Identification of known natural products by querying a carbon-13 nuclear magnetic resonance database-an assessment

Jean-Marc Nuzillard

Identification of known natural products by querying a carbon-13 nuclear magnetic resonance database — an assessment

Author: Jean-Marc Nuzillard (ORCID: 0000-0002-5120-2556)

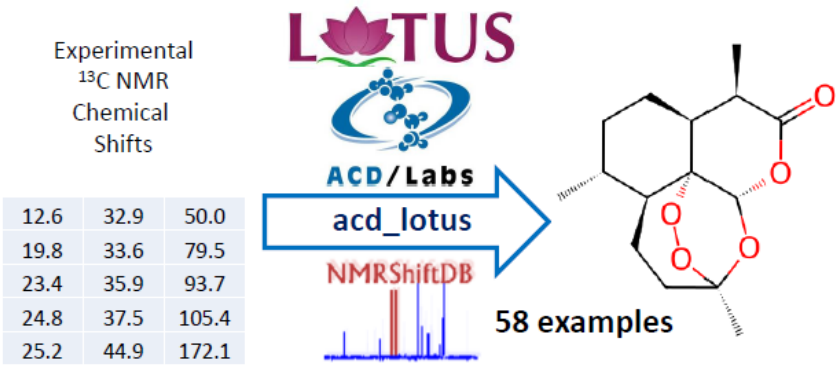
Correspondence: jm.nuzillard@univ-reims.fr

Affiliation: Université de Reims Champagne-Ardenne, CNRS, ICMR UMR 7312, 51097 Reims, France

Keywords: Natural Products ; Nuclear Magnetic Resonance ; Dereplication ; Databases

Abstract: The quick identification of known organic low molecular weight compounds, also known as structural dereplication, is a highly important task in the chemical profiling of natural resource extracts. To that end, a method that relies on carbon-13 nuclear magnetic resonance (NMR) spectroscopy, elaborated in earlier works of the author's research group, requires the availability of a dedicated database that establishes relationships between chemical structures, biological and chemical taxonomy, and spectroscopy. The construction of such a database, called acd_lotus, was reported earlier and its usefulness was illustrated by three examples only. This article presents the results of structure searches carried out starting from 58 carbon-13 NMR data sets recorded on compounds selected in the metabolomics section of the biological magnetic resonance bank (BMRB). Two compound retrieval methods were employed. The first one involves searching in the acd_lotus database using a commercial software. The second one operates through the freely accessible web interface of the nmrshiftdb2 database, that includes the compounds present in acd_lotus and many others. The two structural dereplication methods have proved to be efficient and can be used together in a complementary way.

Graphical abstract:



Introduction

Natural products (NPs) chemistry and metabolomics share almost the same tools, including those for the identification of chemical compounds within complex mixtures.[1] Nuclear magnetic resonance (NMR) spectroscopy plays an important role in these two activities. NMR is well-suited to the identification of known compounds and to the structure elucidation of unknown ones.[2] ^{13}C NMR alone has been reported as an efficient method for the identification of known compounds of small molecular weight such as NPs when the available sample amount is not limiting.[3,4] Avoiding to duplicate the identification work for already reported compounds is known as structural dereplication.[5,6] Collecting and associating NP molecular structures, chemical or biological taxonomic information, and spectroscopic data are necessary to build databases for structural dereplication.[7] The creation of exhaustive catalogs of NP structures was undertaken only recently and lead to the COlleCtion of Open Natural ProdUcTs (COCONUT) and to the natural prOducTs occUrrences databaSe (LOTUS).[8,9] LOTUS collects publicly accessible structures and natively includes the related chemical and biological taxonomy data. The spectroscopic pillar of dereplication was given to LOTUS by the adjunction of predicted NMR chemical shift values. Choosing predicted values instead of experimental ones[3] is the result of the scarce availability of the latter.[10] The acd_lotus database was created using LOTUS structural and taxonomic data and the Advanced Chemistry Development, Inc. (ACD/Labs) ^{13}C NMR chemical shift validation tool for prediction.[11] The process that led to the creation of acd_lotus was presented in a recent article.[5] Its use was illustrated by the retrieval of only three structures from published sets of experimental chemical shift values.[5] The present article reports an assessment of the same method through the attempted retrieval of 56 structures of metabolites.

