13C NMR Dereplication Using MixONat Software: A Practical Guide to Decipher Natural Products Mixtures

Antoine Bruguière, Séverine Derbré, Dimitri Bréard, Félix Tomi, Jean-Marc Nuzillard, Pascal Richomme

To cite this version:

Antoine Bruguière, Séverine Derbré, Dimitri Bréard, Félix Tomi, Jean-Marc Nuzillard, et al.. 13C NMR Dereplication Using MixONat Software: A Practical Guide to Decipher Natural Products Mixtures. Planta Medica, 2021, 10.1055/a-1470-0446 . hal-03220850

HAL Id: hal-03220850
https://hal.univ-reims.fr/hal-03220850v1

Submitted on 30 Jun 2021

^{13}C NMR dereplication using MixONat software: a practical guide to decipher natural products mixtures[#]

Antoine Bruguière[1], Séverine Derbré[1,]*, Dimitri Bréard,[1] Félix Tomi,[2] Jean-Marc Nuzillard,[3] Pascal Richomme[1,*]

[1] Univ Angers, SONAS, SFR QUASAV, Faculty of Health Sciences, Dpt Pharmacy, Angers, France

[2] Université de Corse-CNRS, UMR 6134 SPE, Equipe Chimie et Biomasse, Ajaccio, France

[3] Université de Reims Champagne Ardenne, CNRS, ICMR UMR 7312, Reims, France

* Correspondence
Dr Séverine Derbré, PharmD, PhD
Univ Angers, SONAS, SFR QUASAV, Faculty of Health Sciences, Dpt Pharmacy
16 Bd Daviers, F-49000 Angers, France
Phone: +33 249 180 440 ; fax : +33 241 226 634
severine.derbre@univ-angers.fr
Prof Pascal Richomme, PhD
Univ Angers, SONAS, SFR QUASAV, Faculty of Health Sciences, Dpt Pharmacy
16 Bd Daviers, F-49000 Angers, France
Phone: +33 249 180 437 ; fax : +33 241 226 634
pascal.richomme@univ-angers.fr

[#] Dedicated to Professor Arnold Vlietinck on the occasion of his 80th birthday

Abstract

The growing use of herbal medicines worldwide requires ensuring their quality, safety and efficiency to consumers and patients. Quality controls of vegetal extracts are usually undertaken according to pharmacopoeial monographs. Analyses may range from simple chemical experiments to more sophisticated but more accurate methods. Nowadays, metabolomic analyses allow a fast characterization of complex mixtures. In the field, beside mass spectrometry (MS), nuclear magnetic resonance (NMR) spectroscopy tends to gain importance in the direct identification of natural products in complex herbal extracts. For a decade, automated dereplication processes based on ^{13}C-NMR have been emerging to efficiently identify known major compounds in mixtures. Though less sensitive than MS, ^{13}C-NMR has the advantage of being appropriate to discriminate stereoisomers. Since NMR spectrometers nowadays provide useful dataset in a reasonable time frame, we have recently made available MixONat, a software that processes ^{13}C as well as DEPT 135 and 90 data allowing carbon multiplicity (*i.e.* CH_3, CH_2, CH and C) filtering as a critical step. MixONat requires experimental or predicted chemical shifts (δ_C) databases (DB) and displays interactive results that can be refined based on the user's phytochemical knowledge. The present article provides step-by-step instructions to use MixONat starting from DBs creation with freely available and/or marketed δ_C datasets. Then, for training purpose, the reader is led through a 30-60 min procedure consisting in the ^{13}C-NMR based dereplication of a peppermint essential oil.

Key words:

Database, Essential oils, Lamiaceae, *Mentha piperita*, MixONat, ^{13}C-NMR-based dereplication

Abbreviations

DB: Database

DEPT: Distortionless enhancement by polarization transfer

$\delta_{C\text{-}SDF}$: Carbon 13 chemical shifts in the DB

δ_C: Recorded carbon 13 chemical shifts

EO: Essential oil

NPs: Natural products

Introduction

Currently, among alternative medicines, the use of herbal drugs and dietary supplements continues to expand worldwide with many people resorting to them to either prevent or cure various minor illnesses or in a quest for well-being [1-3]. In this framework, the quality of raw materials and plant extracts contributes to their safety and effectiveness. Indeed, for a given species, the chemical content of a medicinal plant, and thus its biological effects, may vary depending on chemotypes, growing conditions (*e.g.* soils, weather), harvest time as well as post-harvest handling (*e.g.* drying, storage) [4,5]. Quality controls of herbal drugs are usually based on scientific benchmarks such as pharmacopoeias aiming at controlling their identity (botanical and chemical criteria), purity [*i.e.* assays such as ash values, loss on drying or thin layer chromatography (tlc) to detect contaminants] and, finally, content of active constituents or markers (*i.e.* assays using analytical methods) to check quantitatively their composition. Addressing these issues, besides spectrophotometric analyses, different chromatographic methods such as High-Pressure Liquid Chromatography with UV Detector (HPLC-UV) or Gas Chromatography with Flame Ionization Detector (GC-FID) are routinely used to rapidly determine complex mixtures.

Metabolomic analyses allow to rapidly decipher complex mixtures of organic compounds, including vegetal extracts. In this field, beside liquid and gas chromatography hyphenated to mass spectrometry (LC-MS and GC-MS) [6,7], nuclear magnetic resonance (NMR) spectroscopy tends to gain importance in the direct identification of natural products (NPs) in complex herbal matrices. For a decade, automated dereplication processes based on ^{13}C-NMR have been emerging to efficiently identify major compounds in mixtures using moderate field instruments (400 MHz), freely available automation procedures and dedicated software [8-11]. While a higher sensitivity is a major advantage of MS, ^{13}C-NMR is indeed highly suitable for the discrimination of diastereomers. Additionally MS analysis usually requires a separation step of the mixture using appropriate columns and chromatographic conditions as well as standards or benchmarks to do so. It should be noted that commercial (*e. g.* ACD/Labs [12]) and open source (*e. g.* CSEARCH [13]) solutions allow efficient computer-assisted peer reviewing of pure NPs based on their NMR spectra. However, such tools were not developed to perform dereplication analyses of crude extracts.

In this context, we recently proposed a freely distributed algorithm, namely MixONat, for dereplication analyses of major NPs in crude extracts or in less complex fractions. MixONat analyses a single {^{1}H}-^{13}C NMR spectrum which may be optionally combined with DEPT-135 and DEPT-90 data to discriminate between CH_3, CH_2, CH, and C, as well as with molecular weight filtering. The software requires predicted or experimental carbon chemical shifts ($\delta_{C\text{-}SDF}$) databases (DBs) and displays results that can be refined interactively [11,14] (Figure 1). MixONat has demonstrated its effectiveness in elucidating the composition of various mixtures containing alkaloids, di- and triterpenes or xanthones [11].

