* ensuring compatibility of experimental conditions), while the latter enables in-depth data exploration but requires advanced expertise in chemometrics.

Hence, our strategy revolves around the cross-comparison of annotations from both ^{13}C NMR and LC-HRMS^2 spectral datasets. Despite the high simplicity of our cross-comparison process, two major challenges persist. On one hand, automation should make conventional manual procedures more efficient. On the other hand, the expected gain in confidence is noticeable only for compounds detected by both techniques.

Consequently, compound annotation by a single technique remains ambiguous and requires more traditional strategies, such as exploiting multidimensional NMR data. Additionally, the quality of the annotation process, which heavily relies on spectral comparison, depends on the availability of relevant high-quality spectral or structural databases.

The initial step of the developed workflow involved the automatic compilation of tailored structural and predicted spectral databases, including the biological origin of the sample. This approach grants the crucial role of contextualization in NP research by narrowing down the pool of candidate structures whose affiliation with a specific plant species, genus, or family has been previously documented. It is also imperative to clarify that this step relies on queries from the LOTUS natural product database, built directly from Wikidata. The data are organized in triples, containing referenced structure-organism pairs that establish relationships between distinct molecular structures and the living organisms from which they were identified. Furthermore, the information comprised in LOTUS is associated to research findings provided by the scientific community, and as a result, it can suffer from the voluntary or involuntary omission of certain compounds in the chemical description of a given species, genus, or family. Nevertheless, to our knowledge, LOTUS represents the most extensive structural database available, including numerous metadata and is well-suited for automated online queries.

Consequently, integrating the Pinaceae family structural database with MS annotation tools resulted in a notable increase in annotated features. Interestingly, the Pinaceae contextualization of databases employed in NAP and SIRIUS processes enabled us to achieve 25% and 10% annotation rates, respectively, surpassing the 5% annotation rate achieved using the GNPS spectral library. However, the *in silico* annotation strategy is still limited by the accuracy of spectral prediction algorithms (*e.g., the* MS^2 predicted database generated by CFM-ID 4.0). According to LOTUS, this trend was also observed with NMR, where 16 of the 25 identified compounds were associated with the Pinaceae family. The remaining nine compounds were manually verified not to be genuinely related to any Pinaceae species: only dianthoside was an exception.(21) Furthermore, the last eight molecules correspond to substances that can be considered handling pollutants, such as butanol. Another concerns tannin, for which the complete molecular structure has not been elucidated. Others are not previously documented compounds, such as the myrtenic acid glycosides derivatives. The last molecules, piceatannol-3'-o-glucoside (*i.e.,* a positional isomer of *trans*-astringin previously associated with Pinaceae), 2-[2,4-dihydroxy-6-(4-hydroxybenzoyl)oxyphenyl]acetic acid glycerol acetate and its hydrolysis product acetic acid, are likely the only compounds in the set that have never been described in Pinaceae, despite previous documentation. Given its meticulous compilation work, these results affirm our choice of considering LOTUS as the optimal starting point for contextualizing our databases.

The second step of this workflow involved efficiently scoring candidates through the CATHEDRAL script, using a custom confidence level system. Hence, a confidence level was assigned to 52 compounds, annotated with at least two annotation tools. This original scoring strategy markedly enhances the compound annotation reliability and reinforces the structure-organism pairs. This approach also reduces the requirement for tedious manual analysis, including successive purification and structural elucidation steps to confirm the presence of compounds in the studied matrix. As a result, it accelerates chemical profiling through dereplication.

Recent research has ultimately shown that expanding analytical dimensions for multidimensional metabolite annotation (*e.g.* Rt, m/z, chiroptical data, *etc.*) offers a significant boost, enhancing the accuracy and coverage of metabolites. More specially, ion mobility (IM) MS has emerged as a powerful technology that allows for the metabolite measurement of collision cross-section values (CSS). Consequently, the CCS aids in isomer characterization by providing an additional separation dimension to mass, increasing the identification confidence. (22–25) Similar to the four-dimensional untargeted metabolomic concept introduced by Cai *et al.*(26), including multidimensional data and/or metadata (*i.e.,* chemical ontology, scoring strategies, *etc.*) would significantly benefit our assessment methodology and, consequently, the consistency of identifications.

In turn, data from new identified compounds will participate to the completion of existing databases, thereby improving our understanding of the diversity of specialized metabolites biosynthetic pathways. This knowledge will contribute to optimizing the production of high added value compounds in the pharmaceutical industry, as seen in the case of artemisinin production in *Artemisia annua* through the up-regulation of amorphadiene synthase (multiplying yield by 2.3).(27)

Finally, the past and present habits in natural product chemistry research and publishing have led to issues, such as poorly defined chemical composition of plants. Insufficient description and standardization in the natural product practices, coupled with the irresponsible and inefficient management of experimental reference data, can render further research unreliable due to inherent inconsistencies in the substance reporting. This has contributed to the relative neglect of pharmacological studies on natural substances.(28) Additionally, it is essential to enhance the user-friendliness of tools to encourage phytochemists to persevere with better practices.

Despite the need for basic programming skills, this user-friendly approach contributes to global standardization in natural product research, following the ChemBioPrint® framework. Its consistency in providing high confidence levels for identified components in complex mixtures of natural products eases the selection of authentic molecules for pharmacological studies, such as reverse docking, thereby reducing compound to drug development time.

References

1. Beutler JA, Alvarado AB, Schaufelberger DE, Andrews P, McCloud TG. Dereplication of phorbol bioactives: Lyngbya majuscula and croton cuneatus. J Nat Prod [Internet]. 1990 Jul [cited 2019 Nov 25];53(4):867–74. Available from: https://pubs.acs.org/doi/abs/10.1021/np50070a014
2. Hubert J, Nuzillard JM, Renault JH. Dereplication strategies in natural product research: How many tools and methodologies behind the same concept? Phytochemistry Reviews. 2017;16(1):55–95.
3. Letertre MPM, Dervilly G, Giraudeau P. Combined Nuclear Magnetic Resonance Spectroscopy and Mass Spectrometry Approaches for Metabolomics. Vol. 93, Analytical Chemistry. American Chemical Society; 2021. p. 500–18.
4. Sumner LW, Amberg A, Barrett D, Beale MH, Beger R, Daykin CA, et al. Proposed minimum reporting standards for chemical analysis: Chemical Analysis Working Group (CAWG) Metabolomics Standards Initiative (MSI). Metabolomics. 2007;3(3):211–21.
5. Marshall DD, Powers R. Beyond the paradigm: Combining mass spectrometry and nuclear magnetic resonance for metabolomics. Vol. 100, Progress in Nuclear Magnetic Resonance Spectroscopy. Elsevier B.V.; 2017. p. 1–16.
6. Schymanski EL, Jeon J, Gulde R, Fenner K, Ruff M, Singer HP, et al. Identifying Small Molecules via High Resolution Mass Spectrometry: Communicating Confidence. Environ Sci Technol. 2014 Feb 18;48(4):2097–8.
7. Wenk M, Nuzillard JM, Steinbeck C. Sherlock—A Free and Open-Source System for the Computer-Assisted Structure Elucidation of Organic Compounds from NMR Data. Molecules. 2023 Feb 2;28(3):1448.
8. Bergold AN, Heaton P. Does filler database size influence identification accuracy? Law Hum Behav. 2018 Jun 1;42(3):227–43.
9. Nuzillard JM. Taxonomy-Focused Natural Product Databases for Carbon-13 NMR-Based Dereplication. Analytica. 2021 Jun 28;2(3):50–6.
10. Schmid R, Petras D, Nothias LF, Wang M, Aron AT, Jagels A, et al. Ion identity molecular networking for mass spectrometry-based metabolomics in the GNPS environment. Nat Commun. 2021 Dec 1;12(1).
11. Nothias LF, Petras D, Schmid R, Dührkop K, Rainer J, Sarvepalli A, et al. Feature-based Molecular Networking in the GNPS Analysis Environment. bioRxiv. 2019 Oct;812404.
12. da Silva RR, Wang M, lix Nothias LF, J van der Hooft JJ, Mauricio Caraballo-Rodríguez A, Fox E, et al. Propagating annotations of molecular networks using in silico fragmentation. 2018;
13. Hubert J, Nuzillard JM, Purson S, Hamzaoui M, Borie N, Reynaud R, et al. Identification of natural metabolites in mixture: A pattern recognition strategy based on 13C NMR. Anal Chem [Internet]. 2014 Mar 18 [cited 2021 Jun 17];86(6):2955–62. Available from: https://pubmed.ncbi.nlm.nih.gov/24555703/
14. Cordonnier J, Remy S, Renault JH, Nuzillard JM. Versa DB: Assisting 13 C NMR and MS/MS Joint Data Annotation Through On-Demand Databases. 2023; Available from: https://doi.org/10.1002/cmtd.202300020
15. Network Annotation Propagation (NAP) - GNPS Documentation [Internet]. [cited 2023 Jan 13]. Available from: https://ccms-ucsd.github.io/GNPSDocumentation/nap/#structure-database
16. Heemann V, Francke W. Gaschromatographisch–Massenspektrometrische Untersuchungen der Flüchtigen Rindeninhaltsstoffe von Picea abies (l.) Karst. Planta Med. 1977 Dec 13;32(08):342–6.
17. Olivon F, Elie N, Grelier G, Roussi F, Litaudon M, Touboul D. MetGem Software for the Generation of Molecular Networks Based on the t-SNE Algorithm. Anal Chem. 2018 Dec 4;90(23):13900–8.
18. Wang M, Carver JJ, Phelan V V., Sanchez LM;, Garg N;, Peng Y;, et al. Sharing and community curation of mass spectrometry data with GNPS. Nat Biotechnol. 2017;34(8):828–37.
19. Vanii Jayaseelan K, Moreno P, Truszkowski A, Ertl P, Steinbeck C. Natural product-likeness score revisited: an open-source, open-data implementation. BMC Bioinformatics. 2012 Dec 20;13(1):106.
20. Wolf S, Schmidt S, Müller-Hannemann M, Neumann S. In silico fragmentation for computer assisted identification of metabolite mass spectra [Internet]. 2010. Available from: http://www.biomedcentral.com/1471-2105/11/148
21. PLOUVIER V. Recherche du dianthoside chez les Pinacées et quelques autres groupes botaniques, et de l'érigéroside chez les Composées-Astérées. Comptes Rendus de l'Académie des Sciences, Série III. 1984;298:749–52.
22. Paglia G, Smith AJ, Astarita G. Ion mobility mass spectrometry in the omics era: Challenges and opportunities for metabolomics and

lipidomics. Mass Spectrom Rev [Internet]. 2022 Sep;41(5):722–65. Available from: https://analyticalsciencejournals.onlinelibrary.wiley.com/doi/10.1002/mas.21686
23. Luo MD, Zhou ZW, Zhu ZJ. The Application of Ion Mobility-Mass Spectrometry in Untargeted Metabolomics: from Separation to Identification. J Anal Test [Internet]. 2020 Jul 25;4(3):163–74. Available from: https://link.springer.com/10.1007/s41664-020-00133-0
24. Zheng X, Smith RD, Baker ES. Recent advances in lipid separations and structural elucidation using mass spectrometry combined with ion mobility spectrometry, ion-molecule reactions and fragmentation approaches. Curr Opin Chem Biol [Internet]. 2018 Feb;42:111–8. Available from: https://linkinghub.elsevier.com/retrieve/pii/S1367593117301540
25. Picache JA, Rose BS, Balinski A, Leaptrot KL, Sherrod SD, May JC, et al. Collision cross section compendium to annotate and predict multi-omic compound identities. Chem Sci [Internet]. 2019;10(4):983–93. Available from: http://xlink.rsc.org/?DOI=C8SC04396E
26. Cai Y, Zhou Z, Zhu ZJ. Advanced analytical and informatic strategies for metabolite annotation in untargeted metabolomics. Vol. 158, TrAC - Trends in Analytical Chemistry. Elsevier B.V.; 2023.
27. Wen W, Yu R. Artemisinin biosynthesis and its regulatory enzymes: Progress and perspective. Pharmacogn Rev [Internet]. [cited 2023 Nov 11];(10). Available from: www.phcogrev.com
28. Lai CT, Shan JJ, Rodgers K, Sutherland SK. Challenges in Natural Health Product Research: The Importance of Standardization [Internet]. Vol. 50, Proc. West. Pharmacol. Soc. 2007. Available from: https://www.researchgate.net/publication/5246233
29. Darme P, Spalenka J, Hubert J, Escotte-Binet S, Debelle L, Villena I, et al. Investigation of Antiparasitic Activity of 10 European Tree Bark Extracts on Toxoplasma gondii and Bioguided Identification of Triterpenes in Alnus glutinosa Barks [Internet]. 2022. Available from: https://journals.asm.org/journal/aac
30. Pluskal T, Castillo S, Villar-Briones A, Orešič M. MZmine 2: Modular framework for processing, visualizing, and analyzing mass spectrometry-based molecular profile data [Internet]. 2010. Available from: http://www.biomedcentral.com/1471-2105/11/395
31. Dührkop K, Fleischauer M, Ludwig M, Aksenov AA, Melnik A V., Meusel M, et al. SIRIUS 4: a rapid tool for turning tandem mass spectra into metabolite structure information. Nat Methods [Internet]. 2019 Apr 1 [cited 2021 Jun 17];16(4):299–302. Available from: https://www.nature.com/articles/s41592-019-0344-8
32. Stein SE, Scott DR. Optimization and Testing of Mass Spectral Library Search Algorithms for Compound Identification.

Supporting information

The relevant source code of CATHEDRAL script for the assignment of a confidence level score determined based on the outcomes of the comparative analysis comparison of annotations, can be found on GitHub at https://github.com/Jcrdnr/CATHEDRAL.

Acknowledgements

Julien Cordonnier thanks the Grand-Est Region for its financial support, and the European Regional Development Fund (ERDF) for financing the Biomolecules and Biomaterials for regional Bioeconomy (3BR) project.
This paper was typeset with the bioRxiv word template by @Chrelli: www.github.com/chrelli/bioRxiv-word-template

Competing interest statement

The authors have no conflicts of interest to declare.

Keywords

Dereplication – Mass spectrometry – Nuclear Magnetic Resonance – Natural products

Materials and Methods

The dereplication workflows are summarised in Fig. 4.

Plant material

The Nagoya Protocol was adopted by France in 2010 and entered into force in 2014. The French “Office National des Forêts” (ONF) granted the authors permission to collect samples and to use bark extracts at the University of Reims Champagne-Ardenne. *Larix decidua* bark was collected on a standing tree by M. Verdeaux in Signy-L’Abbaye (Ardennes, France), route forestière Roban, 70th plot on the 11th of May 2017. The sample was dried for three days at 30°C before being finely powdered by a hammer mill (VEM Motors GmbH, Germany). The botanical identification of trees was made following the phenotypic characteristics, such as the shape, arrangement, and contours of leaves.(29)

Sample preparation

Collected Larix decidua barks were cleaned to remove unwanted materials such as moss and soil particles, and the crude extract was prepared by maceration of 70 g bark powder in 1 l ethyl acetate (EtOAc) at room temperature for 24 under magnetic stirring. After filtration under vacuum on a 40 µm sintered glass, the solvent was evaporated under vacuum to leave a dry EtOAc extract. The extraction of Larix decidua bark by EtOAc yielded 4.34 g (6.2 %) of crude extract.

Sample fractionation

The crude extract was fractionated by Centrifugal Partition Chromatography (CPC). It is based on liquid–liquid compound partitioning between two non-miscible liquid phases that stay close to their thermodynamic equilibrium. The absence of solid chromatographic support facilitates the fractionation of complex samples, eliminates the risk of stationary phase overloading or clogging, and avoids the potential deterioration of chemical compounds, thus leading to mass recovery rates close to 100 %.

CPC apparatus

CPC experiments were performed on an FCPC200 ® apparatus (Rousselet Robatel Kromaton) equipped with a rotor containing 800 cells (260 ml total column capacity) and connected to Smartline Preparative Pump 1800 (Knauer). The eluent was collected with a LABOCOL Vario-4000 fraction collector (Knauer) in fractions of 20 ml.