Material and methods

All calculations were carried out on a DELL Precision 3530 laptop computer with 16 GB of RAM memory and an Intel® Core™ i5-8400H CPU @ 2.5 GHz running Windows 10 Education version 20H2. The ACD/C+H NMR Predictors and DB 2021.1.0 software was purchased from ACD/Labs (Toronto, Ontario, Canada).

The metabolites involved in the present assessment task were selected from the Biological Magnetic Resonance Bank (BMRB).[12] The home page of its web site proclaims: "BMRB collects, annotates, archives, and disseminates spectral and quantitative data derived from NMR spectroscopic investigations of biological macromolecules and metabolites". A dedicated web page leads to a list of "biologically relevant small molecules in the BMRB".[13] This page contains hypertext links that quickly leads to chemical structures and to the corresponding NMR data. Structures, NMR spectra, and related metadata are available in the Self-defining Text Archival and Retrieval (STAR) format as NMR-STAR files.[14] The simplicity of fetching structures and experimental NMR data of small molecules at BMRB motivated the focusing on this data source in view of the planned evaluation work. The metabolites subset in BMRB include the primary and secondary ones, as well as xenobiotics. A selection of 56 secondary metabolites belonging to diverse chemical classes was arbitrarily constituted to assess the structural dereplication process based on acd_lotus. Two structure retrieval assays were carried out for each experimental chemical shift set drawn from BMRB, one using the ACD/Labs software, and the other one with nmrshiftdb2.[15,16]

The acd_lotus database is provided in Structure-Data File (SDF) format[17] and can be downloaded from zenodo.org[18] and imported in an ACD/Labs database file, possibly named acd_lotusv9.NMRUDB. Querying this database was achieved by copying the targeted ^{13}C NMR chemical shift values in the dedicated text entry area, selecting by default a tolerance for a 4 ppm difference, and searching for at

least as many carbon atoms as there are resonances to find. Duplicated experimental chemical shift values, occurring by accident or by symmetry, were left duplicated in the list of queried values. The search options "Search through Unambiguous Assignment" and "Do not Match One Chemical Shift for Several Shift Queries" well left inactive. Results were sorted according to "HQI Based on Minimal Distances". Each search produces a list of structure proposals whose length, or number of hits, is noted. A structure in LOTUS and in acd_lotus is associated to a Wikidata reference, or Q-Id, so that the validity of a query result can be easily checked.[19] Wikidata is a wide scope knowledge graph that contains nodes for many chemical compounds, including those present in LOTUS. Validity of a query result was defined by identity of compound names and InChI keys.[20] A query may fail, as indicated by the retrieval of structures in which carbon atoms were not associated to ^{13}C NMR resonances. Finding the expected structure can then be achieved by enlarging the chemical shift tolerance for the matching between queried and predicted values. Each compound selected in BMRB was associated to a list of queries to acd_lotus, often containing only a single element and characterized by the tolerance value. A query was identified by a minimal number of queried chemical shifts, a tolerance value, a number of hits, and a list of results. A result was characterized by a Boolean (true or false value) for compound matching success, a rank in the list of proposals, a link to Wikidata for the proposed structures, the InChI and the InChI key identifiers from Wikidata, and a remark as free text.

The structures and predicted NMR data in acd_lotus were made available for compound search in the nmrshiftdb2 web site.[5] The lists of chemical shifts drawn from BMRB that were used for searches in acd_lotusv9.NMRUDB were also submitted for searches in nmrshiftdb2. The latter operate completely unsupervised, meaning that no number of expected chemical shift matchings and no tolerance value has to be provided. Each search result from nmrshiftdb2 was characterized by a spectrum similarity index as a percentage value, a rank in the list of proposals, a Boolean value indicating compound matching success, a link to Wikidata, the InChI and the InChI key identifiers from Wikidata, and a remark as free text.

Each compound selected in BMRB was characterized in a formal way by its name, its Chemical Abstract Service Registry Number (CAS RN), its SMILES, InChI, and InChIKey identifiers, the link to the associated NMR-STAR file, the list of its ^{13}C NMR chemical shift values, the sample supplier, the compound reference by the supplier, and the characteristics, as stated here above, of the queries submitted to acd_lotus through the ACD/Labs software and to nmrshiftdb2 as well as the outcome of these queries.