Essential oils (EOs) are complex mixtures of volatile and odorous principles, including monoterpenes, a significant proportion of them consisting of chiral compounds [4]. GC-FID and GC-MS are usually used to identify and/or quantify such volatiles in EOs [15]. They require comparisons with standards or spectral libraries. Moreover, several analytical problems may occur: The absence of elution or co-elution of volatiles and sometimes their thermosensitivity may impair identifications [16-18]. One drawback of MS detection is also its reduced ability to distinguish between diastereomers and positional isomers. Due to these limitations, we think that other spectroscopic tools such as $\{^{1}H\}$-^{13}C-NMR associated with dereplication software may be useful to accurately characterize major volatiles in EOs [19].

Thus, for training purpose, the present paper presents a workflow which enables to use MixONat software to analyse peppermint EO using ^{13}C-NMR data as well as DEPT-135 and DEPT-90 experiments. The process starts from the building of appropriate DBs consisting in either experimental δ_C or free and/or commercial predicted $\delta_{C\text{-}SDF}$ datasets whereas practical information for exporting input files from NMR spectra are given. Finally, it helps the user to correctly interpret the results suggested by MixONat. The ability of the process to rapidly characterize major monoterpenes from essential oils, including diastereomers, is eventually demonstrated.

Results and Discussion

As a reference, peppermint EO was analysed using both GC-FID and GC-MS following the conditions described by the European Pharmacopoeia. As expected, menthol (**1**), menthone (**2**), 1,8-cineole (**3**), menthyl acetate (**4**), isomenthone (**5**), limonene (**6**) and menthofurane (**7**) (Figure 2) were identified as major monoterpenes by comparison of their retention times with those of authentic samples, and by computer matching of their fragmentation patterns against the NIST mass spectral library (Table 1, Figure 1S). In the present work, we evaluated ^{13}C NMR and MixONat software as an alternative tool to identify monoterpenes **1**-**7** in peppermint EO through a step-by-step procedure.

The first step consists in creating DBs. As far as experimental data are concerned, when DBs containing NPs of interest and their experimental δ_C are available in the selected deuterated solvent, their use obviously increases the odds for better matches [20]. However, such comprehensive DBs, already shaped to be used with MixONat are not yet available though researchers usually keep spectral data including δ_C in their laboratories and share them through academic publishing. Thus, even if the task is tedious, for given botanical genera or families of interest, using dedicated software, or even a simply free text editor (Chart 1), small DBs of NPs associated with their δ_C may be manually built [9,10] and ultimately easily shared with the scientific community [21]. As far as volatile compounds from EOs are concerned, such an approach has been initiated years ago [22]. Thus, Mentha DB1 (30 NPs) including the $\delta_{C\text{-}SDF}$ was manually built from the experimental δ_C of the monoterpenes usually described in peppermint oil [23] using ACD NMR predictors (C,H) software.

Alternatively, in a first approach, $\delta_{C\text{-}SDF}$ may be predicted when data are not easily available or when the number of NPs of interest is too large to achieve the task is a reasonable time. The first step to achieve this goal consists in finding the best way to collect NPs of interest as a structure-data file (SDF), comprising, for each compound, the MDL molfile with associated data (*e.g.* name, CAS number, molecular formula, molecular weight, source …*etc.*). As aforementioned, considering dereplication of plant or microorganism extracts, a chemotaxonomic-based selection is relevant [24]: all NPs previously isolated from a genus or a family are easily exported as SDF using various DBs accessible through subscription. For the dereplication of peppermint EO, a search on SciFinder [25] using the keyword Lamiaceae allowed to select more than 10000 references. They were further reduced to 3499 referring to “Natural products” ”Pharmaceutical natural products” “Essential oils” *etc.* using an analysis of the references with the filter “CA concept heading” proposed by SciFinder. After an additional refining by “categories” (*i.e.* Analytes and matrices) and a filtering based on the molecular weight, all relevant NPs were subsequently exported as SDF to obtain the NPs from Lamiaceae DB2 (982 NPs) (Figure 2S). A similar selection is also possible using the upgraded version Scifinder-n or Reaxys [26].

Alternatively, the freely available KnapsackSearch program (KS) was used [27]. KS is closely related to the KNApSAcK project which contains 129 662 species–metabolite relationships encompassing 23 911 species and 53 032 metabolites. [28,29]. KS associates a set of NPs related to a list of genera with their sources. Running KS with the set of genera described in

Lamiaceae family as a keyword input results in a SDF of 958 NPs which was used to obtain easily Lamiaceae DB3 and DB4 during the following step.

The second step consists in predicting the necessary $\delta_{C\text{-}SDF}$ values. Useful prediction tools are either commercial or freely available. Therefore, as an example, we used both ACD NMR predictors (C,H) [10,12] and nmrshiftdb2 [30,31] to predict $\delta_{C\text{-}SDF}$ and obtain Lamiaceae DB2-3 and DB4 respectively.

Finally, whether starting from a DB of experimental or predicted $\delta_{C\text{-}SDF}$, the use of the "CtypeGen" tab in MixONat software (Figure 3S) is mandatory to sort each $\delta_{C\text{-}SDF}$ by carbon type (*i.e.* Cq, CH, CH_2, CH_3) so that the DB will be readable by MixONat (see SI: Create a DB readable by MixONat).

In a second step, NMR spectra are recorded, processed and δ_C exported in required format. The $\{^1H\}$-^{13}C NMR spectrum (1024 scans) of peppermint EO (90 mg) was recorded in $CDCl_3$ (0.6 mL) using a routine 400 MHz NMR spectrometer. Though not mandatory for the software, DEPT-135 and DEPT-90 spectra were also recorded since carbon multiplicity was previously shown as a powerful discriminant filter [11]. From there the present ^{13}C NMR based dereplication process requires 30 to 60 min only until completion.