Solvents system

The CPC fractionation of the extract was performed in two consecutive steps (CPC1 and CP2) to obtain the broadest possible diversity of fraction polarity.

Gradient elution system n-heptane/ethyl acetate/methanol/water:

The initial mobile phase (lower phase of system 1: n-heptane/ethyl acetate/methanol/water (5:5:5:5, v/v)), the final mobile phase (lower phase of system 3: n-heptane/ethyl acetate/methanol/water (7:3:7:3, v/v)) and the organic stationary phase (upper phase of system 2: n-heptane/ethyl acetate/methanol/water (6:4:6:4, v/v)) were prepared separately.

Isocratic elution system ethyl acetate/acetonitrile/water:

The system composed of ethyl acetate/acetonitrile/water (3:3:4) (system 3) was used in the normal-phase mode (upper phase mobile).

Injection and CPC operating procedure

CPC 1

The column was filled with the stationary phase of the gradient elution system and equilibrated with the initial mobile phase at 10 ml/min and 1600 rpm. The crude extract (3.81 g) was dissolved in 3 ml of lower phase + 3 ml of upper phase and injected into the CPC column through a 20 ml loop. As described in Table, the mobile phase was pumped in descending mode for 110 minutes. 3. The column was extruded by switching the mode selection valve for 15 minutes at 20 ml/min. Fractions of 20 ml were collected over the whole experiment (elution and extrusion) and combined according to their thin layer chromatography profiles (TLC).

CPC 2

The column was filled with the stationary phase (upper phase) of the isocratic system and equilibrated with the lower phase at 10 ml/min and 1800 rpm. Then the CPC1 first fraction (252 mg) was dissolved in 3 ml of lower phase + 3 ml of upper phase and injected into the CPC column by a 20 ml loop. The mobile phase was pumped in ascending mode for 90 minutes. The column was extruded by switching the mode selection valve for 20 minutes at 20 ml/min. Fractions of 20 ml were collected over the elution step and combined according to their TLC profiles.

TLC of CPC fractions

The 20 ml fractions of the two CPC experiments (elution and extrusion) were characterised by TLC. TLC was performed on pre-coated silica gel 60 F254 Merck plates, with the migration solvent system consisting of ethyl acetate/toluene/acetic acid/formic acid (4/6/1/1, v/v). Compound migration was visualised under UV light at 254 nm and 366 nm and revealed by spraying consecutively the dried plates with 50% (v/v) H_2SO_4 acid and vanillin (10 g / l) ethanolic solution followed by heating. As a result, six fractions were obtained from CPC1. The first fraction (consisting in the seven first elution minutes (t0 of the experiment)) was used for CPC2. The other fractions were noted F08, F09, F10, F11 and F12 (Figure S2). The fractions from CPC2 were arranged in reverse order of collection (because CPC1 was performed in descending mode while CPC2 was performed in the ascending mode) to obtain a chemical profile continuity between the two fractionation steps. The fractionation of the bark extract resulted in twelve chemically simplified fractions of polarity ranging from the highest (F01) to the lowest (F12) (Fig. S2), ten produced by elution and two by subsequent column extrusion.

The fractions were then analysed in parallel by LC-HRMS^2 and ^{13}C NMR. Experimental spectral data were then processed and annotated while taking advantage of purposely created databases. A comparison of the resulting candidate structures from both workflows was carried out to enhance confidence in the final profiling outcomes.

Generation of custom structural and spectral databases: VersaDB

The VersaDB GUI was employed for constructing customised structural and spectral databases. The methodology previously outlined was subsequently applied. (14) Briefly, the process involved an automated structural inquiry utilising the 'for all selected categories' function of VersaDB. This function allowed the transmission of HTTP requests *via* the LOTUS API, applying a singular chosen criterion: 'T: All_Taxonomy_DB: family: Pinaceae'. This approach entailed a comprehensive search for structures documented within botanical species affiliated with the Pinaceae family. This was executed by leveraging the entirety of available taxonomic databases originating from the LOTUS Natural Product Occurrence Database.

The resultant compilation of structures, each accompanied by its respective LOTUS ID, was archived in two analogous files: cfmidinput.txt and structuraldb.txt. These files served as input for the MS^2 and ^{13}C NMR spectra prediction or were employed as the structural database within the SIRIUS and NAP platforms. However, it requires formatting adjustments before it can be utilised in the NAP workflow. To achieve this, the file was subjected to processing using the http://dorresteinappshub.ucsd.edu/ web server. The original data, initially presented in the 'SMILES - CompoundID' format, undergo enhancement and transformation into a novel arrangement: "MonoisotopicMass InChI SMILES Identifier InChIKey2 InChIKey1 MolecularFormula kingdom_name superclass_name class_name subclass_name.". Additional metadata from LOTUS, encompassing chemical and physical attributes, taxonomic classification, and chemontology information according to the NPClassifier ontology, were consolidated and stored within a file named "cfmid input.tsv.".

The "predict both properties" functionality within VersaDB was subsequently employed to predict both MS^2 and ^{13}C NMR spectral properties.

Concerning ^{13}C chemical shifts prediction, the VersaDB system incorporated an adaptation of Kuhn and Nuzillard's approach, leveraging the nmrshiftdb2 packages. The outcome was the "13CNMRDatabase.sdf" file containing the structures, their ^{13}C nmrshiftdb2-predicted chemical shifts in the style of ACD/Labs CNMR Predictor, and all compound-related metadata gathered from LOTUS, encompassing biological, chemical taxonomy, and physicochemical properties. This "13CNMRDatabase.sdf" file was then imported into ACD/NMR Workbook Suite 2012 (ACD/Labs, Ontario, Canada), where the predicted ^{13}C chemical shifts underwent validation using the "check chemical shifts" option.

The prediction of MS^2 spectra was conducted using the CFM-ID 4.2.6.0 docker image. MS^2 spectra were predicted across three collision energies, employing the pre-trained CFM-ID models. This process culminated in the creation of the definitive custom *in silico* mass spectral database file, denoted as "MSMSspectraDatabase.mgf", along with the "annotationGNPSformat.tsv" file containing the compound's metadata. These outputs were fashioned to meet the requirements for publishing the database on the GNPS platform.

^{13}C NMR dereplication: CaraMel workflow

Data acquisition

All samples were analysed using the same acquisition and processing parameters. Fractions were dried under vacuum, and aliquots (up to about 20 mg when possible) of CPC fractions were dissolved in 600 µl DMSO-d_6 and analysed by nuclear magnetic resonance (^{1}H, ^{13}C, HSQC, HMBC, and COSY) at 298 K on a Bruker Avance AVIII-600 spectrometer (Karlsruhe, Germany) equipped with a TCI cryoprobe. ^{13}C NMR spectra were acquired at 150.91 MHz using a standard udeft pulse sequence with an FID acquisition time of 0.36 s, a relaxation delay of 3.00 s, and the accumulation of 1,024 scans.

Data processing

The absolute intensities of all ^{13}C NMR signals detected in all spectra were collected by automatic peak picking, after spectra processing (manually phased, baseline corrected, and referenced by setting the central resonance of DMSO-d_6 at δ 39.80 ppm) using the TOPSPIN 4.1.3 software (Bruker, Rheinstetten, Germany). Each peak list was converted into a text file containing peak positions and absolute peak intensities. Peaks positions from the whole set of spectra were then aligned by an algorithm written in the Python language. The principle was to divide the ^{13}C spectral width (from 0 to 240 ppm) into regular bins of 0.2 ppm width and to place the absolute intensity of each ^{13}C peak into the corresponding bin. The bins without any signal, regardless of fraction, were removed from the bin list. The resulting global table contains 12 columns, each corresponding a CPC fraction, and 403 rows, corresponding to the NMR spectral buckets for which at least one ^{13}C NMR peak was detected in at least one spectrum. This table was imported into the PermutMatrix software (version 1.9.3, LIRMM, Montpellier, France) for hierarchical clustering analysis (HCA). This operation reordered the rows of the global peak table so that similar rows, corresponding to similar chromatographic emergence profiles, were grouped together and lead to define clusters of chemical shifts values, with ideally one cluster defined per compound contained in the EtOAc extract of *Larix decidua* bark. The similarity between table rows was measured by the Euclidian distance, and data agglomeration was performed with the Ward's method. The resulting clusters of ^{13}C NMR chemical shifts were visualised as dendrograms on a heat map.

Each ^{13}C NMR chemical shift cluster obtained from HCA was manually submitted to the VersaDB-generated database containing the structures and corresponding predicted NMR chemical shifts values (nmrshiftdb2) of 2790 natural metabolites found in the Pinaceae family (April 2022), via the structure search engine from ACD/NMR Workbook Suite 2012 (ACD/Labs, Ontario, Canada). This dereplication procedure was described in a previous article.(13) A ^{13}C NMR chemical shift difference between the predicted and experimental spectra was tolerated between 2 and 3 ppm, and the minimum number of ^{13}C query shift values to match was set at about 80% of the number of chemical shifts in the cluster. Finally, each structure proposal provided by the database query and its associated relative configurations was confirmed by interpretation of 1D and 2D NMR data (^{1}H, ^{13}C, HSQC, HMBC, COSY).

LC-HRMS^2 dereplication

Data acquisition

LC analyses were performed with Waters QSM Acquity, equipped with an UPTISPHERE Strategy C18 column (2.2µm x 150 mm x 2.1 µm, Interchim). The eluent consisted of H_2O + 0.1% formic acid (A) and MeCN (B), following a gradient 5-30 % B in 4 min, then 30-80 % B in 14 min, then 80-100 % in 0.50 min, then maintaining 100 % B for 9.5 min at a flow rate of 0.5 ml/min. The wavelength range of the UV detector was set from 210 to 400 nm.

Mass data were acquired with a Waters SYNAPT G2-Si (QTof) mass spectrometer. The electrospray ionisation source was set as follow: positive mode, source temperature 100°C, capillary 3 kV, desolvation temperature 450°C, nebuliser gas flow 5 Bar, desolvation gas flow 700 l/h. MS scans were performed in full-scan mode from m/z 100 to 1200 (scan time 0.1 sec) with a resolution of 40 000 (FWHM). An MS^1 scan was followed by MS^2 scans of the three most intense ions above a threshold of 3000 counts (exclusion window 3 sec). The selected parent ions were fragmented according to the following energy ramp: low mass start: 35 eV, low mass end: 55 eV, high mass start: 70 eV high mass end: 130 eV. Leucine-enkephalin (1 ng/µL) was used as a reference mass *via* a lock spray interface at a flow rate of 10 µl/min for positive ion mode monitoring ($[M + H]^+$ = 556.277).

MZMine 3 data processing

The 12 MS^2 raw files were converted from the .raw (Waters) standard data format to .mzml format using the MSConvert software, part of the ProteoWizard package (version 3.0.21349, Palo Alto, CA). All .mzml were processed by MZMine 3 v0.21 beta(30) in batch mode.

The mass detection was realised by keeping the noise level at 50. The ADAP chromatogram builder was used with a minimum group size of scans of 5, a group intensity threshold of 5000, a minimum highest intensity of 7500, and m/z tolerance of 0.005 (or 20 ppm). The ADAP feature resolver was used for the deconvolution step and the wavelets deconvolution algorithm was applied with the following standard settings: S/N threshold = 5, minimum feature height = 7500, coefficient/area threshold = 20, peak duration range 0.01−1 min, RT wavelet range 0.01−0.12 min. MS^2 scans were paired using an m/z tolerance range of 0.05 Da and RT tolerance range of 0.5 min. Isotopologues were grouped using ^{13}C isotope filter (isotopic peak grouper) algorithm with an m/z tolerance of 0.005 (or 20 ppm) and a retention time (RT) tolerance of 0.2 min and a maximum charge of 3. The peak alignment was performed using the join aligner module [m/z tolerance = 0.004 (or 10 ppm), weight for m/z = 2, weight for RT = 1, absolute RT tolerance 0.5 min]. The peak list was gap-filled with the same RT and m/z range gap filler module [m/z tolerance of 0.004 (or 10 ppm)]. The feature list was filtered using the feature filter algorithm as follow: Duration 0-3 min, data points 3-10000.

Features were then filtered using the row filter algorithm as minimum features in an isotope pattern set to 2.

Finally, the metaCorrelate algorithm was applied to the dataset with the following parameters, RT tolerance 0.1 min, feature height correlation and intensity correlation threshold were set to 0. Min sample in all 1, Min % intensity overlap 40%, Exclude estimated features (gap-filled) turn on. Concerning correlation grouping parameter, min datapoints 5, min data points on edge 2, measure Pearson, min feature shape correlation 85%. Concerning the feature height correlation parameters, min datapoints 3, measure Pearson, min correlation 60%. For the next step, the Ion Identity Networking algorithm was used as follow: m/z tolerance 0.001 (10 ppm), check all features, min-height 0, MS mode positive, maximum charge 2, maximum molecules/cluster 2, adducts: [M+H]+, $[M+Na]^+$, $[M+NH4]^+$, $[M+2H]2^+$, $[M\text{-}H+2Na]^+$, $[M+H+Na]2^+$, modifications: [M-H_2O], [M-2H_2O]. Less common adducts were then added: [M+K]+, $[M+Ca]2^+$, $[M+Fe]2^+$, $[M+H+NH_4]2^+$, $[M+H+K]2^+$, $[M+Ca\text{-}H]^+$, $[M+Fe\text{-}H]^+$. Finally, adducts that tend to form clusters/in-source fragments were added: $[M+H]^+$, $[M+NH_4]^+$, $[M+2H]2^+$, and modifications: [M+HFA], [M+ACN], [M-H_2O], [M-2H_2O], [M-3H_2O], [M-4H_2O]. Ion identity network refinement was applied using the following parameters: minimum size 3, delete smaller networks: link threshold 5. All ion identities were checked with m/z tolerance (MS^2) 0.002 (15 ppm), with checking for multimers and neutral losses ($MS^1 \rightarrow MS^2$).

Eventually, the .mgf preclustered spectral data file and its corresponding .csv metadata files (for RT, areas, and formulas integration and edges MS^1 annotation) were exported using the dedicated "Export for GNPS" built-in options. Processed spectral data were exported to SIRIUS with m/z tolerance of 0.002 (10 ppm) without merging MS/MS.

SIRIUS annotation

SIRIUS 4 is considered a state-of-the-art metabolite annotation solution, which combines molecular formula calculation and the prediction of a molecular fingerprint of a query compound from its fragmentation tree and spectrum.(31) Although the SIRIUS algorithm is integrated into GNPS, it was executed locally with Sirius 4.8.2 GUI, as it allows to use an in-house database for the CSI FingerID contrarily to GNPS. The parameters used to proceed to the spectral data analysis were the following for SIRIUS molecular formula calculation: possible ionisation $[M+H]^+$, $[M+K]^+$, $[M+Na]^+$, instrument: Q-TOF, ppm tolerance 10 ppm, Top molecular formula candidates: 5, filter: formulas from all available DBs, Elements allowed in Molecular Formula: H, C, O, N (max: 3). Zodiac algorithm was used for re-ranking predicted formula as follow: [candidates 300m/z 10, candidates 800 m/z 50, use 2-step approach, Edge Threshold 0.95, min local connections 10, iterations 20 000, burn-in 2000, separates runs 10]. For the CSI: FingerID step, the parameters were the following: possible adducts: $[M+H]^+$, $[M+K]^+$, $[M+Na]^+$, $[M\text{-}H_2O+H]^+$, filter: compounds present in the in-house PINACEAE DB, the maximal number of returned candidate structures: unlimited. Eventually, the CANOPUS algorithm was applied to predict the chemical class of compounds based on NPClassifier chemical ontology.