Results encoding was achieved in practice by means of the formalism brought by the JavaScript Object Notation (JSON).[21] The adoption of JSON for data representation was motivated by its simplicity of being read and written by humans and computers. Extensible Markup Language (XML) was found less practical than JSON for this purpose.[22]

The structures of all compounds were collected in a single SDF file in which the 2D atomic coordinates were calculated from SMILES chains using the RDKit library of cheminformatic functions.[23] Structure depictions in the Portable Network Graphics (PNG) format of all compounds were also created by means of RDKit.

Results

Scheme 1 shows the structures, names, and numbers of all the compounds selected for the assessment of acd_lotus as a tool for structural dereplication. Table 1 reports the results of the tests carried out on these compounds, presented as the ranking of the targeted compound in the list of the proposals returned by the compound search tools in ACD/Labs and nmrshiftdb2. More details on compounds,

quieries, and answers to queries are available as Supplementary material from zenodo.org. Among the 58 tested structures, 39 were placed in first position by the compound search tool of the ACD/Labs software operating on acd_lotus on the basis of chemical shift list similarity. Moreover, in 8 cases an increase of the chemical shift tolerance either to 6 or to 8 ppm was required to obtained this result. Compound search in nmrshiftdb2 lead to 32 rankings of the targeted compounds in first position. Target ranking in first position by the two search tools simultaneously occurred 28 times.

Structural dereplication was considered to have failed in the three following cases. 4-Ipomeanol **34** is the only compound among the 58 selected ones from BMRB that is present neither in acd_lotus nor in nmrshiftdb2. Cytidine **18** is present in these databases only as a tautomeric form originating from an improper decoding of an InChI string from LOTUS, a form that is not considered as a correct answer to the compound identification queries. The same decoding problem arose for uracil **58**. However, the nmrshiftdb2 database contains a structure of uracil that is not related to LOTUS and for which NMR data are experimental, thus allowing for proper structural dereplication. Particular situations that were encountered during the present study are discussed in more details hereafter. Information about compound structures were searched in public sources Wikidata[19] and PubChem[24] as well as in SciFinder^n.[25]

Abscissic acid **1** was analyzed by BMRB using a sample of natural (+)-abscisic acid. Its molecule contains a single asymmetric carbon atom and two double carbon-carbon bonds. The first proposal from ACD/Labs, obtained with a chemical shift tolerance of 6 ppm, has no defined configuration of the asymmetric center and correct configurations of the double bonds. This proposal is considered as correct because NMR is unable to distinguish between enantiomers in standard spectrum recording conditions. Structures completely lacking of asymmetric center configuration data are present in acd_lotus and are referred to as flat structure. The second proposal is a single enantiomer but the geometry of one the double bonds is left unspecified. Such a structure is not considered as a valid answer to the query. The third proposal is as valid as the first one because all the elements of the 3D geometry are properly defined. Nmrshiftdb2 provides an answer to the query that is similar to the one by the ACD/Labs software.

(-)-Alpha-santonin **2** was retrieved by the ACD/Labs search tool at first rank. Its Wikidata Q-Id, Q413166, according to acd_lotus, leads to a compound simply named santonin in Wikidata but with (-)-alpha-Santonin as synonym. Nmrshiftdb2 ranks at the first place Q105328952, the enantiomer of Q413166 named alpha-Santonin in Wikidata, and Q413166 at the second place. The structures of the two enantiomers are considered valid for dereplication purpose. The presence of the two enantiomers in LOTUS and in Wikidata is apparently surprising, as the natural occurence of only one of the two enantiomers is the most likely hypothesis. The only santonin known to SciFinder^n is Q413166, in which the four asymmetric centers are of the (S) kind. Compound naming in Wikidata appears as possibly misleading in particular cases.

The case of ascochitin **4** illustrates the confusion in compound identification brought by the existence of tautomeric forms. The molecular structure of ascochitin contains a single asymmetric center and a series of conjugated double bonds. The first proposal from the ACD/Labs software is a tautomer, Q110168575 in Wikidata, of the structure declared by BMRB and confirmed by SciFinder^n on the basis of the trivial name ascochitine. The second proposal, Q77573394, has its single and double bonds at the expected positions and would have been the correct proposal if its asymmetric center configuration were defined. This proposal was accepted because the corresponding ^{13}C NMR spectrum matches the one of the enantiomerically pure ascochitine. The expected structure is neither present in acd_lotus nor in nmrshiftdb2 but it is referenced Q27275562 in Wikidata. Even though the InChI code was created in order to solve compound identification issues, it does not create different

identifiers for compound related by non-trivial tautomerism such as Q27275562 and Q110168575. Retrieving a tautomer is not considered in this study as a valid result for dereplication.