As usual, ^{13}C NMR, DEPT-135 and DEPT-90 spectra are to be phased and baseline corrected using a dedicated software. Among critical features, the user should obviously reference the ^{13}C-NMR spectrum on the central resonance of the deuterated solvent with caution; concerning DEPT spectra, their careful alignment on a selected δ_C, *i.e.* a chosen CH, is essential for MixONat proper use. Whatever the reason, if it is not the case, the default values of "DEPT alignment" at 0.02 ppm (MixONat, Tab2: Parameters, Figure 4S) should be slightly increased accordingly. Of course, the peak picking process is also critical. It may be manual but the use of a minimum intensity threshold is preferred to automatically collect positive ^{13}C NMR and DEPT-90 signals and positive and negative DEPT-135 signals while avoiding potential noise artifacts. This step is sometimes difficult to implement, for example when the intensities of δ_C from major NPs' quaternary carbons are in the same range than the ones from minor NPs' methyl groups. Back to peppermint EO, a high threshold value was first chosen to pick peaks (Figures 5S-6S). However, the minor signal at δ_C 8.20 ppm is obviously arising from a methyl group, which means that, if the corresponding NP includes quaternary carbons, their intensities will be close to the noise level. Alternatively, a second peak picking was carried out using a lower threshold value (Figures 7S-8S). After eliminating deuterated solvent signals, the lists of chemical shifts and intensities from ^{13}C NMR, DEPT-135 and 90 were exported and/or saved as separated Comma Separated Values (CSV) files, which were then used as input files by the MixONat software (Figure 9S). Thus, in the present practical exercise, two batches of data were used as input files. The first batch consisted of those selected on the basis of a high threshold value, namely Peppermint EO 13C.csv, Peppermint EO DEPT 135.csv, Peppermint EO DEPT 90.csv (Figure 5S, files available in supporting information) corresponding to ^{13}C-NMR, DEPT-135 and DEPT-90 data respectively. The second batch was made of chemical shifts picked using a lower threshold value, namely Peppermint EO Minor 13C.csv, Peppermint EO Minor DEPT 135.csv, Peppermint EO Minor DEPT 90.csv (Figure 7S, files available as supporting information).

In a third and last step, MixONat basic and advanced exploitation can be detailed as follows: A first dereplication analysis was undertaken using the small experimental Mentha DB1 (30 monoterpenes) as well as δ_C and associated DEPT data picked using the high threshold value as input files in tab 1 “Inputs” of the graphical user interface of MixONat (Figure 10S). Before running the matching process with MixONat, a verification step of each file is recommended using the "Check" button. A description of the SDF and each .csv files is thus expected (Figure 11S) to be displayed. If not, a careful examination of the defective file and a formatting step using Notepad++ usually fixes the bug. In the “Parameters” tab, the tolerance value **ε** reflects the accuracy of the used DB. It was set at 0.5 ppm since an experimental DB was used [32]. The “equivalent carbons” were authorized meaning that same δ_C might be matched multiple times if several identical $\delta_{C\text{-SDF}}$ were found. As a result, the software sorted out compounds of the DB by decreasing score and increasing error. The score is defined as the number of carbon chemical shifts in the ^{13}C-NMR spectrum (δ_C) matched with $\delta_{C\text{-SDF}}$ out of the number of carbons of the compound (Figure 1). The error is the cumulated absolute difference between matched signals (i.e. $\Sigma\ |\delta\text{C-SDF}-\delta\text{C}|$). MixONat sorted out five major monoterpenes (**1**-**5**: ranks 1-4 and 6) as well as neomenthol (**8**: rank 5), a diastereomer of menthol with a perfect match (*i.e.* score 1.0). The presence of **8** in peppermint EO was confirmed by a careful examination of δ_C (Table 1S). It should be noted that neither GC-FID nor GC-MS using the method described in the European Pharmacopeia was able to distinguish neomenthol (**8**) from menthol (**3**) because of their co-elution and similar MS fragmentation, pointing to the complementarity of ^{13}C NMR for quality control. Minor monoterpenes such as limonene (**6**: rank 7, score 0.9) and menthofuran (**7**: rank 12, score 0.7) were also identified (Table 1). An automatic peak picking using a lower threshold value allowed to pick additional quaternary carbons for these minor

NPs. Finally, using these δ_C lists, the first seven guess compounds were six monoterpenes (**1**-**6**) expected in peppermint EO according to the European Pharmacopoeia as well as neomenthol (**8**). Menthofuran (**7**) was ranked 8th (score 0.9) as the quaternary C-2 is in the background noise (Table 1, Figures 7S, 12S).

Alternatively, MixONat can work with DBs of predicted $\delta_{C\text{-}SDF}$. As MixONat software ranks all compounds of the DB by decreasing score and increasing error, the result of the dereplication process depends on two crucial parameters, *i.e.* the selected NPs and the accuracy of the $\delta_{C\text{-}SDF}$. Thus, one can wonder about the relevance of such predicted DBs. In the present study, $\delta_{C\text{-}SDF}$ from Lamiaceae DB2-3 and DB4 were predicted using either the commercial ACD NMR predictors (C,H) or the open NMR web database nmrshiftdb2.

Using Lamiaceae DB2 (952 NPs | ACD NMR predictors [C,H]), δ_C obtained with a high threshold value and a tolerance kept at 1.3 ppm [10,11], 8 monoterpenes reached a perfect score. Among them, menthyl acetate (**4**), menthol (**1**), isomenthone (**5**) and 1,8-cineole (**3**) were ranked at positions 1 to 4 respectively. NPs suggested in positions 4 to 8 were either the same monoterpenes without any stereochemistry or isomers of menthyl acetate (Table 1, Figures 13S, 15S). To conclude on the actual presence of these monoterpenes first ranked in the studied EO, a comparison of δ_C picked in the ^{13}C NMR spectrum should be done with experimental data in the same deuterated solvent. This may be easily achieved using SpectraBase [33], a free spectral database including NMR data of various NPs. Alternatively, SciFinder through subscription allows fast access to previous publications describing δ_C of a given NP [25]. On this basis, menthol (**1**), isomenthone (**5**), 1,8-cineole (**3**) and menthyl acetate (**4**) could be rapidly identified with certainty in peppermint EO using ^{13}C NMR (Table S1). Moreover, on ^{13}C NMR spectra, for a given compound, the individual intensities of C, CH, CH_2, CH_3 δ_C are approximately proportional. As a consequence, a way to rapidly select NPs to be checked is to consult the reconstructed ^{13}C NMR spectrum by clicking on the button “Details” for monoterpenes suggested in the first positions (Figure 13S): in the present example, considering the relative intensities of the signals in each reconstructed spectrum, experimental chemical shifts of isomenthone (**2**: rank 3) were searched among δ_C while this was considered useless for oplopanone (rank 9) (Figure 14S). Then, limonene (**6**) and menthone (**2**) were ranked 10th and 11th with a score of 0.9 (Figure 15S). For limonene (**6**), a minor monoterpene from EO, the quaternary C-8 (δ_C 150.1 ppm in $CDCl_3$) was not matched because not picked when using the high threshold value (Figure 5S). But the corresponding δ_C was identified as a minor signal after a careful examination of the ^{13}C NMR spectrum. Regarding menthone (**2**), ketone C-1 (δ_C 212.6 ppm in $CDCl_3$) was inaccurately predicted by ACD / Labs Spectrus processor in

Lamiaceae DB2 ($\delta_{C\text{-}SDF}$ 210.8 ppm, $\Delta\delta > 1.3$ ppm). However, the MixONat software offers an interactive interface (Figure 14S) which allows the user to manually add or remove mismatched signals. After adding the missing signals for limonene (**6**) and menthone (**2**), they moved to positions 2 and 3 respectively, while menthyl acetate (**4**) remained at rank 1 and menthol (**1**), isomenthone (**5**) and 1,8-cineole (**3**) shifted from 2-4 to 4-6 positions respectively.