Ion Identity Molecular Network

The molecular networks are based on matching spectral patterns between paired mass spectra, accomplished by calculating cosine similarity. This metric measures the cosine of the angle between the vectors representing non-zero spectral data points in a multidimensional hyperspace, where the dimensions correspond to the considered mass-to-charge ratio (m/z) variables. The intensity of each data point corresponds to the coordinate value along the respective mass axis within this hyperspace. (26,32)

The molecular networks were created using the online workflow at Global Natural Products Social molecular networking (http://gnps.ucsd.edu). An Ion Identity Molecular Network (IIMN) was created (job: 25b1448341ab454c9002c1767fba98e1) where edges were filtered to have a cosine score above 0.4 and at least 4 matching peaks. Further edges between two nodes were kept in the network if and only if each of the nodes appeared in each other's respective top 10 most similar nodes. All matches kept between the network spectra and the library spectra were required to have a score above 0.7 and at least 4 matched peaks. The default speclibs from the GNPS platform were used for the spectral library search. A Feature Based Molecular Network was realised (job: 0f64e360d2ba4227beb6a2e0a03a5335) to decrease the library search cosine-score threshold from 0.7 to 0.4, with at least 4 matching peaks. The default speclibs from the GNPS platform were used for the spectral library search. The FBMN results were used for Network Annotation Propagation (NAP) step (job: 35984a76a5794491acd88cb694e30843). The parameters used to proceed with analysis were the following: N first candidates for consensus score 10, Accuracy for exact mass candidate search (ppm) 20, acquisition mode Positive, Multiple adduct types $[M+H]^+$, $[M+Na]^+$, $[M\text{-}H_2O+H]^+$, user-provided database: NAP formatted in house Pinaceae structural database. The maximum number of candidate structures in the graph was set to 3.

Merging and comparing SIRIUS annotations with IIMN, NAP

IIMN, FBMN, NAP, and MolNetEnhancer results were downloaded from the GNPS platform as archive folders. Metadata from the different GNPS annotation workflows were merged *via* Cytoscape 3.8.2 software (https://cytoscape.org/), following the method explained in Fig. S3. The IIMN graph was kept visualising the molecular network. Quantification table and the "df_resume_confidence.tsv" file resulting from the CATHEDRAL comparison data were overlayed on the molecular network using custom style.

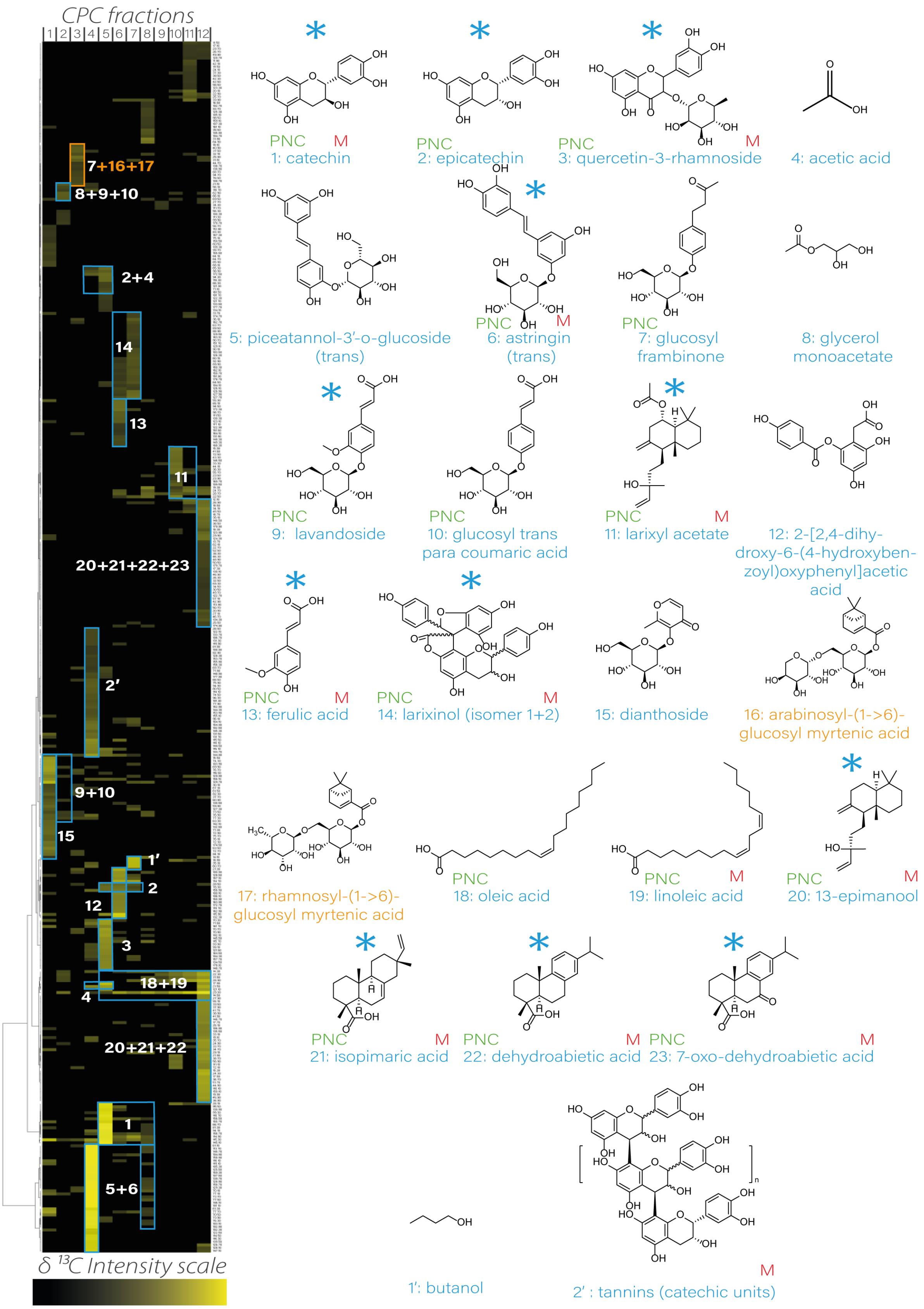


Figure 1: Annotated heatmap, built with ^{13}C NMR chemical shifts of the twelve fractions from EtOAc *L. decidua* extract.
(M) Compound equally annotated through MS2 data, orange compounds: previously undescribed compounds, (PNC) Compound structure described in the ^{13}C NRM predicted Pinaceae database, (Blue star) Compound directly annotated by Caramel workflow

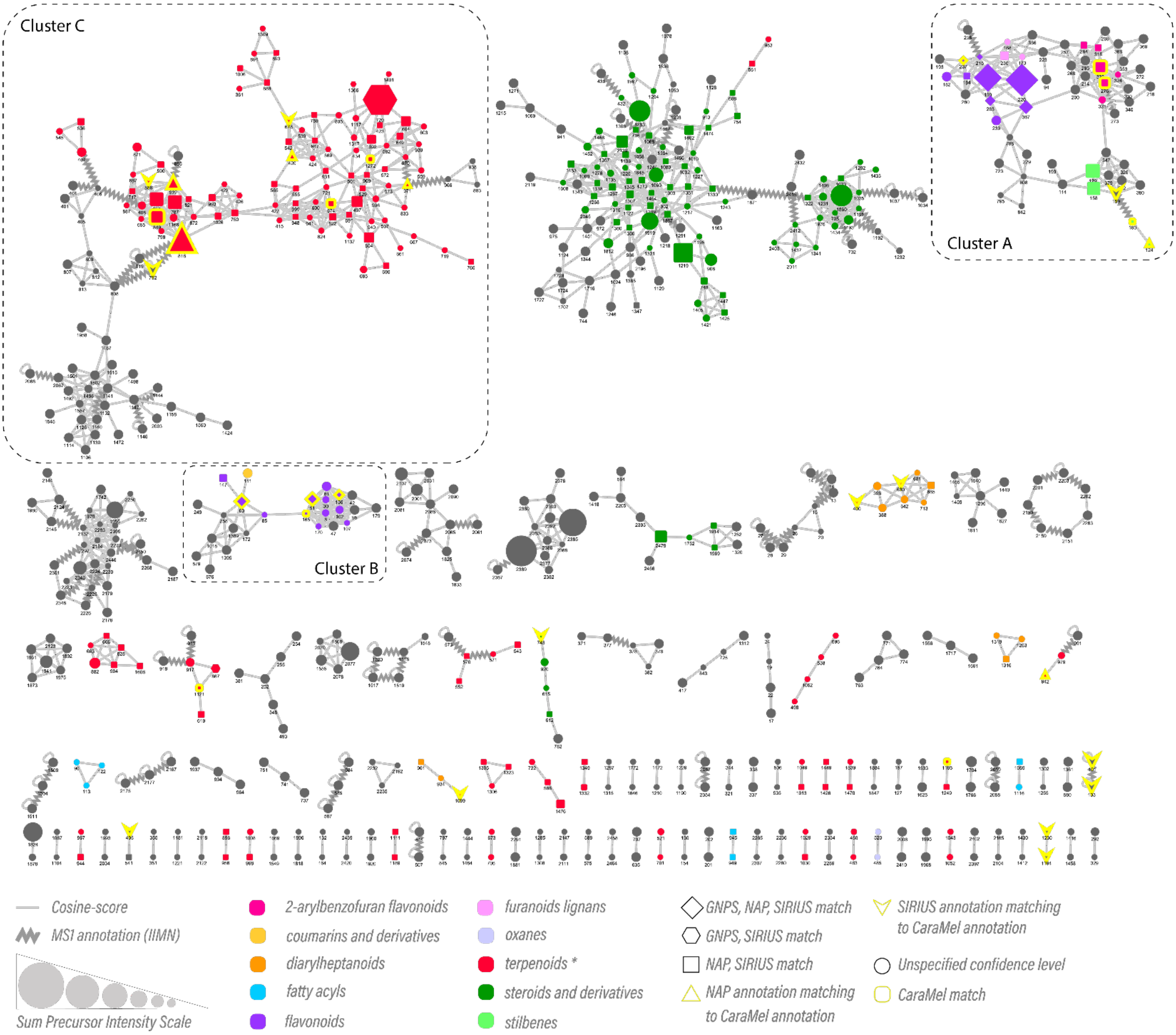


Figure 2: IIMN visualization with CATHEDRAL comparison outcomes overlayed

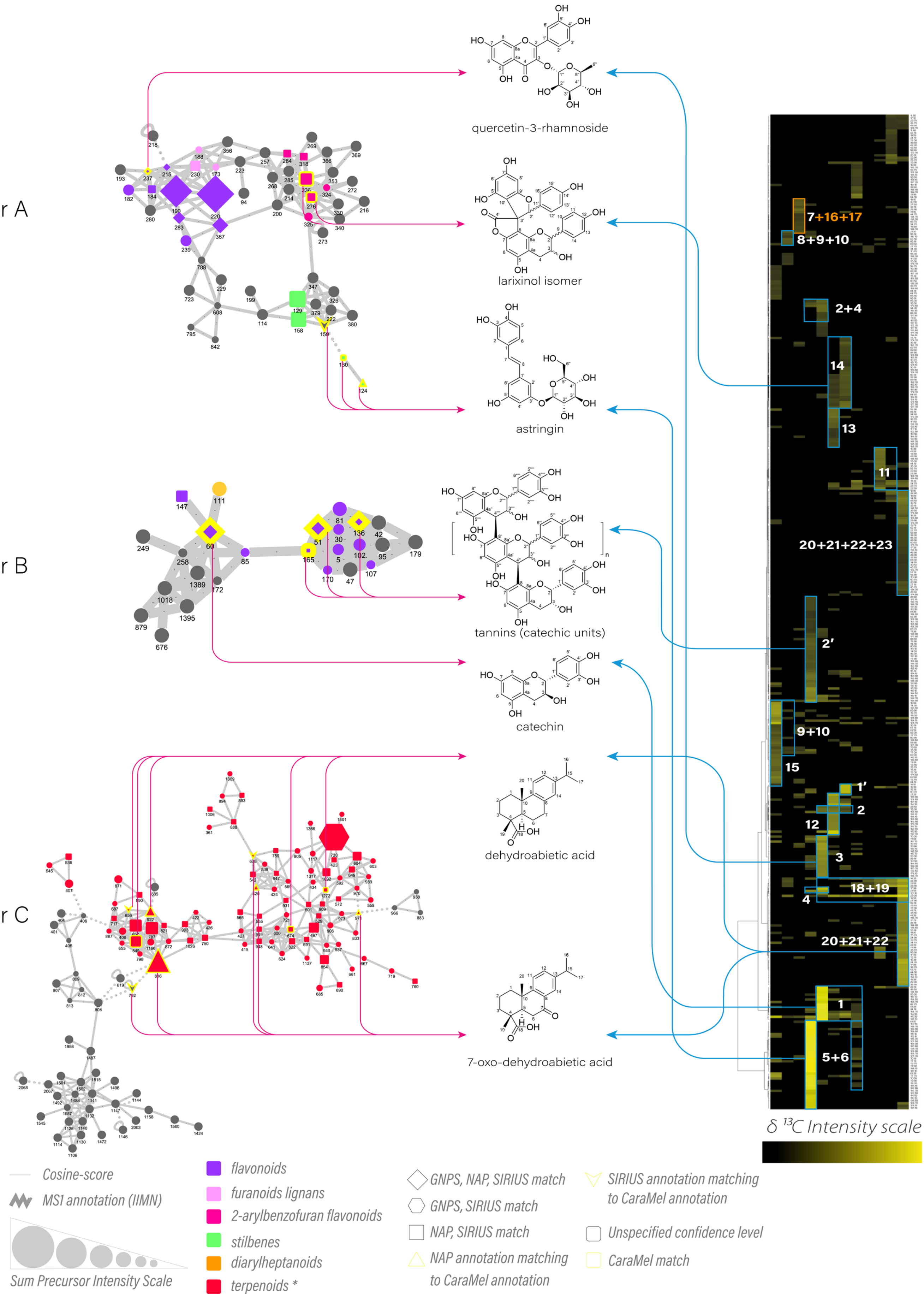


Figure 3: Comparison of MS/MS clusters A to C to the annotated CaraMel heatmap.

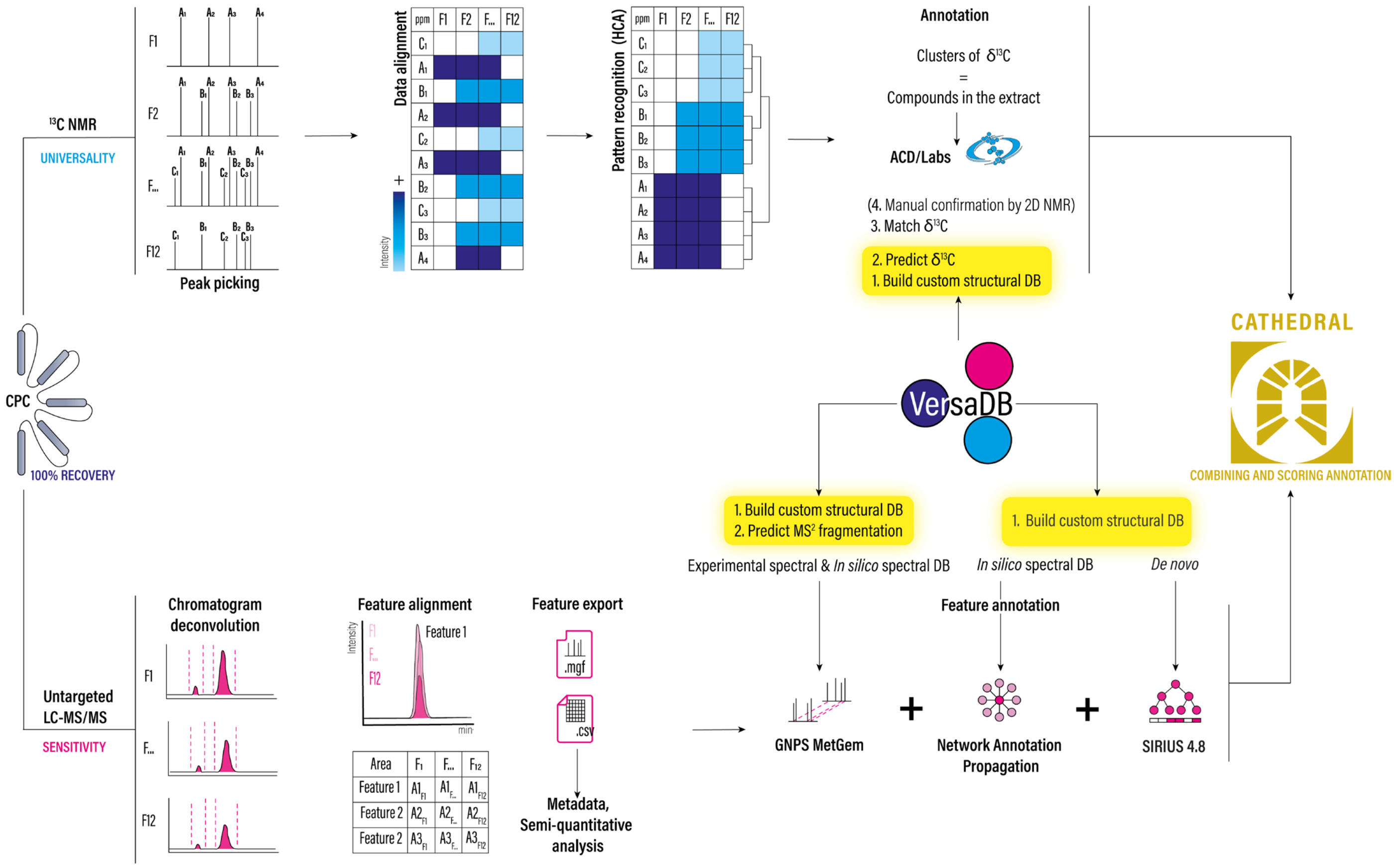


Figure 4: Dereplication workflows based on ^{13}C NMR (CaraMel) and LC-HRMS^2 of *Larix decidua* EtOAc crude extract by Centrifugal Partition Chromatography