Aspergillic acid **5** derives from two aminoacid residues, leucine and isoleucine. SciFinder^n indicates that the absolute configuration of the asymmetric center in the side-chain of isoleucine is not defined, even though it reports a non-zero optical rotation. Interestingly this database contains two entries, one for Aspergillic acid with a Registry Number (RN) 490-02-8, drawn with an explicit pentavalent nitrogen atom and one for Aspergillic acid, DL- of RN 22810-67-9. PubChem and the NCI catalog[26] do not indicate a defined chirality neither, as found for Q4807880 in Wikidata. Moreover, the data sheet of aspergillic acid in BMRB reports a defined configuration that corresponds to the one of Q105120934, retrieved and ranked first by the ACD/Labs software and nmrshiftdb2, even though the process by which this configuration was defined is not clear. A hypothesis can be proposed for the spontaneous apparition of a new chiral center: a flat 2D structure as the one derived from a non-isomeric SMILES chain can be transformed into a 3D structure by means of a cheminformatic toolkit so that an asymmetric center is created with a randomly assigned absolute configuration. Transforming back this 3D structure into textual chemical descriptors then initiates the propagation of a fake information.

Atropine **6** is another case in which configuration information plays an important role. The sample analyzed by BMRB is (+/-)-atropine, also named (+/-)-hyoscyamine in Wikidata and ranked first as such by the ACD/Labs compound search tool. The molecule of atropine contains four asymmetric carbons but the planar symmetry of the non-aromatic tropine part makes a configuration inversion at the position close to the aromatic ring sufficient to create enantiomers. In BMRB and in PubChem the geometry of the tropine part is incorrect as being reported in the *exo* (or β) form instead of the *endo* (or α) form. Moreover, BMRB fully defines the configuration of all the asymmetric center, in contradiction with the data sheet provided by the sample supplier. The identification of atropine is complicated by the fact that it is a scalemic natural compound, possibly occuring neither in enantiopure nor in racemic form.

Betulin **10** in its flat form was ranked first by the ACD/Labs and nmrshiftdb2 search tools. The exact structure was ranked at the third place by ACD/Labs and at the twelfth by nmrshiftdb2 but the fact that many stereoisomers were reported for this molecules makes the ranking not significant.

Bicuculline is the name in Wikidata for bicuculline **12** and for its flat form. Two enantiomers of it are present in Wikidata but SciFinder^n knows only about one.

Cytidine **18** is not present in acd_lotus in the appropriate di-amide tautomeric form. The predicted ^{13}C NMR chemical shifts predicted from the di-iminol tautomer are too wrong to allow for structure retrieval from experimental chemical shifts. The nmrshiftdb2 search tool is not constrained by a user-supplied chemical shift tolerance and finds the cytidine di-iminol tautomer at rank 87. In that particular case, prediction was carried out using the HOSE code approach, as explained in [5].

Eburnamonine **21** was isolated in its two enantiomeric forms, depending on the nature of the investigated plant. BMRB analyzed (-)-eburnamonine, strangely written (~)-eburnamonine in the BMRB web site and designated as vinburnine in Wikidata. Its enantiomer is also present in Wikidata and is named eburnamonin. The two compound search tools ranked at first place one or the other enantiomer.

Nmrshiftdb2 ranked ergosterol 23 at position 10 but compounds ranked from position 1 to position 10 bear the same similarity index to targeted chemical shifts, thus making ranking meaningless.

The InChI code of eucalyptol **25** in BRMB erroneously indicates the presence of two asymmetric centers. The correct identifiers were substituted to the proposed ones and reported in the JSON file. InChI and SMILES for fastigillin B **26** (possibly also written fastigilin) are wrong in BMRB but match together; they were replaced by the correct ones from PubChem. SMILES in BMRB is wrong for harmalol **31** and replaced in the JSON file by the one in PubChem. The InChI, InChI Key and SMILES in BMRB are wrong for himbacine **32** and were replaced in the JSON file by the ones in PubChem.

4-Ipomeanol **34** is present neither in LOTUS nor in nmrshiftdb2 and was therefore not retrieved in this study. The absolute configuration of the asymmetric center is defined as being (R) in BMRB only and nowhere else. The exact configuration might have never been determined, as suggested by SciFinder^{n}. Flat 4-ipomeanol is present in Wikidata as Q27291230.