Using the Lamiaceae DB3 (958 NPs, KnapsackSearch, ACD NMR predictors [C,H]), the same approach first suggested neomenthol (**8**) (Figure 16S), a diastereomer of menthol (**1**), which was not identified by GC in the condition described by the European Pharmacopoeia. It was not suggested while using Lamiaceae DB2 as unfortunately the approach to collect NPs from Lamiaceae using SciFinder (see above) did not select this monoterpene. Nonetheless, it should be noted that neomenthyl acetate was ranked 7th with Lamiaceae DB2. This exemplifies the relevance of using various DBs in such ^{13}C NMR based dereplication approach. As with DB2, the use of Lamiaceae DB3 suggested menthyl acetate (**4**), menthol (**1**), isomenthone (**5**), menthone (**2**), limonene (**6**) and 1,8-cineole (**3**) among the twelve first hypotheses (Figure 16S). Finally, Lamiaceae DB4 was constituted of the same 958 NPs as Lamiaceae DB3 but with predicted $\delta_{C\text{-}SDF}$ using combination of KnapsackSearch program and nmrshiftdb2. As a result, even if the $\delta_{C\text{-}SDF}$ prediction is less accurate (Table S1), this free solution succeeded in suggesting the presence of all major monoterpenes (**1**-**3**, **5**-**6**, **8**) in peppermint EO among the eleven first suggestions, except for menthyl acetate (**4**: rank 37). It should be noted that both enantiomers were usually suggested (*i.e.* rank 2: (-)-menthone and rank 3: (+)-menthone) (Table 1, Figure 17S).

Through this practical case, we exemplified the simplicity and efficacy of ^{13}C-NMR-based dereplication using the freely available MixONat and KnapsackSearch software to identify the major products in complex extracts such as EOs. The process requires ^{13}C-NMR and DEPT data recorded thanks to a routine NMR spectrometer and DBs inventorying structures of interest associated with their δ_C. A chemotaxonomic approach is proposed to build reasonable sized libraries from selected NPs. When available, experimental δ_C lead to the best outcomes but suggestions using predicted values calculated by commercial and free programs are precise enough to rapidly identify the major NPs. The chosen example, peppermint EO, also showed that dereplication by ^{13}C-NMR distinguishes menthol diastereomers whereas GC-FID or GC-MS using the method described in the European Pharmacopeia fail to do so. ^{13}C-NMR-based dereplication processes may thus be considered for the study as well as for the quality control of EOs and medicinal plants.

Materials and Methods

Chemicals

Mentha × *piperita* L. essential oil (16020152/K) was purchased from Laboratoire Cooper.

Apparatus and operation conditions

Gaz chromatography hyphenated with flame ionization detector (GC-FID) of peppermint oil was performed as described in the European Pharmacopoeia [15] with a 6890 GC system (Agilent Technologies) equipped with a Phenomenex Zebron ZB-5 column (30 m × 0.25 mm × 0.25 µm film thickness). The temperature program started with a 10 min period at 60 °C, then temperature was increased to 180 °C at a rate of 2 °C/min and finally stabilized at 180 °C for 5 min before returning to the initial value. The carrier gas was helium (1.5 ml/min); 1 µL of sample (2% in methanol) was injected; the split ratio was 10:1. Identification of the monoterpenes was based on the comparison of the retention times with those of authentic samples.

GC hyphenated with Mass Spectrometry (GC-MS) analysis of peppermint essential oil was performed with a GCMS-QP2010 apparatus (Shimadzu) in the same conditions as those described for GC-FID analyses. The ionic source and interface temperatures were 220 and 200 °C respectively, operating in the Electron Impact (EI) ionization mode (ionization energy at -70 eV). Identification of the monoterpenes was based on computer matching against the commercial NIST 11 and 11S mass spectral libraries.

Peppermint essential oil (90 mg) was dissolved in 600 µL of $CDCl_3$. NMR analyses were performed at 298 K on a JEOL 400MHz YH spectrometer (JEOL Europe) equipped with an inverse 5 mm probe (ROYAL RO5). For ^{13}C NMR (100 MHz) spectra, a WALTZ-16 decoupling sequence was used with an acquisition time of 1.04 s (32768 complex data points) and a relaxation delay of 2 s. 1024 scans were collected for 90 mg of essential oil to obtain a satisfactory S/N ratio. A 1 Hz exponential line broadening filter was applied to each FID prior to the Fourier transformation. Spectra were manually phased and baseline corrected using the MestReNova software (Mestrelab Research) and referenced on the central resonance of the deuterated solvent [34] at δ_C 77.16. For DEPT experiments, 512 scans were required for 90 mg of essential oil and alignments with the ^{13}C spectrum were made using a given δ_C. A minimum intensity threshold was then used to automatically collect positive ^{13}C NMR and DEPT-90 signals and positive and negative DEPT-135 signals while avoiding potential noise artefacts.

Procedure

Building a database of predicted $\delta_{C\text{-}SDF}$. To create a DB of molecules and their δ_C that can be used by MixONat, the first step is to gather the structures of the compounds of interest (*e.g.* NPs previously identified in a genus or a botanical family). The easiest way consists in downloading them from various DBs accessible through subscription (*e.g.* SciFinder [25], Dictionary of Natural Products [20]) or from freely available ones (*e.g.* KNApSAcK [28], Universal Natural Products Database [35], LOTUS [36,37]). Once the individual files of each molecule (.mol, .cdx, .sk2) are collected in a structure data file (.sdf), their $\delta_{C\text{-SDF}}$ are predicted using a NMR prediction software under license (*e.g.* ACD NMR predictors [C,H]) [10,12] or not (*e.g.* nmrshiftdb2) [30,31]. From such DBs containing NPs together with their δ_C, the CTypeGen routine included in MixONat (Figure 3S) creates a suitable DB: it reads the SDF and sorts chemical shifts by carbon type. A new SDF is then created. The latter contains, for each compound of the DB, the predicted δ_C values organized as methyl, methylene, methine or quaternary carbons. The creation of such a DB is required for the MixONat algorithm to work properly [38].