CPC fractions	Mass (mg)	% Crude extract	Composition
01 – elution	42	10.8 %	dianthoside (Maj – cluster 15)
02 – elution	12	3.1 %	glucosyl-frambinone (Min – cluster 7); glycerol-monoacetate (Maj – cluster 8); lavandoside (Min – cluster 9); glucosyl *trans-para*-coumaric acid (Min – cluster 10); rhamnosyl-(1->6)-glucosyl-Myrtenic acid (Min – cluster 17)
03 – elution	18	4.7 %	glucosyl-frambinone (Med – cluster 7); arabino-syl-(1->6)-glucosyl-myrtenic acid (Maj – cluster 16); rhamnosyl-(1->6)-glucosyl-myrtenic acid (Maj – cluster 17)
04 – elution	85	22.0 %	acetic acid (Min – cluster 4); astringin (trans) (Maj – cluster 6); piceatannol-3'-O-glucoside (Maj – cluster 5); tannins (catechin unit) (Med – cluster 2')
05 – elution	29	7.5 %	catechin (Maj – cluster 1); epicatechin (Min – cluster 2); quercetin-3-rhamnoside (Med – clus-ter 3); acetic acid (Min – cluster 4); oleic acid (Min – cluster 18); linoleic acid (Min – cluster 19)
06 – elution	26	6.7 %	epicatechin (Min – cluster 2); acetic acid (Min – cluster 4); 2-[2,4-dihydroxy-6-(4-hydroxybenzoyl)oxyphenyl]acetic acid (Maj - cluster 12); ferulic acid (Med – cluster 13); lar-ixinol isomer 1 (Min – cluster 14); quercetin-3-rhamnoside (Min – cluster 3); oleic acid (Min – cluster 18); linoleic acid (Min – cluster 19)
07 – elution	38	9.8 %	larixinol isomer 1 (Maj - cluster 14); larixinol isomer 2 (Med – cluster 14); acetic acid (Min – cluster 4); oleic acid (Min – cluster 18); linoleic acid (Min – cluster 19)
08 – elution	16	4.1 %	catechin (Med – cluster 1); astringin (*trans*) (Min – cluster 6); piceatannol-3'-O-glucoside (Min – cluster 5); glycerol-monoacetate (Min – cluster 8); ferulic acid (Med – cluster 13); oleic acid (Min – cluster 18); linoleic acid (Min – cluster 19)
09 – elution	8	2.1 %	oleic acid (Med – cluster 18); linoleic acid (Min – cluster 19); 7-oxodehydroabietic acid (Min – cluster 23)
10 – elution	17	4.4 %	glycerol-monoacetate (Min – cluster 8); larixyl acetate (Maj – cluster 11); oleic acid (Min – clus-ter 18); linoleic acid (Min – cluster 19); 7-oxodehydroabietic acid (Min – cluster 23)
11 – extrusion	83	21.4 %	larixyl acetate (Min – cluster 11); oleic acid (Med – cluster 18); linoleic acid (Med – cluster 19); dehydroabietic acid (Min – cluster 22)
12 – extrusion	13	3.4 %	glycerol-monoacetate (Min – cluster 8); oleic acid (Med – cluster 18); linoleic acid (Min – clus-ter 19); 13-epimanool (Maj – cluster 20); isopi-maric acid (Maj – cluster 21); dehydroabietic acid (Med – cluster 22)

Table 1: Mass and global composition of the CPC fractions. (Maj=major; Med=medium; Min=minor). Compounds are related to their corre-sponding cluster on Fig.1.

Feature jointly annotated with	Confidence level
FBMN, NAP, SIRIUS, CaraMel	1
FBMN, SIRIUS, CaraMel	2
NAP, SIRIUS, CaraMel (+ other FBMN annotation)	3+
NAP, SIRIUS, CaraMel (no other FBMN annotation)	3
FBMN, NAP, CaraMel	4
FBMN, CaraMel	5
SIRIUS, CaraMel	6
NAP (one structure candidate of the *in silico* fragmentation search with MetFrag), CaraMel	7
CaraMel	8
FBMN, NAP, SIRIUS	9
FBMN, SIRIUS	10
NAP, SIRIUS (+ other FBMN annotation)	11+
NAP, SIRIUS (no other FBMN annotation)	11
FBMN, NAP	12

Table 2: **Custom confidence levels established according to the tools for which the feature was jointly annotated with.**

Time (min)	Flow rate (ml / min)	% Lower phase system 1	% Lower phase sytem2
0	1	100 %	0 %
3	10	100 %	0 %
15	10	100 %	0 %
90	10	0 %	100 %
110	10	0 %	100 %
125	Extrusion (20 ml / min)	Extrusion	Extrusion

Table 3: Mobile phase composition during CPC 1 fractionation.

ChemRxiv®

Improving the chemical profiling of complex natural extracts by joint ^{13}C NMR and LC-HRMS^2 analysis and the querying of *in silico* generated chemical

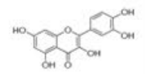
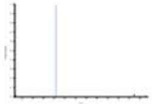
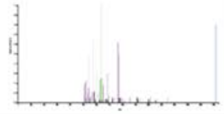
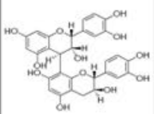
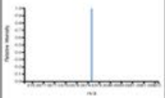
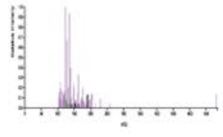
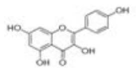
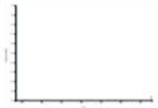
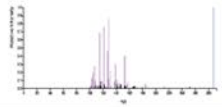
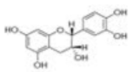
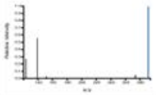
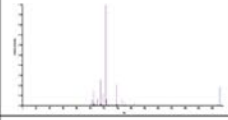
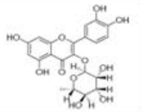
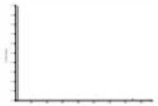
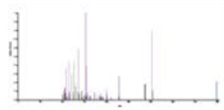
Julien Cordonnier,[a,b] Simon Remy,[b*] Alexis Kotland,[d] Ritchy Leroy,[b] Pierre Darme,[a,b], Benjamin Bertaux,[b] Charlotte Sayagh,[b] Agathe Martinez,[b] Nicolas Borie,[b] Jane Hubert,[d] Dominique Aubert,[a,c] Isabelle Villena,[a,c] Jean-Marc Nuzillard,[b] Jean-Hugues Renault[b*]

[a]University of Reims Champagne Ardenne, ESCAPE EA7510, 51097 Reims, France
[b] University of Reims Champagne Ardenne, CNRS, ICMR 7312, 51097 Reims, France
[c] University of Reims Champagne Ardenne, CRB National reference Centre on Toxoplasmosis, 51097 Reims, France
[d] NatExplore, 51140 Prouilly, France
*Correspondence should be addressed to S.R. (simon.remy@univ-reims.fr)

Table of contents

Confidence Level	FBMN	NAP	SIRIUS	CararMel	Nb Features	Nb unique compound	Compound(s) name	Cumulative nb compounds	Graphical shape	Related document
1	X	X	X	X	4	3	catechin / epicatechin quercetin-3-rhamnoside catechic tanin	3	Yellow diamond	Table. S2
2	X		X	X	NaN	NaN	NaN	3	NaN	NaN
3+		X	X	X	1	1	dehydroabietic acid	4	Yellow rectangle	Table. S3
3		X	X	X	6	3	dehydroabietic acid larixinol isomers 7-oxodehydroabietic acid	6	Yellow rectangle	Table. S4
4	X	X		X	NaN	NaN	NaN	6	NaN	NaN
5	X			X	NaN	Nan	NaN	6	NaN	NaN
6			X	X	42	9	catechin / epicatechin linoleic acid dehydroabietic acid ferulic acid isopimaric acid 7-oxo-dehydroabietic acid astringin 13-epimanool larixyl acetate	12	V shape	Table. S5
7		X		X	4	3	astringin 7-oxodehydroabietic acid dehydroabietic acid	12	Triangle	Table. S6
8				X	NaN	11	acetic acid piceatannol-3'-o-glucoside glucosyl frambinone glycerol monoacetate lavandoside glucosyl trans para coumaric acid 2-[2,4-digydroxy-6-(4-hydroxybenzoyl)oxyphenyl]acet ic acid dianthoside arabinosyl-(1->6)-glucosyl myrtenic acid rhamnosyl-(1->6)-glucosyl myrtenic acid oleic acid	23	Yellow higlighted	Table. 1
9	X	X	X		4	2	quercetin kaempferol	25	Diamond	Table. S2
10	X		X		12	5	danielic acid rutin kaempferol quercetin-3-galactoside quercetin	28	Hexagon	Table. S7
11+		X	X		15	5 features with a unique MetFrag candidate ⊠ 3 unique compounds	1,4a-dimethyl-7-(prop-1-en-2-yl)-1,2,3,4,4a,9,10,10 a-octahydrophenanthrene-1-carboxylic acid 19-nor-4-hydroxyabieta-8,11,13-trien-7-one 18-nor-4,15-dihydroxyabieta-8,11,13-trien-7-one	31	Rectangle	Table. S3
11		X	X		103	41 features with a unique MetFrag candidate ⊠ 21 unique compounds	piceatannol cycloeucalenone picealactone A 13-hydroxy-8,11,13-podocarpatrien-18-oic acid meridinol octabenzone methyl dehydroabietate 19-nor-4-hydroxyabieta-8,11,13-trien-7-one (9beta,23R)-23-Hydroxy-3-oxo-5alpha-lanosta-7,24-di en-26-oic acid lactone methyl 3-[5-(2-hydroxypropan-2-yl)-1,4',11,13-tetra methyl-5'-oxospiro[16-oxapentacyclo[9.7.0.02,8.06,8 .012,17]octadecane-15,2'-oxolane]-6-yl]propanoate 4-(3-methoxy-4,4,10,13,14-pentamethyl-2,3,5,6,7,8,1 2,15,16,17-decahydro-1H-cyclopenta[a]phenanthren-17 -yl)pentanoic acid 1,4a-dimethyl-7-(prop-1-en-2-yl)-1,2,3,4,4a,9,10,10 a-octahydrophenanthrene-1-carboxylic acid meso-3,4-divanillyltetrahydrofuran epiafzelechin abiesadine Q dehydroabietan piceanolol B (2E,5R,9S,10R,13R,14R,17R)-2-ethylidene-4,4,10,13,1 4-pentamethyl-17-[(2R)-4-oxopentan-2-yl]-5,6,9,11,1 2,15,16,17-octahydro-1H-cyclopenta[a]phenanthren-3- one (r)-1,7-diphenyl-3-heptanol 18-nor-4,15-dihydroxyabieta-8,11,13-trien-7-one (1S,4aS,10aS)-1,4a-dimethyl-7-propan-2-yl-2,3,4,9,1 0,10a-hexahydrophenanthren-1-ol	52	Rectangle	Table.S3
12	X	X			NaN	NaN	NaN	52	NaN	NaN

Table S1: Summary of compounds annotated during chemical profiling of *Larix decidua bark* crude extract, after the comparison of the candidate coming from each annotation tool (*c.f. Excel file Table_S1.xlsx*).

Feature ID	Structure	Name	Mol. Weight	MS1 spectrum	GNPS spectrum ID	FBMN matched mirror spectra	NAP - MetFrag candidate (LOTUS ID)	MS/MS spectrum	SIRIUS annotation	NPClassifier Superclass	CaraMel
190 220		quercetin	302.24		CCMSLIB00005739139	https://gnps.ucsd.edu/ProteoSAFe/result.jsp?task=0f64e360d2ba4227beb6a2e0a03a5335&view=view_all_annotations_DB	LTS0004651		REFJWTPEDVJJIY C15H10O7[M + H]+	Flavonoids	NO
51 136		catechic tannin	578.52		CCMSLIB00005742589	https://gnps.ucsd.edu/ProteoSAFe/result.jsp?task=0f64e360d2ba4227beb6a2e0a03a5335&view=view_all_annotations_DB	LTS0066122		XFZJEEAOWLFHDH C30H26O12[M + H]+	Flavonoids	YES
283 367		kaempferol	286.24		CCMSLIB00005748053	https://gnps.ucsd.edu/ProteoSAFe/result.jsp?task=0f64e360d2ba4227beb6a2e0a03a5335&view=view_all_annotations_DB	LTS0155822		IYRMWMYZSQPJKC C15H10O6[M + H]+	Flavonoids	NO
60		catechin	290.27		CCMSLIB00005742701	https://gnps.ucsd.edu/ProteoSAFe/result.jsp?task=0f64e360d2ba4227beb6a2e0a03a5335&view=view_all_annotations_DB	LTS0265245		PFTAWBLQPZVEMU C15H14O6[M + H]+	Flavonoids	YES
237		quercetin-3-rhamnoside	448.38		CCMSLIB00000085852	https://gnps.ucsd.edu/ProteoSAFe/result.jsp?task=0f64e360d2ba4227beb6a2e0a03a5335&view=view_all_annotations_DB	LTS0186298		OXGUCUVFOIWWQJ C21H20O11[M + H]+	Flavonoids	YES

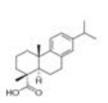
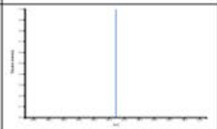
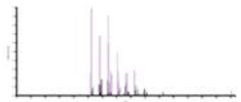
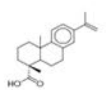
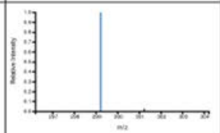
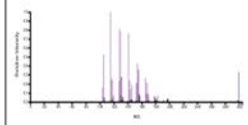
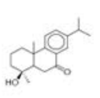
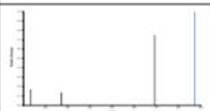
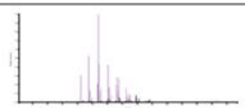
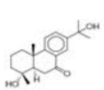
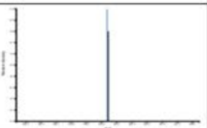
Feature ID	Structure	Name	Mol. Weight	MS1 spectrum	GNPS spectrum ID	FBMN matched mirror spectra	NAP - MetFrag candidate (LOTUS ID)	MS/MS spectrum	SIRIUS annotation	NPClassifier Superclass	CaraMel
1121		dehydroabietic-acid	300.44		Other	Other	LTS0034017		NFWKVWVWBFBAOV C20H28O2[M + H]+	Diterpenoid	YES
497 750 854		1,4a-dimethyl-7-(prop-1-en-2-yl)-1,2,3,4,4a,9,10,10a-octahydrophenanthrene-1-carboxylic acid	298.42		Other	Other	LTS0088733		NGBRPGLXCQJIPU C20H26O2[M + H]+	Diterpenoid	NO
822		19-nor-4-hydroxyabieta-8,11,13-trien-7-one	286.41		Other	Other	LTS0108097		PTQFIYQNKVSVGM C19H26O2[M + H]+	Diterpenoid	NO
893		18-nor-4,15-dihydroxyabieta-8,11,13-trien-7-one	302.41		Other	Other	LTS0171116		SGCHZBKQDFNHSL C19H26O3[M + Na]+	Diterpenoid	NO

Table S2: Summary of compounds jointly annotated by all annotation tools (*c.f. Excel file Table_S2.xlsx*).