The InChI of methyl jasmonate **43** in BMRB is wrong, with a *trans* double bond, but the InChI key is correct. The SMILES in BMRB does not include the double bond geometry. The SMILES, InChI and InChI key from Pubchem, identical to those in Wikidata, were copied in the JSON file.

Nmrshiftdb2 ranked raffinose **50** at the first place because it contains experimental NMR data for this compound that match well with those reported by BMRB. However, the InChI key associated to the retrieved compound does not match with the one of raffinose. The origin of this mismatch lies in the drawing of the molecule that contains sugar rings in chair form, thus preventing a correct automatic interpretation of the structure from 2D atomic coordinates. Sugars rings are best drawn as regular hexagons for database storage purpose, even though this results in ring conformation data loss.[27]

Rosmarinic **51** is ranked first by nmrshiftdb2 on the basis of experimental data. However, these data are related to a correct compound name but to a structure in which the configuration of the asymmetric center is not defined.

BMRB analyzed *trans*-nerolidol **57**, which is either a racemic compound or a compound for which the absolute configuration of its asymmetric center was not determined. The SMILES, InChI, and InChI key identifiers proposed by BMRB were replaced by those from PubChem derived from the flat structure with a correct geometry of the double bonds.

The correct di-amide tautomer of uracil **58** is not present in acd_lotus, which contains only the di-iminol form. The correct tautomer was found in nmrshiftdb2, ranked at the second place behind a non-natural boronic acid derivative, because the uracil data it contains do not originate from acd_lotus. Uracil in its di-amide form is Q182990 in Wikidata and the corresponding InChI is unexpectedly decoded by RDKit into the di-amide tautomer, even though a warning message is issued during decoding. Unexpectedly, stated in the previous sentence, refers to the usual InChI software behavior that produces iminol structures instead of amides. So, two different InChI strings may be converted into two different tautomers that in turn are back converted to the same InChI string. Using PerkinElmer ChemDraw or ACD/Labs ChemSketch interactive structure drawing software to carry out back and forth structure to InChI conversions produces the same result as the one obtained with RDKit.

Supplementary material

The JSON files and the structure depiction of all compounds were deposited at https://doi.org/10.5281/zenodo.8023745 .

Conclusion

The structures of the compounds arbitrarily selected by the author from the BMRB website for ^{13}C NMR based structural dereplication were on the average satisfactorily identified. Compound identification by means of the nmrshiftdb2 database only requires having access to a web browser. It allows to refine searches by assigning a number of directly bound hydrogen atoms to each targeted chemical shift value, a possibility that has been purposely ignored in the present study in order to maintain some fairness in the comparison with the ACD/Labs compound search tool. Moreover, nmrshiftdb2 contains compounds that do not result from acd_lotus data importation and for which experimental NMR data are available, thus enhancing its identification capability. Resorting on the acd_lotus database through the ACD/Labs compound search tool results often in a better ranking of the expected structures among the set of proposals, even though the reason for this cannot be easily investigated. This slight supplement in ranking quality takes place at the price of purchasing the necessary software.

References.

1. NMR in Metabolomics and Natural Products Research: Two Sides of the Same Coin. SL Robinette, R Brüschweiler, FC Schroeder, AS Edison; Acc. Chem. Res. 45 (2012) 288−297; DOI: 10.1021/ar2001606

2. Sherlock—A Free and Open-Source System for the Computer-Assisted Structure Elucidation of Organic Compounds from NMR Data, M Wenk, J-M Nuzillard, C Steinbeck; Molecules 28 (2023) 1448. DOI: 10.3390/molecules28031448

3. Identification of Natural Metabolites in Mixture: A Pattern Recognition Strategy Based on ^{13}C NMR. J Hubert, J-M Nuzillard, S Purson, M Hamzaoui, N Borie, R Reynaud, J-H Renault; Anal. Chem. 86 (2014) 2955−2962; DOI : 10.1021/ac403223f

4. MixONat, a Software for the Dereplication of Mixtures Based on ^{13}C NMR Spectroscopy. A Bruguière, S Derbré, J Dietsch, J Leguy, V Rahier, Q Pottier, D Bréard, S Suor-Cherer, G Viault, A-M Le Ray, F Saubion, P Richomme; Anal. Chem. 92 (2020) 8793−8801; DOI: 10.1021/acs.analchem.0c00193