Specific DBs. Lamiaceae DB2 was built by searching for compounds described in the Lamiaceae family on SciFinder, resulting in a database of 982 NPs. δ_C were predicted using ACD NMR predictors (C,H). Lamiaceae DB3 contains the 958 NPs from Lamiaceae according to KNApSAcK. δ_C were also predicted using ACD NMR predictors (C,H). Finally, Lamiaceae DB4 contains the same 958 NPs but δ_C were predicted using nmrshiftdb2 and was automatically assembled using the KnapsackSearch (KS) program.

The KS program. KS, available for free from https://github.com/nuzillard/KnapsackSearch/, is a tool for the construction of focused NPs libraries that relate together structure, biological taxonomy, and predicted ^{13}C NMR data. In this context, a focused library is defined by a user-supplied list of organism genera, possibly related to a taxonomic family. As clearly stated by its name, KnapsackSearch is related to the KNApSAcK project [28]. Searching in KNApSAcK for an organism according to its genus returns a list of pairs constituted by the organism's binomial name and by the KNApSAcK compound identifier. Searching in KNApSAcK for a compound identifier returns structural descriptors of this compound. These two types of searches are combined by KS to associate a set of compounds related to a list of genera with the organisms in which they have been reported. Running KS with a set of genera as input results in an SDF in which 2D stereo-aware structures are derived from the SMILES chains stored in KNApSAcK using the cheminformatics toolkit RDKit [39]. The final SDF contains tags that define the molecular properties such as the compound's name, its molecular formula, molecular weight, CAS registry number, InChI key, InChI code, SMILES chain, KNApSAcK identifier, the associated list of organism binomial names, and the calculated NMR data. The latter associate each carbon atom index with the ^{13}C NMR chemical shift predicted by nmrshiftdb2 [30]. KS is written as a collection of python scripts and is run from the command line interface. Assuming that the list of the genera from the Lamiaceae family is stored in a file named lamiaceae_genera.txt, the command "python process lamiaceae" automatically produces an SDF named lamiaceae_knapsack.sdf. It should also be noted that the PNMRNP DB is now available and can be used to create DBs based on chemotaxonomic or phytochemical criteria. It consists of a SDF file that reports to date the structure, properties and classification of 211 280 NPs as well as their predicted $\delta_{C\text{-}SDF}$ using ACD/Labs C+H NMR Predictors and DB [40,41].

NMR data export: The peak list and intensity data obtained from each spectrum were exported as a .csv file using Microsoft Excel (Microsoft) software and used as an input file in MixONat software. The file consists of a list of δ_C ordered in decreasing order associated with their intensities on the same line, separated by a comma.

Supporting information

Additional figures including GC-MS chromatogram, NMR spectra, screenshots of MixONat tabs and obtained results as well as practical processes are available in supporting information. A table comparing predicted and experimental δ_C for each major monoterpene of peppermint EO is also included.

Example datasets. For training purposes, all NMR spectra (fid and .mnova), peak lists (.csv) files and databases (.sdf) used in the present paper are accessible in supporting information.

Acknowledgements

A.B. gratefully acknowledges the Ministère de l'Enseignement Supérieur, de la Recherche et de l'Innovation (MESRI) for his PhD grant.

Conflicts of Interest

The authors declare no conflict of interest.

Authors' Contributions: This article is based on the results obtained by A.B., supervised by S.D. and P.R. during his PhD thesis. A.B., S.D., D.B., J.-M.N. and P.R. performed algorithms writing and/or analyses. F.T. provided the experimental spectral data of the peppermint monoterpenes. A.B., S.D., J.-M.N. and P.R. prepared the figures and tables and wrote the manuscript together. All authors discussed the results from the experiments and commented on the manuscript.

References

1. Tilburt JC, Kaptchuk TJ. Herbal medicine research and global health: an ethical analysis. Bulletin of the World Health Organization 2008; 86: 577-656
2. Eisenberg DM, Davis RB, Ettner SL, Appel S, Wilkey S, Van Rompay M, Kessler RC. Trends in alternative medicine use in the United States, 1990-1997. Results of a Follow-up National Survey. JAMA 1998; 280: 1569-1575
3. Welz AN, Emberger-Klein A, Menrad K. The importance of herbal medicine use in the German health-care system: prevalence, usage pattern, and influencing factors. BMC Health Services Research 2019; 19: 952
4. Bruneton J. Pharmacognosy : Phytochemistry, medicinal plants. 2nd ed. Londres: Tec & Doc Lavoisier; 2008
5. World Health Organization. Quality control methods for herbal materials. Geneva: World Health Organization; 2011
6. Wang M, Carver JJ, Phelan VV, Sanchez LM, Garg N, Peng Y, Nguyen DD, Watrous J, Kapono CA, Luzzatto-Knaan T, Porto C, Bouslimani A, Melnik AV, Meehan MJ, Liu W-T, Crüsemann M, Boudreau PD, Esquenazi E, Sandoval-Calderón M, Kersten RD, Pace LA, Quinn RA, Duncan KR, Hsu C-C, Floros DJ, Gavilan RG, Kleigrewe K, Northen T, Dutton RJ, Parrot D, Carlson EE, Aigle B, Michelsen CF, Jelsbak L, Sohlenkamp C, Pevzner P, Edlund A, McLean J, Piel J, Murphy BT, Gerwick L, Liaw C-C, Yang Y-L, Humpf H-U, Maansson M, Keyzers RA, Sims AC, Johnson AR, Sidebottom AM, Sedio BE, Klitgaard A, Larson CB, Boya P CA, Torres-Mendoza D, Gonzalez DJ, Silva DB, Marques LM, Demarque DP, Pociute E, O'Neill EC, Briand E, Helfrich EJN, Granatosky EA, Glukhov E, Ryffel F, Houson H, Mohimani H, Kharbush JJ, Zeng Y, Vorholt JA, Kurita KL, Charusanti P, McPhail KL, Nielsen KF, Vuong L, Elfeki M, Traxler MF, Engene N, Koyama N, Vining OB, Baric R, Silva RR, Mascuch SJ, Tomasi S, Jenkins S, Macherla V, Hoffman T, Agarwal V, Williams PG, Dai J, Neupane R, Gurr J, Rodríguez AMC, Lamsa A, Zhang C, Dorrestein K, Duggan BM, Almaliti J, Allard P-M, Phapale P, Nothias L-F, Alexandrov T, Litaudon M, Wolfender J-L, Kyle JE, Metz TO, Peryea T, Nguyen D-T, VanLeer D, Shinn P, Jadhav A, Müller R, Waters KM, Shi W, Liu X, Zhang L, Knight R, Jensen PR, Palsson BØ, Pogliano K, Linington RG, Gutiérrez M, Lopes NP, Gerwick WH, Moore BS, Dorrestein PC, Bandeira N. Sharing and community curation of mass spectrometry data with Global Natural Products Social Molecular Networking. Nature Biotechnol 2016; 34: 828
7. Mimica-Dukić N, Božin B, Soković M, Mihajlović B, Matavulj M. Antimicrobial and antioxidant activities of three *Mentha* species essential oils. Planta Med 2003; 69: 413-419
8. Bakiri A, Hubert J, Reynaud R, Lanthony S, Harakat D, Renault JH, Nuzillard JM. Computer-aided ^{13}C NMR chemical profiling of crude natural extracts without fractionation. J Nat Prod 2017; 80: 1387-1396
9. Hubert J, Nuzillard JM, Purson S, Hamzaoui M, Borie N, Reynaud R, Renault JH. Identification of natural metabolites in mixture: a pattern recognition strategy based on ^{13}C NMR. Analytical chemistry 2014; 86: 2955-2962
10. Bruguière A, Derbré S, Coste C, Le Bot M, Siegler B, Leong ST, Sulaiman SN, Awang K, Richomme P. ^{13}C-NMR dereplication of *Garcinia* extracts: Predicted chemical shifts as reliable databases. Fitoterapia 2018; 131: 59-64
11. Bruguière A, Derbré S, Dietsch J, Leguy J, Rahier V, Pottier Q, Bréard D, Suor-Cherer S, Viault G, Ray A-ML, Saubion F, Richomme P. MixONat, a software for mixtures dereplication based on ^{13}C NMR experiments. Anal Chem 2020; 92: 8793-8801