Table S3: Summary of compounds jointly annotated by NAP, SIRIUS and eventually CaraMel, with another FBMN candidate (*c.f. Excel file Table_S3.xlsx*).

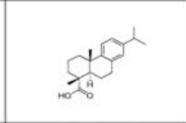
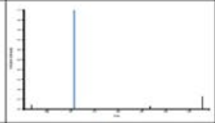
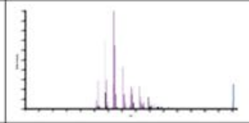
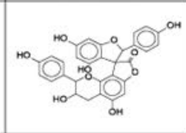
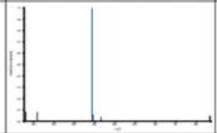
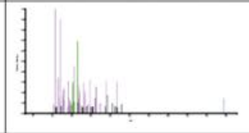
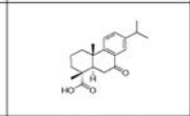
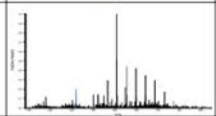
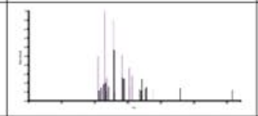
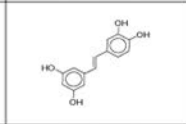
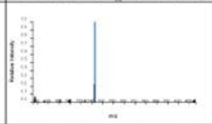
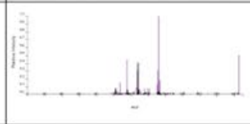
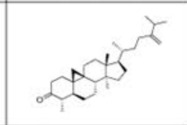
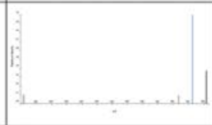
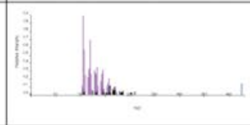
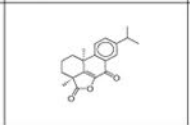
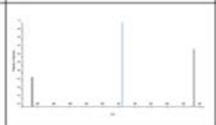
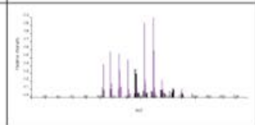
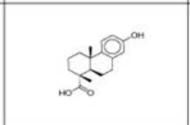
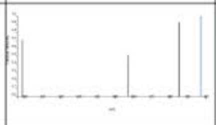
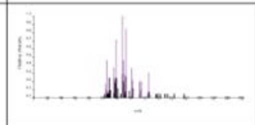
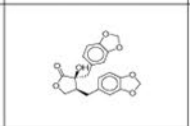
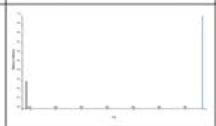
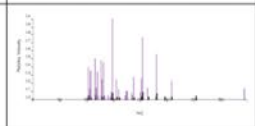
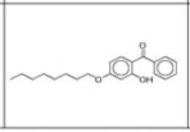
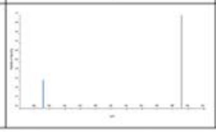
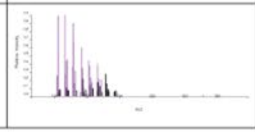
Feature ID	Structure	Name	Mol. Weight	MS1 spectrum	GNPS spectrum ID	FBMN matched mirror spectra	NAP - MetFrag candidate (LOTUS ID)	MS/MS spectrum	SIRIUS annotation	NPClassifier Superclass	CaraMel
674 849 1272		dehydroabietic-acid	300.44		Other	Other	LTS0034017		NFWKVWVWBFBAOV C20H28O2[M + H]+	Diterpenoid	YES
276 336		larixinol isomers	542.49		Other	Other	LTS0269975		RNDNBGULZNCSNB C30H22O10[M + H]+	Flavonoid	YES
1185		7-oxo-dehydroabietic acid	314.42		Other	Other	LTS0229171		MSWJSDLNPCSSNW C20H26O3[M + H]+	Diterpenoid	YES
129 158		piceatannol	244.24		Other	Other	LTS0094372		CDRPUGZCRXZLFL C14H12O4[M + H]+	Stilbenoid	NO
1267		cycloeucalenone	424.70		Other	Other	LTS0224467		NFRXSIOHGADFRG C30H48O[M + H]+	Triterpenoid	NO
888		picealactone A	310.39		Other	Other	LTS0079469		UKOKENMLXFNPEJ C20H22O3[M + H]+	Diterpenoid	NO
541		13-hydroxy-8,11,13-podocarpatrien-18-oic acid	274.36		Other	Other	LTS0083733		DWHTYLMRWXUGJL C17H22O3[M + Na]+	Diterpenoid	NO
230		meridinol	370.35		Other	Other	LTS0044016		OTWLSQPCSOEBAY C20H18O7[M + Na]+	Lignan	NO
529		octabenzone	326.43		Other	Other	LTS0194549		QUAMTGJKVDWJEQ C21H26O3[M + H]+	Phloroglucinol	NO

Table S4 (Part 1.): Summary of compounds jointly annotated by NAP, SIRIUS and eventually CaraMel, without any FBMN candidate (*c.f. Excel file Table_S4.xlsx*).

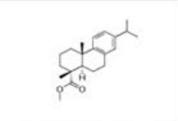
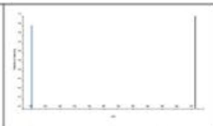
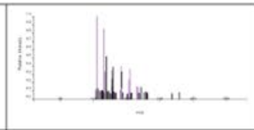
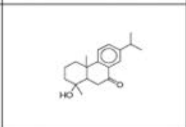
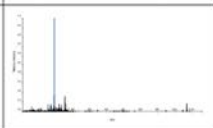
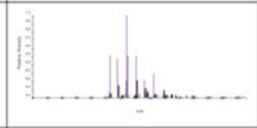
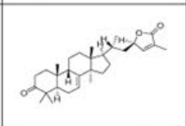
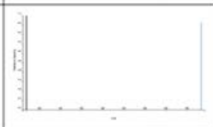
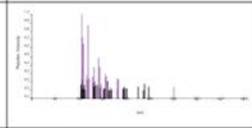
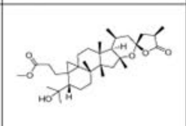
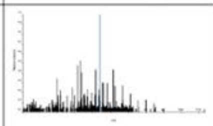
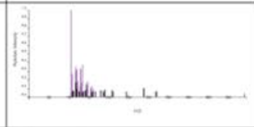
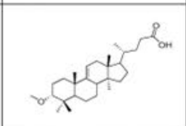
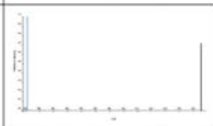
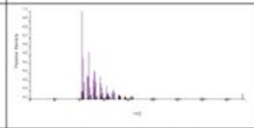
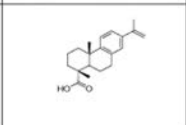
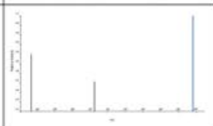
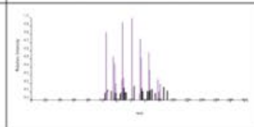
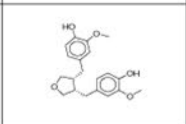
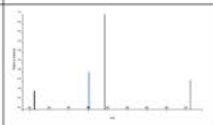
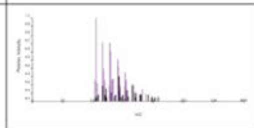
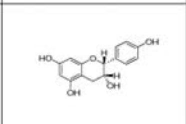
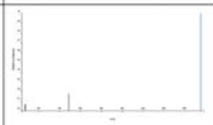
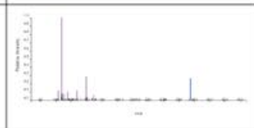
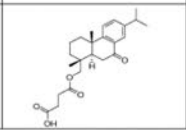
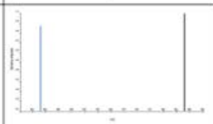
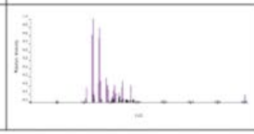
1249		methyl dehydroabietate	314.46		Other	Other	LTS0080951		PGZCJOPTDHWYES $C_{21}H_{30}O_{2}$[M + H]+	Diterpenoid	NO
590 657 821		19-nor-4-hydroxyabieta-8,11,13-trien-7-one	286.41		Other	Other	LTS0108097		PTQFIYQNKVSVGM $C_{19}H_{26}O_{2}$[M + H]+	Diterpenoid	NO
716 1136 1380		(9beta,23R)-23-Hydroxy-3-oxo-5alpha-lanosta-7,24-dien-26-oic acid lactone	452.67		Other	Other	LTS0013015		JWBQJSCILWCBJO $C_{30}H_{44}O_{3}$[M + H]+	Triterpenoid	NO
1013		methyl 3-[5-(2-hydroxypropan-2-yl)-1,4',11,13-tetramethyl-5'-oxospiro[16-oxapentacyclo[9.7.0.02,8.06,8.012,17]octadecane-15,2'-oxolane]-6-yl]propanoate	516.71		Other	Other	LTS0049703		YBQWVVDWBNHYPN $C_{31}H_{48}O_{6}$[M + Na]+	Triterpenoid	NO
1002 1080 1821		4-(3-methoxy-4,4,10,13,14-pentamethyl-2,3,5,6,7,8,12,15,16,17-decahydro-1H-cyclopenta[a]phenanthren-17-yl)pentanoic acid	430.66		Other	Other	LTS0002592		SIYBXEYFWKPLJC $C_{28}H_{46}O_{3}$[M + H]+	Triterpenoid	NO
619 754 993 1026		1,4a-dimethyl-7-(prop-1-en-2-yl)-1,2,3,4,4a,9,10,10a-octahydrophenanthrene-1-carboxylic acid	298.42		Other	Other	LTS0088733		NGBRPGLXCQJIPU $C_{20}H_{26}O_{2}$[M + H]+	Diterpenoid	NO
572		meso-3,4-divanillyltetrahydrofuran	344.40		Other	Other	LTS0255421		ROGUIJKVZZROIQ $C_{20}H_{24}O_{5}$[M + H]+	Lignan	NO
147		epiafzelechin	274.27		Other	Other	LTS0203074		RSYUFYQTACJFML $C_{15}H_{14}O_{5}$[M + H]+	Flavonoid	NO
873		abiesadine Q	400.51		Other	Other	LTS0092634		ZSHVSBCOLVBOLB $C_{24}H_{32}O_{5}$[M + H]+	Diterpenoid	NO

Table S4 (Part 2.): Summary of compounds jointly annotated by NAP, SIRIUS and eventually CaraMel, without any FBMN candidate (*c.f. Excel file Table_S4.xlsx*).

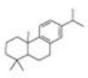
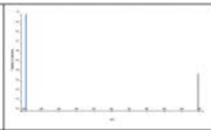
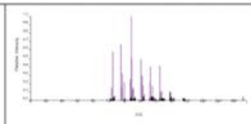
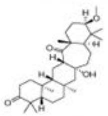
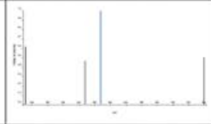
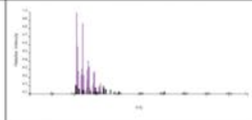
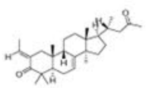
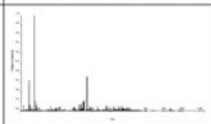
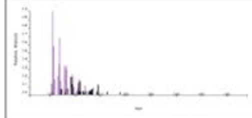
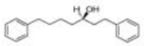
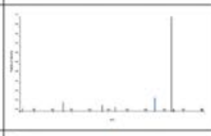
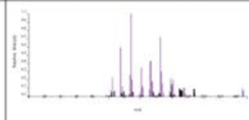
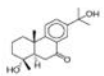
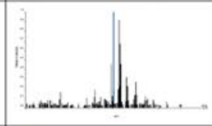
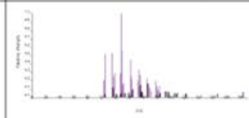
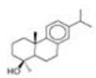
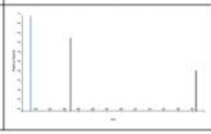
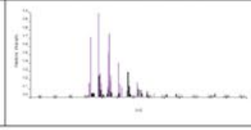
690 828 904 998 1219 1425 1447		dehydroabietan	270.45		Other	Other	LTS0210076		QUUCYKKMFLJLFS C20H30[M + H]+	Diterpenoid	NO
1347		piceanolol B	486.73		Other	Other	LTS0041481		KUMASSYJGMPIHK C31H50O4[M + H]+	Triterpenoid	NO
1356		(2E,5R,9S,10R,13R,14R,17R)-2-ethylidene-4,4,10,13,14-pentamethyl-17-[(2R)-4-oxopentan-2-yl]-5,6,9,11,12,15,16,17-octahydro-1H-cyclopenta[a]phenanthren-3-one	424.66		Other	Other	LTS0087433		SKHNEXWRHKHJFZ C29H44O2[M + H]+	Triterpenoid	NO
665 961 1316		(r)-1,7-diphenyl-3-heptanol	268.39		Other	Other	LTS0155650		FITFAFMCWGCVDQ C19H24O[M + H]+	Lignan	NO
423 565 616		18-nor-4,15-dihydroxyabieta-8,11,13-trien-7-one	302.41		Other	Other	LTS0171116		SGCHZBKQDFNHSL C19H26O3[M + H]+	Diterpenoid	NO
949		(1S,4aS,10aS)-1,4a-dimethyl-7-propan-2-yl-2,3,4,9,10,10a-hexahydrophenanthren-1-ol	272.43		Other	Other	LTS0118921		SOJWLJKPIIODOH C19H28O[M + H]+	Diterpenoid	NO

Table S4 (Part 3.): Summary of compounds jointly annotated by NAP, SIRIUS and eventually CaraMel, without any FBMN candidate (*c.f. Excel file Table_S4.xlsx*).

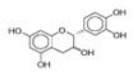
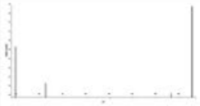
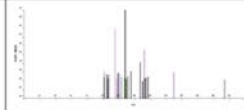
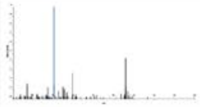
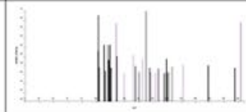
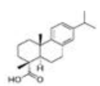
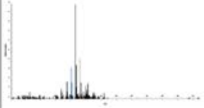
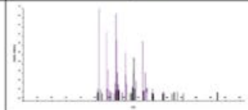
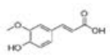
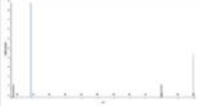
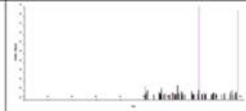
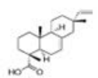
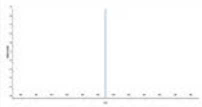
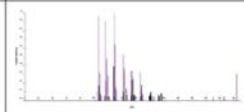
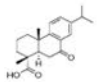
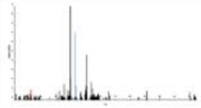
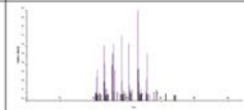
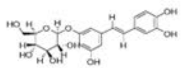
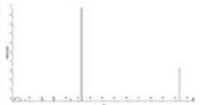
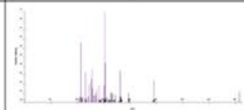
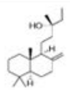
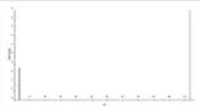
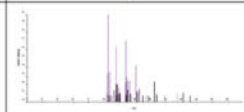
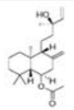
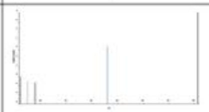
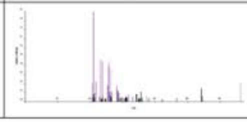
Feature ID	Structure	Name	Mol. Weight	MS1 spectrum	GNPS spectrum ID	FBMN matched mirror spectra	NAP - MetFrag candidate (LOTUS ID)	MS/MS spectrum	SIRIUS annotation	NPClassifier Superclass	CaraMel
76 250		catechin / epicatechin	290.27		NaN	NaN	NaN		PFTAWBLQPZVEMU C15H14O6[M - H2O + H]+	Flavonoid	YES
209 958 959 960		linoleic acid	280.45		NaN	NaN	NaN		OYHQOLUKZRVURQ C18H32O2[M + Na]+	Fatty Acid	YES
436 743 856 935 1184 1253 2014		dehydroabietic acid	300.44		NaN	NaN	NaN		NFWKVWVWBFBAOV C20H28O2[M + H]+	Diterpenoid	YES
57 196 700 1036 1194 1230 1328 1617 1910		feurlic acid	194.18		NaN	NaN	NaN		KSEBMYQBYZTDHS C10H10O4[M - H2O + H]+	Phenylpropanoid (C6-C3)	YES
630 978 1073 1099 1160		isopimaric acid	302.45		NaN	NaN	NaN		MXYATHGRPJZBNA C20H30O2[M - H4O2 + H]+	Diterpenoid	YES
386 400 460 461 494 495 583 678 1279 1340		7-oxo-dehydroabietic acid	314.42		NaN	NaN	NaN		MSWJSDLNPCSSNW C20H26O3[M + H]+	Diterpenoid	YES
159		*trans* astringin	406.38		NaN	NaN	NaN		PERPNFLGJXUDDW C20H22O9[M + H]+	Stilbenoid	YES
1047		13-epimanool	290.48		NaN	NaN	NaN		CECREIRZLPLYDM C20H34O[M - H2O + H]+	Diterpenoid	YES
1182		larixyl acetate	348.52		NaN	NaN	NaN		WCNCDVQMEQJFGH C22H36O3[M - H2O + H]+	Diterpenoid	YES

Table S5: Summary of compounds jointly annotated by SIRIUS and CaraMel (*c.f. Excel file Table_S5.xlsx*).