5. Easy Structural Dereplication of Natural Products by Means of Predicted Carbon-13 Nuclear Magnetic Resonance Spectroscopy Data. S Kuhn, J-M Nuzillard; Chem. Methods 3 (2023) e202200054; DOI: 10.1002/cmtd.202200054

6. Dereplication strategies in natural product research: How many tools and methodologies behind the same concept? J Hubert, J-M Nuzillard, J-H Renault; Phytochem. Rev. 16 (2017) 55–95; DOI: 10.1007/s11101-015-9448-7

7. The Three Pillars of Natural Product Dereplication. Alkaloids from the Bulbs of *Urceolina peruviana* (C. Presl) J.F. Macbr. as a Preliminary Test Case. M Lianza, R Leroy, C Machado Rodrigues, N Borie, C Sayagh, S Remy, S Kuhn, J-H Renault, J-M Nuzillard; Molecules 26 (2021) 637. DOI: 10.3390/molecules26030637

8. COCONUT online: Collection of Open Natural Products database. M Sorokina, P Merseburger, K Rajan, MA Yirik, C Steinbeck; J. Cheminform. 13, (2021) 2; DOI: 10.1186/s13321-020-00478-9

9. The LOTUS initiative for open knowledge management in natural products research. A Rutz, M Sorokina, J Galgonek, D Mietchen, E Willighagen, A Gaudry, JG Graham, R Stephan, R Page, J Vondrášek,

C Steinbeck, GF Pauli, J-L Wolfender, J Bisson, P-M Allard; (2022) eLife 11 (2022) e70780; DOI: 10.7554/eLife.70780

10. For chemists, the AI revolution has yet to happen. Nature 617 (2023) 438; DOI:10.1038/d41586-023-01612-x

11. https://www.acdlabs.com/

12. Biological Magnetic Resonance Data Bank. JC Hoch, K Baskaran, H Burr, J Chin, HR Eghbalnia, T Fujiwara, MR Gryk, T Iwata, C Kojima, G Kurisu, D Maziuk, Y Miyanoiri, JR Wedell, C Wilburn, H Yao, M Yokochi; Nucleic Acids Res. 51 (2023) D368–D376; DOI: 10.1093/nar/gkac1050

13. https://bmrb.io/metabolomics/metabolomics_standards.php?dataset=metabolomics

14. NMR-STAR: comprehensive ontology for representing, archiving and exchanging data from nuclear magnetic resonance spectroscopic experiments. EL Ulrich, K Baskaran, H Dashti, YE Ioannidis, M Livny, PR Romero, D Maziuk, JR Wedell, H Yao, HR Eghbalnia, JC Hoch, JL Markley; J Biomol NMR 73 (2019) 5–9; DOI: 10.1007/s10858-018-0220-3

15. Stereo-Aware Extension of HOSE Codes. S Kuhn, SR Johnson; ACS Omega 4 (2019), 7323–7329; DOI: 10.1021/acsomega.9b00488

16. https://nmrshiftdb.nmr.uni-koeln.de/

17. https://discover.3ds.com/sites/default/files/2020-08/biovia_ctfileformats_2020.pdf

18. https://zenodo.org/record/7124055

19. Wikidata: a free collaborative knowledgebase. D Vrandečić, M Krötzsch; Commun. ACM 57 (2014) 78–85; DOI: 10.1145/2629489

20. InChI, the IUPAC International Chemical Identifier. SR Heller, A McNaught, I Pletnev, S Stein, D Tchekhovskoi; J. Cheminform. 7 (2015) 23; DOI: 10.1186/s13321-015-0068-4

21. https://www.json.org/json-en.html

22. https://www.xml.com/axml/testaxml.htm

23. RDKit: Open-source cheminformatics. https://www.rdkit.org

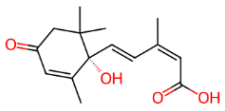
24. https://pubchem.ncbi.nlm.nih.gov/

25. https://scifinder.cas.org

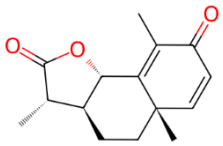
26. https://dtp.cancer.gov/dtpstandard/chemname/index.jsp

27. http://scholle.oc.uni-kiel.de/lind/iteach/kh_struct/kh_struct_eng_kap1.pdf

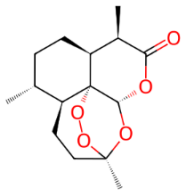
Scheme 1. Structure diagram, trivial name, and structure number of the compounds involved in the assessment of the dereplication procedure based on acd_lotus.