12. ACD/Labs. NMR Spectroscopy Software, https://www.acdlabs.com/products/adh/nmr/ (accessed March 2019).

13. Robien W. CSEARCH Spectral Similarity Search with Ranking, http://c13nmr.at/planta_medica/eval.php (accessed March 2021).

14. Bruguière A, Derbré S. MixONat. ^{13}C-NMR based dereplication software, https://sourceforge.net/projects/mixonat/ (accessed November 2020).

15. Council of Europe, European Directorate for the quality of medicines and healthcare. European Pharmacopoeia. 10th ed. Strasbourg: Council of Europe; 2019

16. Xu T, Gherib M, Bekhechi C, Atik-Bekkara F, Casabianca H, Tomi F, Casanova J, Bighelli A. Thymyl esters derivatives and a new natural product modhephanone from *Pulicaria mauritanica* Coss. (Asteraceae) root oil. Flavour Fragr J 2015; 30: 83-90

17. Ouattara ZA, Boti JB, Ahibo AC, Sutour S, Casanova J, Tomi F, Bighelli A. The key role of ^{13}C NMR analysis in the identification of individual components of *Polyalthia longifolia* leaf oil. Flavour Fragr J 2014; 29: 371-379

18. Baldovini N, Tomi F, Casanova J. Identification and quantitative determination of furanodiene, a heat-sensitive compound, in essential oil by ^{13}C-NMR. Phytochem Anal 2001; 12: 58-63

19. Zavahir JS, Smith JSP, Blundell S, Waktola HD, Nolvachai Y, Wood BR, Marriott PJ. Relationships in Gas Chromatography-Fourier Transform Infrared Spectroscopy. Comprehensive and multilinear analysis. Separations 2020; 7: 27

20. ChemNetBase. Dictionary of Natural Products 29.1. http://dnp.chemnetbase.com (Accessed November 2020).

21. Pupier M, Nuzillard J-M, Wist J, Schlörer NE, Kuhn S, Erdelyi M, Steinbeck C, Williams AJ, Butts C, Claridge TDW, Mikhova B, Robien W, Dashti H, Eghbalnia HR, Farès C, Adam C, Kessler P, Moriaud F, Elyashberg M, Argyropoulos D, Pérez M, Giraudeau P, Gil RR, Trevorrow P, Jeannerat D. NMReDATA, a standard to report the NMR assignment and parameters of organic compounds. Magn Reson Chem 2018; 56: 703-715

22. Tomi F, Bradesi P, Bighelli A, Casanova J. Computer aided identification of individual components of essential oils using carbon ^{13} NMR spectroscopy. Magn Reson Anal 1995; 1: 25–34

23. Lanfranchi DA, Blanc MC, Vellutini M, Bradesi P, Casanova J, Tomi F. Enantiomeric differentiation of oxygenated p-menthane derivatives by ^{13}C NMR using Yb(hfc)3. Magn Reson Chem 2008; 46: 1188-1194

24. Rutz A, Dounoue-Kubo M, Ollivier S, Bisson J, Bagheri M, Saesong T, Ebrahimi SN, Ingkaninan K, Wolfender J-L, Allard P-M. Taxonomically informed scoring enhances confidence in natural products annotation. Front Plant Sci 2019; 10: 1-15

25. CAS. SciFinder, https://www.cas.org/products/scifinder (accessed March 2019).

26. Elsevier. Reaxys, https://www.elsevier.com/solutions/reaxys (accessed November 2020).

27. Nuzillard JM. KnapsackSearch. Automated data search in the KNApSAcK database, https://github.com/nuzillard/KnapsackSearch (accessed June 2020).