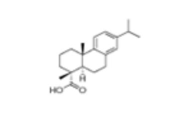
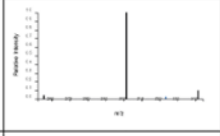
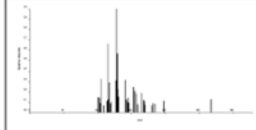
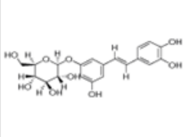
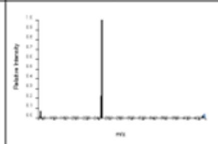
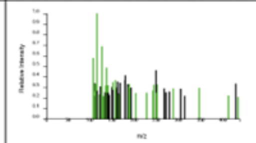
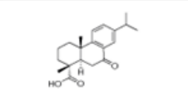
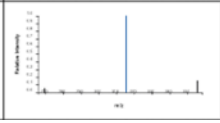
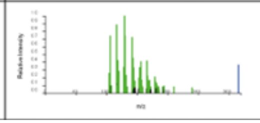
Feature ID	Structure	Name	Mol. Weight	MS1 spectrum	GNPS spectrum ID	FBMN matched mirror spectra	NAP - MetFrag candidate (LOTUS ID)	MS/MS spectrum	SIRIUS annotation	NPClassifier Superclass	CaraMel
942		dehydroabietic acid	300.44		NaN	NaN	LTS0251851		NaN	Diterpenoid	YES
124		*trans* astringin	406.38		NaN	NaN	LTS0251851		NaN	Stilbenoid	YES
420 816		7-oxo-dehydroabietic acid	314.42		NaN	NaN	LTS0229171		NaN	Diterpenoid	YES

Table S6: Summary of compounds jointly annotated by NAP and CaraMel (*c.f. Excel file Table_S6.xlsx*).

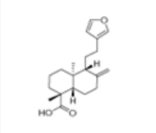
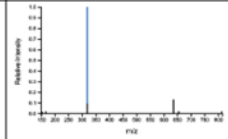
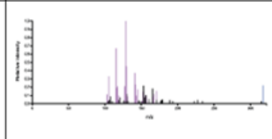
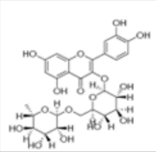
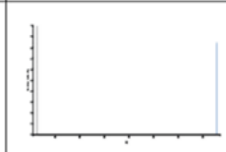
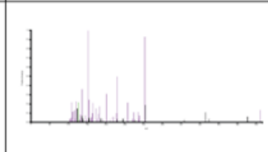
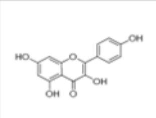
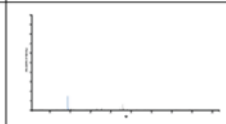
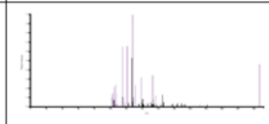
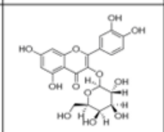
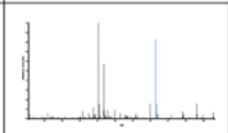
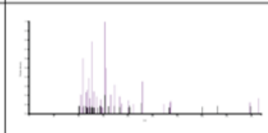
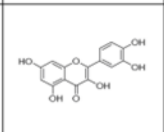
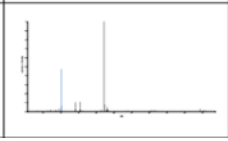
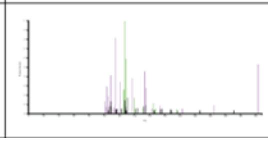
Feature ID	Structure	Name	Mol. Weight	MS1 spectrum	GNPS spectrum ID	FBMN matched mirror spectra	NAP - MetFrag candidate (LOTUS ID)	MS/MS spectrum	SIRIUS annotation	NPClassifier Superclass	CaraMel
334 422 426 488 569 720 867 1366		daniellic acid	316.44		CCMSLIB00006570661 CCMSLIB00006570644	https://gnps.ucsd.edu/ProteoSAFe/result.jsp?task=0f64e360d2ba4227beb6a2e0a03a5335&view=view_all_annotations_DB	NaN		ZQHJXKYYELWEOK C20H28O3[M + H]+	Diterpenoid	NO
182		rutin	610.52		CCMSLIB00000222082	https://gnps.ucsd.edu/ProteoSAFe/result.jsp?task=0f64e360d2ba4227beb6a2e0a03a5335&view=view_all_annotations_DB	NaN		IKGXIBQEEMLURG C27H30O16[M + H]+	Flavonoid	NO
239		kaempferol	286.24		CCMSLIB00005749366	https://gnps.ucsd.edu/ProteoSAFe/result.jsp?task=0f64e360d2ba4227beb6a2e0a03a5335&view=view_all_annotations_DB	NaN		IYRMWMYZSQPJKC C15H10O6[M + H]+	Flavonoid	NO
280		quercetin-3-galactoside	464.38		CCMSLIB00005739276	https://gnps.ucsd.edu/ProteoSAFe/result.jsp?task=0f64e360d2ba4227beb6a2e0a03a5335&view=view_all_annotations_DB	NaN		OVSQVDMCBVZWGM C21H20O12[M + H]+	Flavonoid	NO
343		quercetin	302.24		CCMSLIB00005745873	https://gnps.ucsd.edu/ProteoSAFe/result.jsp?task=0f64e360d2ba4227beb6a2e0a03a5335&view=view_all_annotations_DB	NaN		REFJWTPEDVJJIY C15H10O7[M + H]+	Flavonoid	NO

Table S7: Summary of compounds jointly annotated by FBMN and SIRIUS (*c.f. Excel file Table_S7.xlsx*).

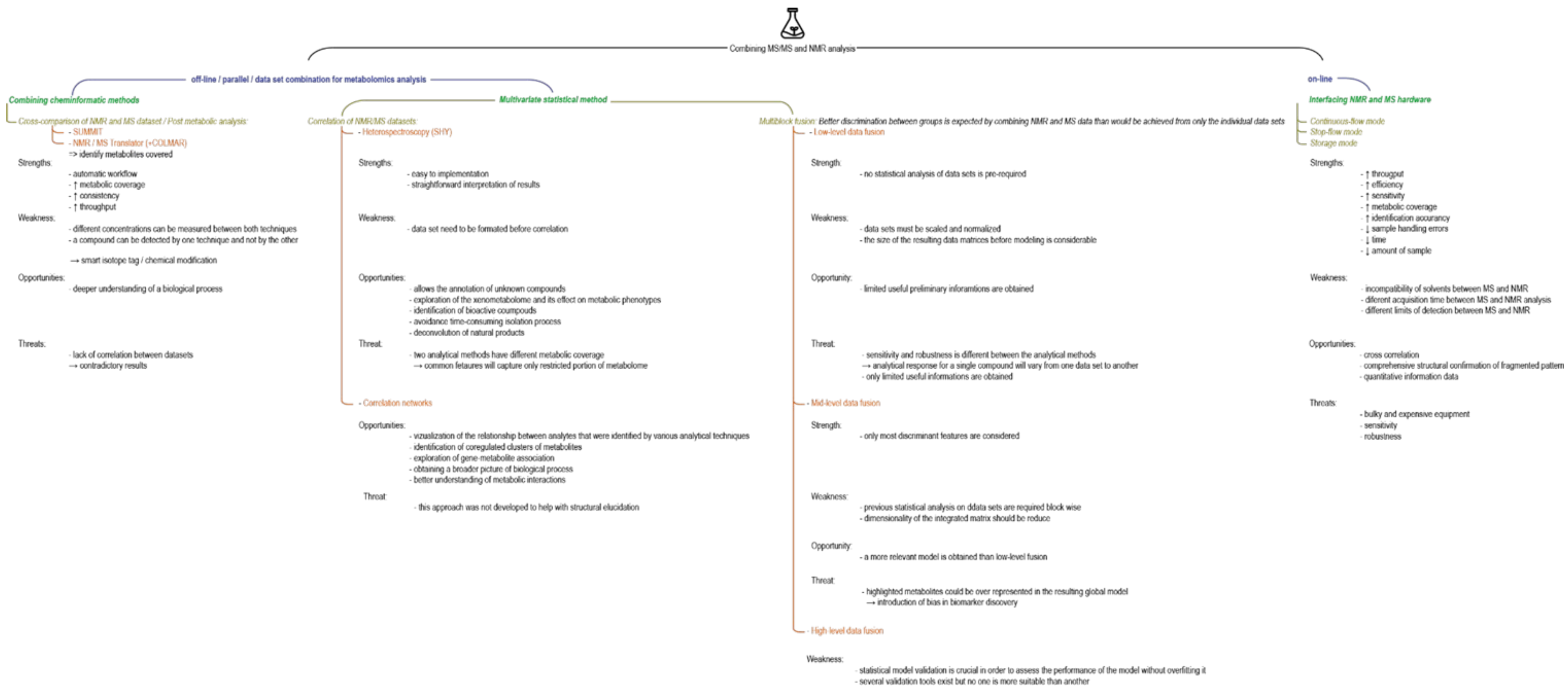
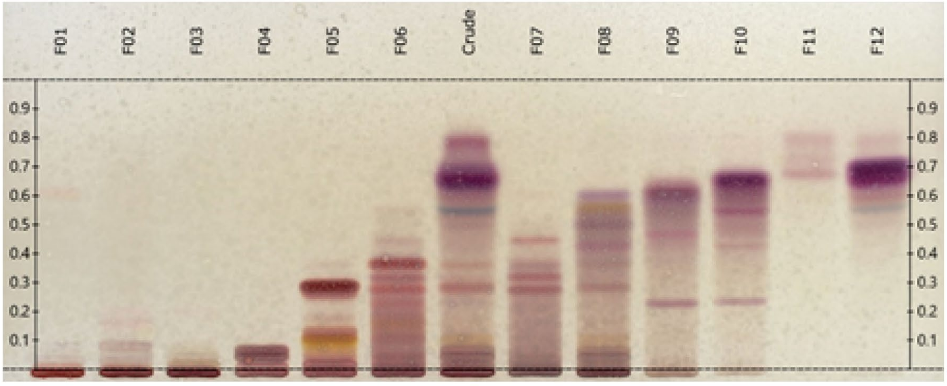


Figure S1: The reported strategies for the joint use of MS and NMR analytical data, with their benefits and constraints

a)



b)

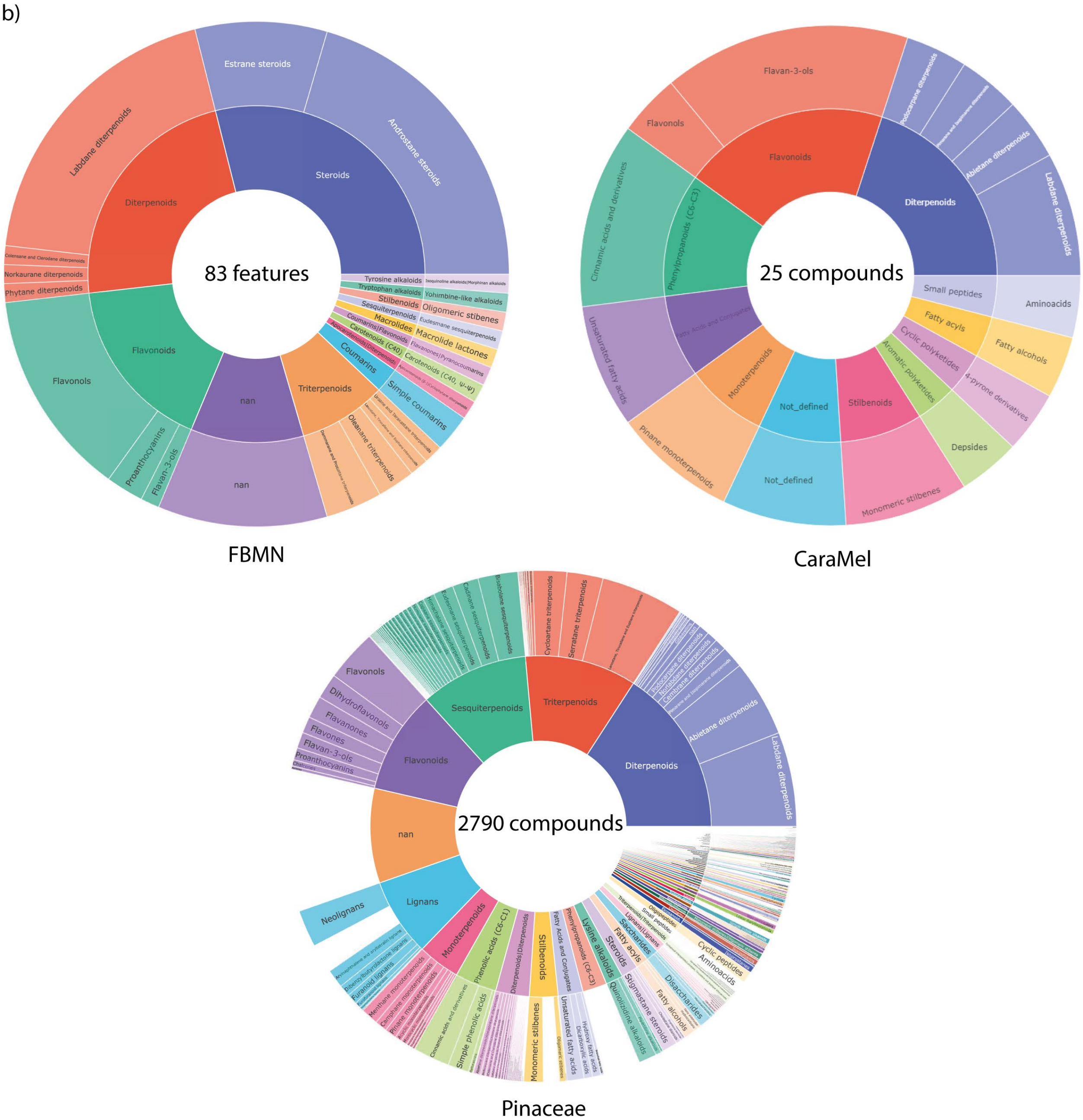


Figure S2: Chemical distributions (A) HPTLC profile of the 12 CPC fractions of the *Larix decidua* EtOAc extract. The plate was sprayed with vanillin/H_2SO_4 derivatization reagent, heated, and observed under white light **(B)** Sunburst charts representing the chemical distribution of the Pinaceae structural database, CaraMel annotations, FBMN annotations

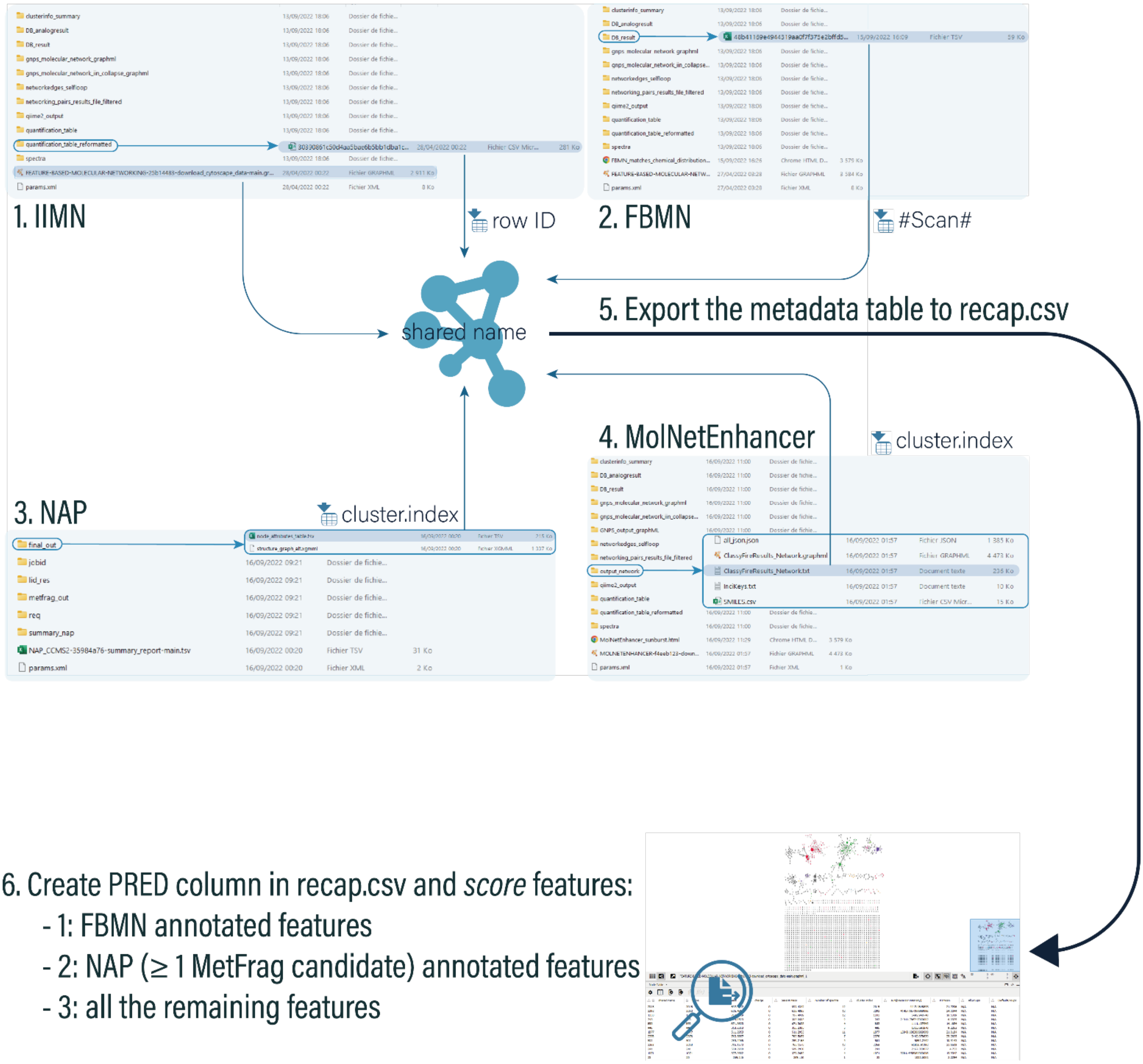


Figure S3: Merging the annotations produced by the MS workflows, leading to the recap.csv file

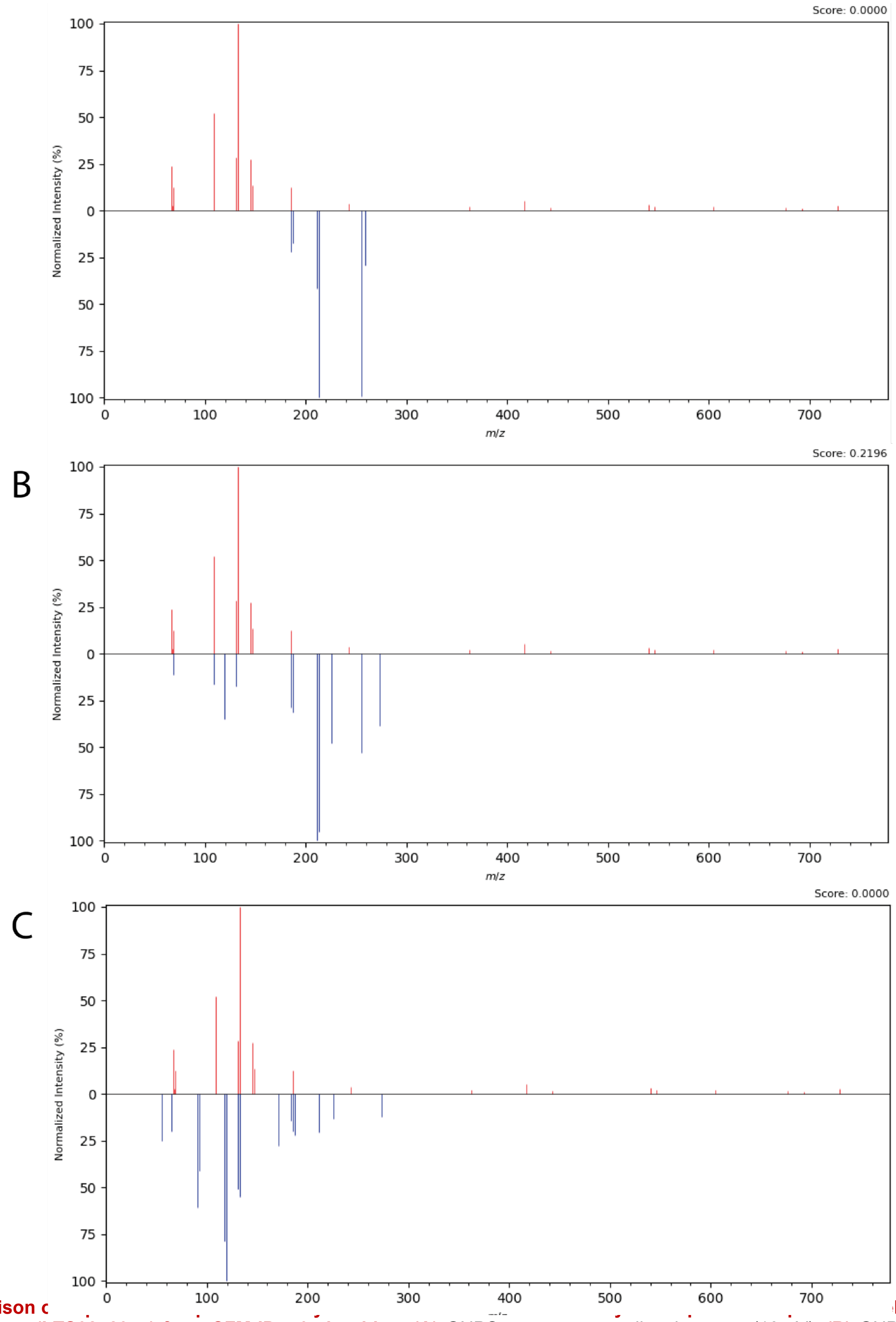
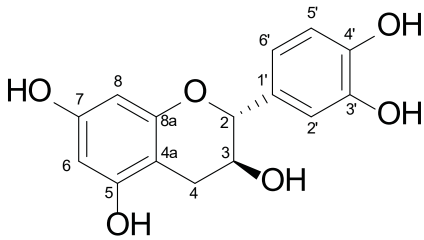


Figure S4: Comparison c hydroabietic acid from LOTUS structure (LTS0252977) from CFM-ID 4.0 algorithm. (A) GNPS spectra vs. predicted spectra (10 eV). **(B)** GNPS spectra vs. predicted spectra (20 eV). **(C)** GNPS spectra vs. predicted spectra (40 eV). Red spectrum: CCMSLIB00000840371 Blue spectra: CFM-ID 4.0 predicted spectrum.

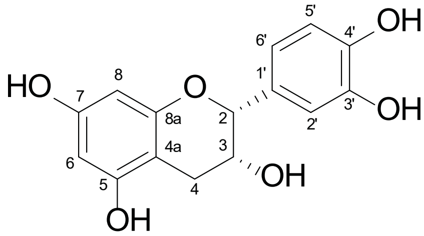
Table S8: Annotated structures by CaraMel workflow, linked to their ^{13}C NMR and ^{1}H chemical shifts.

catechin



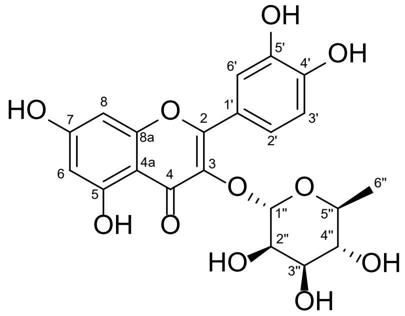
$C_{15}H_{14}O_6$	290.3 g/mol	CAS 154-23-4
Atom number	^{13}C (ppm)	^{1}H (ppm)
2	81.5	4.48
3	66.6	3.81
4	28.3	2.66/2.35
4a	99.6	-
5	156.6	-
6	94.1	5.68
7	156.8	-
8	95.6	5.89
8a	155.7	-
1’	131.0	-
2’	114.9	6.72
3’	145.2	-
4’	145.3	-
5’	115.3	6.68
6’	118.8	6.59

epicatechin



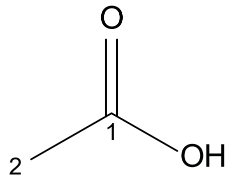
$C_{15}H_{14}O_6$	290.3 g/mol	CAS 490-46-0
Atom number	^{13}C (ppm)	^{1}H (ppm)
2	78.3	4.73
3	65.3	4.00
4	28.7	2.67/2.47
4a	98.9	-
5	156.0	-
6	94.4	5.72
7	156.5	-
8	95.6	5.89
8a	156.1	-
1’	130.9	-
2’	118.3	6.65
3’	144.7	-
3’	144.7	-
4’	144.8	-
5’	115.2	6.66
6’	115.4	6.88

quercetin 3-*O*-rhamnoside



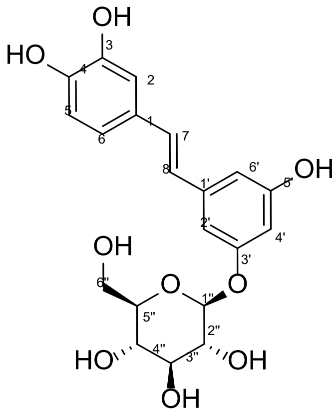
$C_{21}H_{20}O_{11}$	448.4 g/mol	CAS 522-12-3
Atom number	^{13}C (ppm)	^{1}H (ppm)
2	156.8	-
3	134.5	-
4	178.1	-
4a	104.4	-
5	161.6	-
6	99.0	6.21
7	164.6	-
8	94.0	6.40
8a	157.7	-
1’	121.1	-
2’	121.5	7.25
3’	115.8	6.87
4’	148.8	-
5’	145.5	-
6’	116.0	7.29
1’’	102.1	5.25
2’’	70.4	3.98
3’’	70.7	3.51
4’’	71.5	3.15
5’’	71.0	3.20
6’’	17.9	0.81

acetic acid



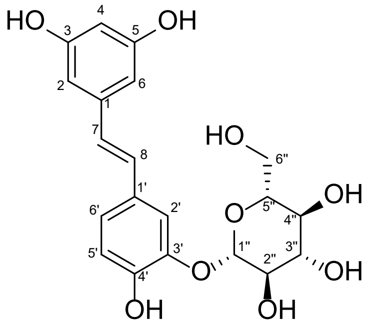
$C_2H_4O_2$	60.5 g/mol	CAS 64-19-7
Atom number	^{13}C (ppm)	^{1}H (ppm)
1	172.3	-
2	21.4	1.91

astringin (*trans*)



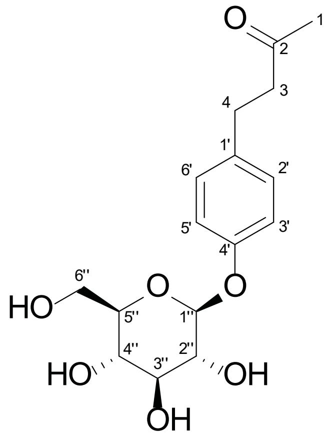
$C_{20}H_{22}O_9$	406.4 g/mol	CAS: 29884-49-9
Atom number	^{13}C (ppm)	^{1}H (ppm)
1	129.0	-
2	113.7	6.99
3	145.8	-
4	146.0	-
5	116.1	6.73
6	119.2	6.85
7	129.3	6.96
8	125.5	6.79
1’	139.8	-
2’	105.2	6.73
3’	159.3	-
4’	103.1	6.34
5’	158.7	-
6’	107.5	6.58
1’’	101.0	4.81
2’’	73.7	3.23
3’’	77.1	3.28
4’’	70.2	3.18
5’’	77.5	3.37
6’’	61.1	3.50/3.74

piceatannol-3'-*O*-glucoside (trans)



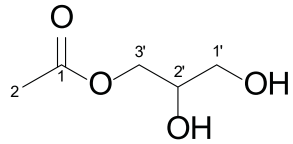
$C_{20}H_{22}O_9$	406.4 g/mol	CAS: 94356-26-0
Atom number	^{13}C (ppm)	^{1}H (ppm)
1	139.6	-
2	104.8	6.40
3	158.9	-
4	102.3	6.14
5	158.9	-
6	104.8	6.40
7	126.8	6.85
8	128.2	6.91
1'	129.3	-
2'	114.6	7.45
3'	146.1	-
4'	147.0	-
5'	116.3	6.80
6'	122.4	7.06
1''	102.9	4.75
2''	73.8	3.34
3''	76.4	3.33
4''	70.5	3.18
5''	77.8	3.43
6''	61.3	3.50/3.79

glucosyl-frambinone



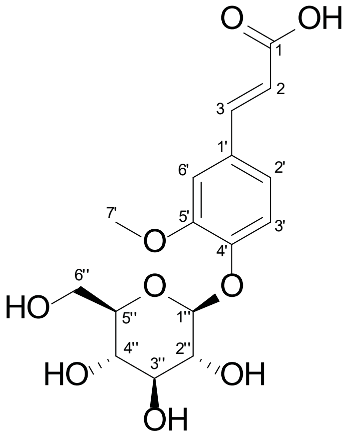
$C_{16}H_{22}O_7$	326.3 g/mol	CAS: 38963-94-9
Atom number	^{13}C (ppm)	^{1}H (ppm)
1	30.2	2.09
2	208.2	-
3	44.7	2.72
4	28.7	2.72
1’	134.8	-
2’/6’	129.3	7.11
3’/5’	116.5	6.93
4’	156.0	-
1’’	100.9	4.79
2’’	73.5	3.22
3’’	76.7	3.24
4’’	70.0	3.15
5’’	77.3	3.28
6’’	60.9	3.47/3.68

glycerol monoacetate



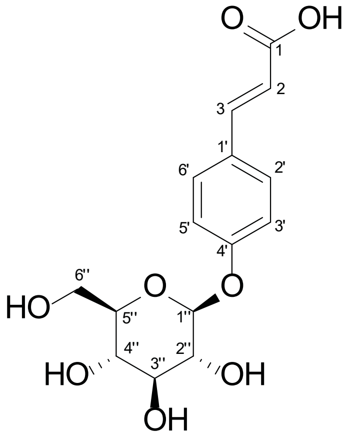
$C_5H_{10}O_4$	134.1 g/mol	CAS 106-61-6
Atom number	^{13}C (ppm)	^{1}H (ppm)
1	170.9	-
2	21.3	2.00
1’	63.1	3.33
2’	69.7	3.63
3’	66.0	3.88/4.02