28. Afendi FM, Okada T, Yamazaki M, Hirai-Morita A, Nakamura Y, Nakamura K, Ikeda S, Takahashi H, Altaf-Ul-Amin M, Darusman LK, Saito K, Kanaya S. KNApSAcK family databases: Integrated metabolite–plant species databases for multifaceted plant research. Plant and Cell Physiology 2011; 53: e1-e1

29. KNApSAcK. KNApSAcK core System, http://www.knapsackfamily.com/knapsack_core/top.php (accessed November 2020).

30. Kuhn S. NMRShiftDB2, https://nmrshiftdb.nmr.uni-koeln.de/ (accessed November 2020).

31. Kuhn S, Schlörer NE. Facilitating quality control for spectra assignments of small organic molecules: nmrshiftdb2 – a free in-house NMR database with integrated LIMS for academic service laboratories. Magn Reson Chem 2015; 53: 582-589

32. Dona AC, Kyriakides M, Scott F, Shephard EA, Varshavi D, Veselkov K, Everett JR. A guide to the identification of metabolites in NMR-based metabonomics/metabolomics experiments. Comput Struct Biotechnol J 2016; 14: 135-153

33. John Wiley & Sons Inc. SpectraBase; https://spectrabase.com/ (accessed January 2021).

34. Gottlieb HE, Kotlyar V, Nudelman A. NMR chemical shifts of common laboratory solvents as trace impurities. J Org Chem 1997; 62: 7512-7515

35. Gu J, Gui Y, Chen L, Yuan G, Lu H-Z, Xu X. Use of natural products as chemical library for drug discovery and network pharmacology. PLoS One 2013; 8: e62839

36. LOTUS. the naturaL prOducTs occUrrences databaSe, https://lotus.naturalproducts.net/ (accessed March 2021).

37. Rutz A, Sorokina M, Galgonek J, Mietchen D, Willighagen E, Graham J, Stephan R, Page R, Vondrášek J, Steinbeck C, Pauli GF, Wolfender J-L, Bisson J, Allard P-M. Open natural products research: Curation and dissemination of biological occurrences of chemical structures through wikidata. bioRxiv 2021, DOI: 10.1101/2021.02.28.433265: 2021.2002.2028.433265

38. Please note that the program has been optimized for DBs created with ACD/Labs and hence may not work properly with a different type of DB.

39. Landrum G. An overview of the RDKit, https://www.rdkit.org/docs/Overview.html (accessed November 2020).

40. Lianza M, Leroy R, Machado Rodrigues C, Borie N, Sayagh C, Remy S, Kuhn S, Renault J-H, Nuzillard J-M. The three pillars of natural product dereplication. Alkaloids from the bulbs of *Urceolina peruviana* (C. Presl) J.F. Macbr. as a preliminary test case. Molecules 2021; 26: 637

41. Nuzillard JM, Leroy R, Kuhn S. Predicted carbon-13 NMR data of Natural Products (PNMRNP), https://zenodo.org/record/4420849#.YFNMy9zjJPY (accessed January 2021).

42. ChemAxon. Marvin. MarvinView - View your molecules, https://chemaxon.com/products/marvin (accessed March 2019).

43. Ho D. Notepad++. What is Notepad++, https://notepad-plus-plus.org/ (accessed March 2019).

Table and figures' legends

Table 1. Major monoterpenes in peppermint essential oil determined by GC-MS as well as their ranks using ^{13}C-NMR based dereplication, the MixONat software and various DBs of experimental or predicted $\delta_{\text{C-SDF}}$.

Figure 1. Schematic representation of the ^{13}C-NMR based dereplication process. MixONat (orange, middle) requires appropriate DBs including either experimental δ_{C} or freely available and/or commercial predicted $\delta_{\text{C-SDF}}$ datasets (green, right) as well as peak lists (.csv files) exported from experimental data (blue, left). *Optional

Figure 2. Structures and atom numbering of major monoterpenes from peppermint EO identified using GC-MS

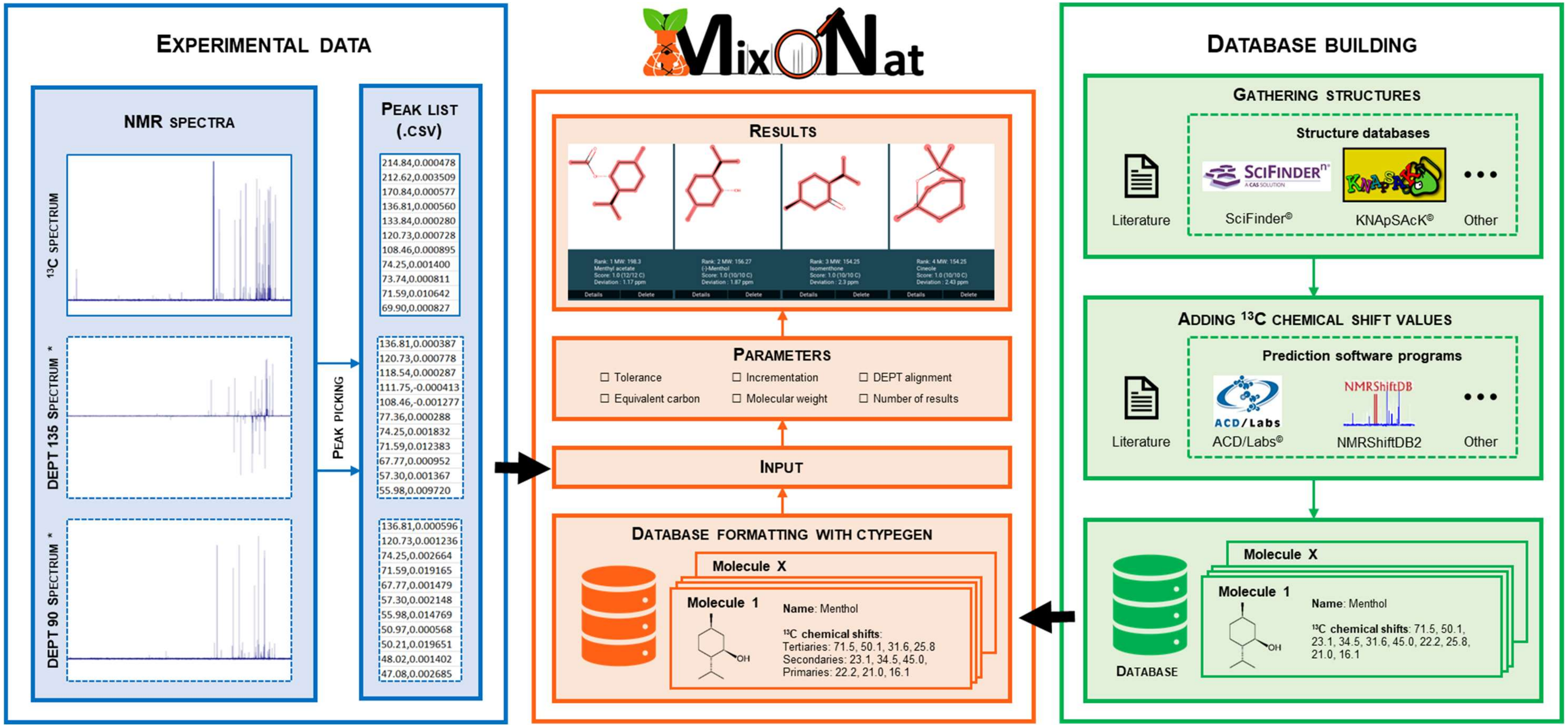
Tables

Table 1. Major monoterpenes in peppermint essential oil determined by GC-MS as well as their ranks using ^{13}C-NMR based dereplication, the MixONat software and various DBs of experimental or predicted $\delta_{C\text{-}SDF}$.

	Compound	**Amount**	**Peppermint EO DB1**	**Lamiaceae DB2**	**Lamiaceae DB3**	**Lamiaceae DB4**
	name	**(%)**	**Experimental δ_C**	**SciFinder / ACD**	**Knapsack / ACD**	**Knapsack / nmrshiftDB**
		GC-MS	**30 NPs**	**952 NPs**	**958 NPs**	**958 NPs**
			(High | Low threshold values)	(High threshold value)	(High threshold value)	(High threshold value)
1	Menthol	41.6	3 | 1	2 and 6	4	11
2	Menthone	27.0	6 | 4	11	7 and 8	2 and 3
3	1,8-Cineole	6.1	1 | 2	4	12	1
4	Menthyl acetate	6.1	2 | 3	1 and 8	2 and 3	37
5	Isomenthone	5.7	4 | 5	3	5 and 6	4 and 5
6	Limonene	1.9	7 | 7 (Score 0.9)	10	10 and 11	7 and 8
7	Menthofurane	2.9	12 | 8 (score 0.9)	207	40	58
8	Neomenthol	[a]	5 | 6	[b]	1	10

[a]Not detected; [b]Not in Lamiaceae DB2 | Neomenthol acetate suggested in rank 7 (score 1.0);

Figures



```
214.84,0.000478
212.62,0.003509
170.84,0.000577
136.81,0.000560
133.84,0.000280
120.73,0.000728
108.46,0.000895
74.25,0.001400
73.74,0.000811
71.59,0.010642
69.90,0.000827
```

```
136.81,0.000387
120.73,0.000778
118.54,0.000287
111.75,-0.000413
108.46,-0.001277
77.36,0.000288
74.25,0.001832
71.59,0.012383
67.77,0.000952
57.30,0.001367
55.98,0.009720
```

```
136.81,0.000596
120.73,0.001236
74.25,0.002664
71.59,0.019165
67.77,0.001479
57.30,0.002148
55.98,0.014769
50.97,0.000568
50.21,0.019651
48.02,0.001402
47.08,0.002685
```

Rank: 1 MW: 198.3
Menthyl acetate
Score: 1.0 (12/12 C)
Deviation : 1.17 ppm

Rank: 2 MW: 156.27
(-)-Menthol
Score: 1.0 (10/10 C)
Deviation : 1.87 ppm

Rank: 3 MW: 154.25
Isomenthone
Score: 1.0 (10/10 C)
Deviation : 2.3 ppm

Rank: 4 MW: 154.25
Cineole
Score: 1.0 (10/10 C)
Deviation : 2.43 ppm

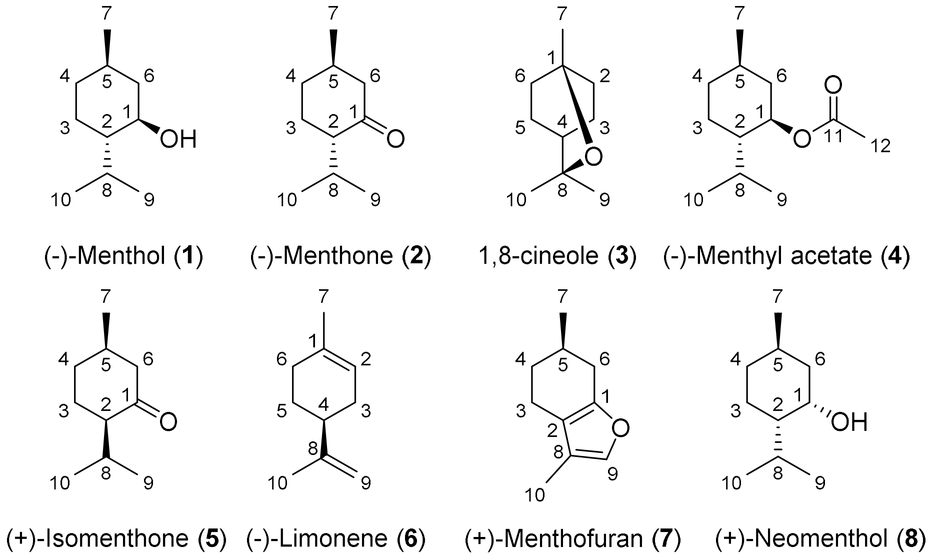
Name: Menthol

^{13}C chemical shifts:
Tertiaries: 71.5, 50.1, 31.6, 25.8
Secondaries: 23.1, 34.5, 45.0,
Primaries: 22.2, 21.0, 16.1

Name: Menthol

^{13}C chemical shifts: 71.5, 50.1, 23.1, 34.5, 31.6, 45.0, 22.2, 25.8, 21.0, 16.1

Figure 1. Schematic representation of the ^{13}C-NMR based dereplication process. MixONat (orange, middle) requires appropriate DBs consisting in either experimental δ_C or free available and/or commercial predicted $\delta_{C\text{-}SDF}$ datasets (green, right) as well as peak lists (.csv files) exported from experimental data (blue, left). *Optional


(-)-Menthol (**1**) (-)-Menthone (**2**) 1,8-cineole (**3**) (-)-Menthyl acetate (**4**)

(+)-Isomenthone (**5**) (-)-Limonene (**6**) (+)-Menthofuran (**7**) (+)-Neomenthol (**8**)

Figure 2. Structures and atom numbering of major monoterpenes from peppermint EO

Chart 1. Useful software and algorithms

- **KnapsackSearch** [29]. This program allows to export the NPs previously isolated from a set of genera as a SDF together with associated data (*e.g.* name, molecular weight, sources, predicted $\delta_{C\text{-}SDF}$ using nmrshiftdb2)
- **MarvinView** [42]. This free advanced chemical viewer allows to visualize compounds from a database (SDF) as well as their associated data (*e.g.* molecular weight, NMR predicted shifts)
- **Notepad++** [43]. This is a free source code editor that supports several languages. It is useful to create DBs suitable for MixONat software.
- **MixONat** [14]. It allows the dereplication of natural products mixtures using ^{13}C-NMR and DEPT data as well as experimental or predicted δ_C DB. It displays interactive results that can be refined based on the user's phytochemical knowledge.