glucosyl-*trans*-*para*-coumaric acid



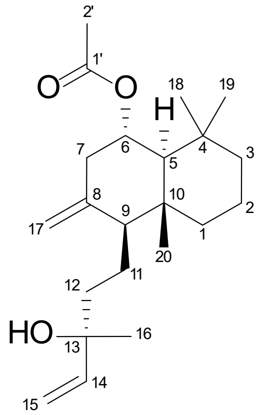
$C_{15}H_8O_8$	326.3 g/mol	CAS 14364-05-7
Atom number	^{13}C (ppm)	^{1}H (ppm)
1	168.3	-
2	117.4	6.41
3	144.1	7.59
1’	128.3	-
2’/6’	130.2	7.65
3’/5’	116.8	7.05
4’	159.3	-
1’’	100.3	4.95
2’’	73.6	3.26
3’’	76.9	3.27
4’’	69.9	3.16
5’’	77.6	3.35
6’’	61.1	3.46/3.68

lavandoside



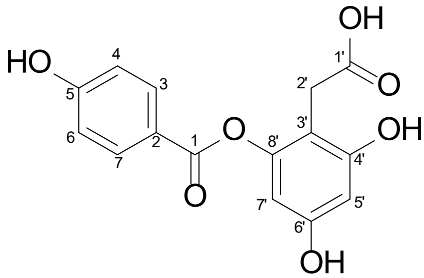
$C_{16}H_{20}O_9$	356.3 g/mol	CAS: 177405-51-3
Atom number	^{13}C (ppm)	^{1}H (ppm)
1	168.3	-
2	118.2	6.46
3	143.7	7.50
1’	128.5	-
2’	115.1	7.09
3’	122.3	7.17
4’	148.5	-
5’	149.3	-
6’	111.3	7.32
7’	55.9	3.82
1’’	99.9	4.97
2’’	73.4	3.27
3’’	77.3	3.29
4’’	70.0	3.16
5’’	77.4	3.35
6’’	60.9	3.45/3.66

larixyl acetate



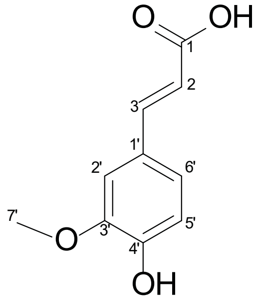
$C_{22}H_{36}O_3$	348.5 g/mol	CAS: 4608-49-5
Atom number	^{13}C (ppm)	^{1}H (ppm)
1	38.8	1.04/1.67
2	18.8	1.45/1.48
3	43.4	1.21/1.32
4	33.4	-
5	56.9	1.44
6	72.6	4.90
7	44.2	2.02/2.54
8	145.0	-
9	55.8	1.61
10	39.7	-
11	17.9	1.27/1.48
12	41.6	1.16/1.56
13	72.1	-
14	146.6	5.84
15	111.1	4.94/5.11
16	28.0	1.13
17	109.5	4.65/4.88
18	36.3	1.00
19	22.4	0.81
20	15.9	0.67
1’	169.8	-
2’	21.9	2.00

2-O-(4-Hydroxybenzoyl)-2,4,6-trihydroxyphenylacetic acid



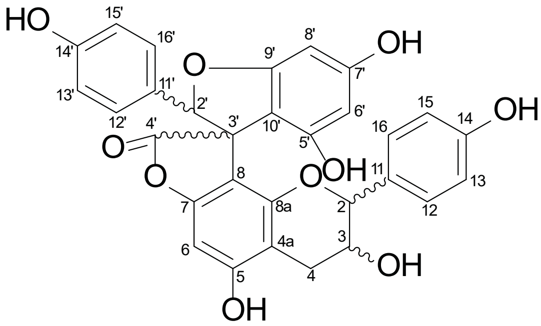
$C_{15}H_{12}O_7$	304.3 g/mol	CAS: -
Atom number	^{13}C (ppm)	^{1}H (ppm)
1	163.9	-
2	119.8	-
3/7	132.4	7.92
4/6	115.9	6.92
5	162.8	-
1’	172.7	-
2’	29.3	3.28
3’	106.1	-
4’	151.0	-
5’	101.0	6.09
6’	157.1	-
7’	100.0	6.24
8’	157.1	-

ferulic acid



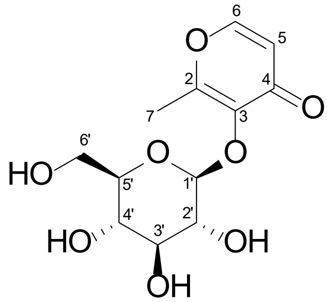
$C_{10}H_{10}O_4$	194.2 g/mol	CAS: 1135-24-6
Atom number	^{13}C (ppm)	^{1}H (ppm)
1	168.2	-
2	115.9	6.36
3	144.8	7.49
1’	126.0	-
2’	111.4	7.28
3’	148.2	-
4’	149.3	-
5’	115.9	6.8
6’	123.0	7.08
7’	56.0	3.82

larixinol isomer 1 + 2



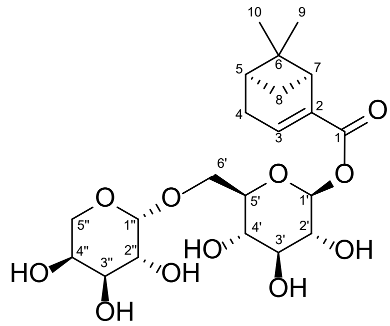
$C_{30}H_{22}O_{10}$	542.5 g/mol	CAS: 101046-79-1
Atom number	^{13}C (ppm)	^{1}H (ppm)
2	77.7 / 78.4	4.60 / 4.90
3	64.5 / 63.7	4.02 / 4.23
4	28.6 / 28.6	2.46/2.51 / 2.67/2.82
4a	103.4 / 103.8	-
5	156.7 / 157.2	-
6	90.1 / 90.7	6.08 / 6.20
7	150.7 / 151.8	-
8	104.2 / 104.4	-
8a	151.9 / 151.7	-
9	129.6 / 129.5	-
10/14	127.7 / 127.9	6.93 / 7.09
11/13	114.8 / 114.8	6.69 / 6.69
12	156.5 / 156.7	-
2’	92.9 / 88.9	5.72 / 6.14
3’	60.1 / 59.9	-
4’	178.7 / 174.7	-
5’	154.1 / 154.9	-
6’	96.1 / 96.1	5.88 / 5.79
7’	160.3 / 160.2	-
8’	89.8 / 89.4	5.90 / 5.88
9’	162.1 / 163.1	-
10’	104.9 / 104.1	-
11’	126.2 / 126.2	-
12’/16’	127.0 / 127.3	6.90 / 6.91
13’/15’	114.8 / 115.2	6.52 / 6.68
14’	157.2 / 157.8	-

dianthoside



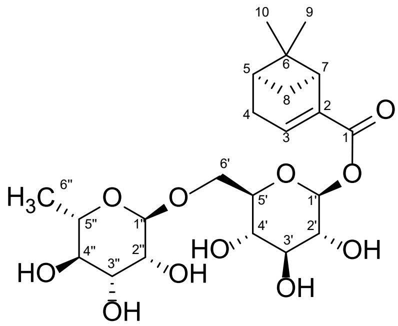
$C_{12}H_{16}O_8$	288.3 g/mol	CAS: 20847-13-6
Atom number	^{13}C (ppm)	^{1}H (ppm)
2	161.8	-
3	142.1	-
4	174.6	-
5	116.5	6.44
6	156.0	8.14
7	15.6	2.37
1’	104.0	4.74
2’	74.3	3.15
3’	76.7	3.19
4’	69.9	3.10
5’	77.7	3.10
6’	31.3	3.44/3.64

arabinosyl-(1->6)-glucosyl-myrtenic acid



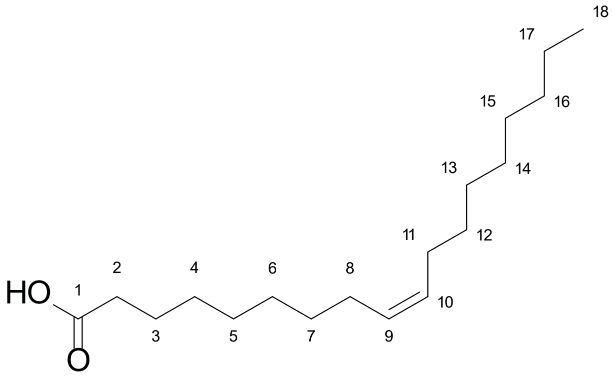
$C_{21}H_{32}O_{11}$	460.5 g/mol	CAS: -
Atom number	^{13}C (ppm)	^{1}H (ppm)
1	164.1	-
2	138.9	-
3	138.6	6.91
4	32.1	2.41/2.48
5	40.0	2.11
6	37.5	-
7	40.9	2.70
8	31.1	1.03/2.46
9	25.9	1.31
10	21.2	0.76
1’	94.6	5.36
2’	72.8	3.16
3’	76.6	3.24
4’	70.1	3.08
5’	76.6	3.41
6’	67.2	3.38/3.88
1’’	108.7	4.71
2’’	82.3	3.79
3’’	77.5	3.63
4’’	84.0	3.70
5’’	31.5	3.41/3.54

rhamnosyl-(1->6)-glucosyl-myrtenic acid



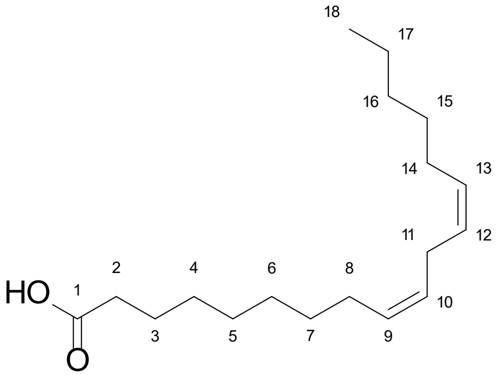
$C_{22}H_{34}O_{11}$	474.5 g/mol	CAS: -
Atom number	^{13}C (ppm)	^{1}H (ppm)
1	164.1	-
2	138.9	-
3	138.6	6.91
4	32.1	2.41/2.48
5	40.0	2.11
6	37.5	-
7	40.9	2.70
8	31.1	1.03/2.46
9	25.9	1.31
10	21.2	0.76
1’	94.7	5.37
2’	72.8	3.16
3’	76.6	3.24
4’	69.8	3.10
5’	76.7	3.36
6’	66.5	3.45/3.78
1’’	100.8	4.53
2’’	70.9	3.40
3’’	70.7	3.59
4’’	72.2	3.17
5’’	68.7	3.42
6’’	18.2	1.10

oleic acid



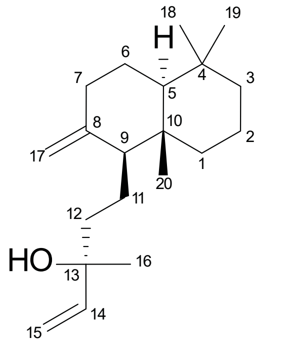
$C_{18}H_{34}O_2$	282.5 g/mol	CAS: 112-80-1
Atom number	^{13}C (ppm)	^{1}H (ppm)
1	174.8	-
2	34.0	2.17
3	24.8	1.46
4-7	28.8-29.4	1.22-1.30
8	26.9	1.97
9	129.9	5.32
10	129.9	5.32
11	26.9	1.97
12-15	28.8-29.4	1.22-1.30
16	31.6	1.23
17	22.3	1.25
18	14.3	0.85

linoleic acid



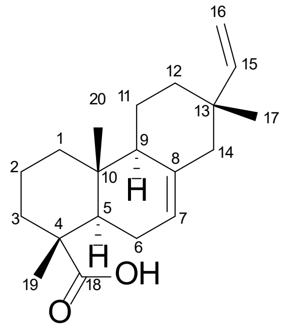
$C_{18}H_{32}O_2$	280.4 g/mol	CAS: 60-33-3
Atom number	^{13}C (ppm)	^{1}H (ppm)
1	174.7	-
2	34.0	2.16
3	24.9	1.48
4	29-30	1.25
5	29-30	1.2-1.3
6	29-30	1.2-1.3
7	29-30	1.31
8	27.0	2.00
9	129.9	5.32
10	128.0	5.30
11	25.6	2.73
12	128.0	5.30
13	129.9	5.32
14	27.0	2.00
15	29-30	1.31
16	31.3	1.25
17	22.4	1.25
18	14.2	0.85

13-epimanool



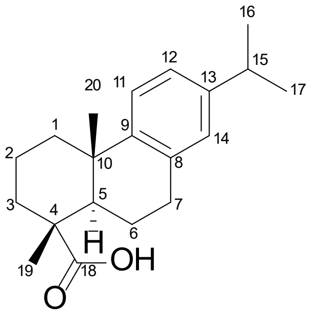
$C_{20}H_{34}O_1$	290.5 g/mol	CAS: 1438-62-6
Atom number	^{13}C (ppm)	^{1}H (ppm)
1	38.7	1.05/1.80
2	19.2	1.44/1.50
3	42.0	1.14/1.36
4	33.5	-
5	55.1	1.08
6	24.2	1.24
7	38.0	1.81/2.32
8	148.5	-
9	56.9	1.49
10	39.6	-
11	17.9	1.48
12	41.6	1.12/1.56
13	72.0	-
14	146.6	5.83
15	111.1	4.93/5.11
16	27.9	1.12
17	106.9	4.52/4.79
18	21.8	0.77
19	33.7	0.85
20	14.6	0.62

isopimaric acid



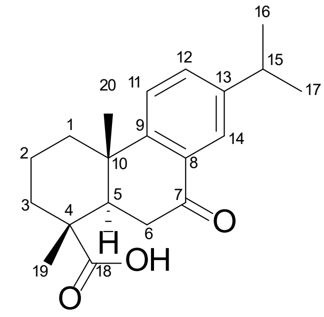
$C_{20}H_{30}O_2$	302.5 g/mol	CAS: 5835-26-7
Atom number	^{13}C (ppm)	^{1}H (ppm)
1	38.8	1.05/1.80
2	17.9	1.47
3	36.9	1.53/1.65
4	45.6	-
5	44.9	1.83
6	25.0	1.55/1.90
7	121.1	5.31
8	135.5	-
9	51.8	1.69
10	34.8	-
11	18.4	1.63/1.71
12	35.8	1.32/1.44
13	36.8	-
14	45.8	1.86/1.93
15	150.2	5.80
16	110.0	4.87/4.93
17	21.5	0.83
18	179.8	-
19	17.5	1.16
20	15.3	0.85

dehydroabietic acid



$C_{20}H_{28}O_2$	300.4 g/mol	CAS: 1740-19-8
Atom number	^{13}C (ppm)	^{1}H (ppm)
1	38.1	1.31/2.30
2	18.4	1.63/1.72
3	36.6	1.56/1.68
4	46.7	-
5	45.0	2.02
6	21.4	1.41/1.75
7	29.8	2.75/2.82
8	134.4	-
9	147.1	-
10	36.7	-
11	124.3	7.15
12	124.0	6.97
13	145.3	-
14	126.7	6.84
15	33.2	2.77
16	24.3	1.16
17	24.3	1.16
18	179.7	-
19	16.7	1.16
20	25.1	1.13

7-oxodehydroabietic acid



$C_{20}H_{26}O_3$	314.4 g/mol	CAS: 18684-55-4
Atom number	^{13}C (ppm)	^{1}H (ppm)
1	37.3	1.47/2.37
2	Not detected	Not detected
3	37.2	1.49/2.37
4	46.0	-
5	44.3	2.53
6	37.9	2.24/2.72
7	198.2	-
8	130.5	-
9	153.8	-
10	37.4	-
11	124.5	7.41
12	132.8	7.50
13	146.5	-
14	124.0	7.70
15	33.2	2.92
16	23.9	1.20
17	23.9	1.20
18	179.8	-
19	16.8	1.20
20	23.9	1.20