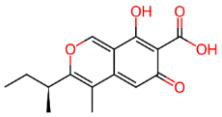
abscisic acid **1**



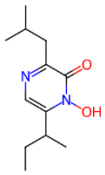
α-santonin **2**



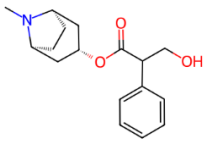
artemisinin **3**



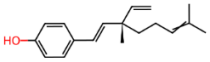
ascochitine **4**



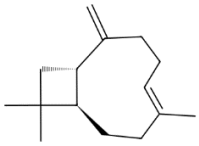
aspergillic acid **5**



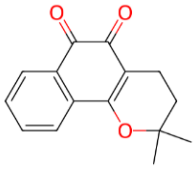
atropine **6**



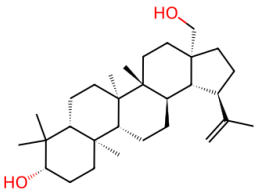
bakuchiol **7**



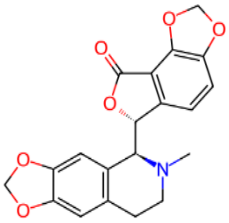
β-caryophyllene **8**



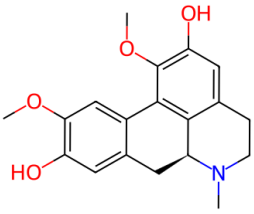
β-lapachone **9**



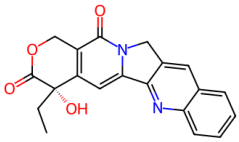
betulin **10**



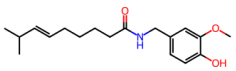
bicuculline **11**



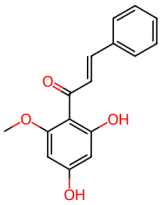
boldine **12**



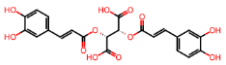
camptothecin **13**



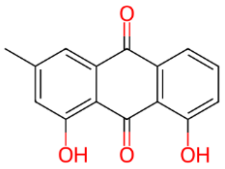
capsaicin **14**



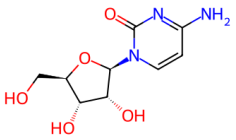
cardamonin **15**



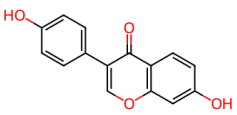
chicoric acid **16**



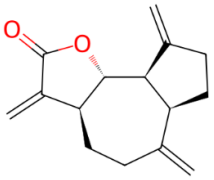
chrysophanol **17**



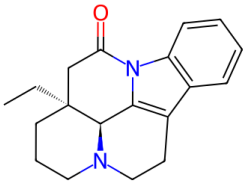
cytidine **18**



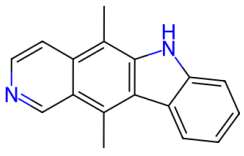
daidzein **19**



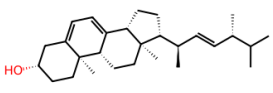
(-)-dehydrocostuslactone **20**



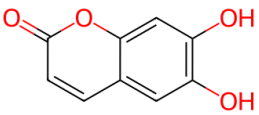
(-)-eburnamonine **21**



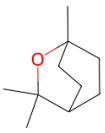
ellipticine **22**



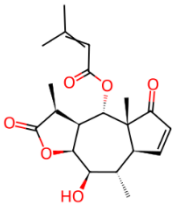
ergosterol **23**



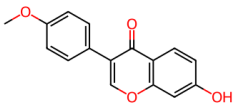
esculetin **24**



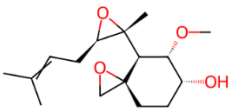
eucalyptol **25**



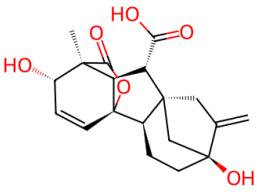
fastigillin B **26**



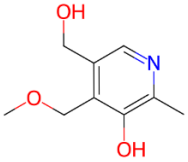
formononetin **27**



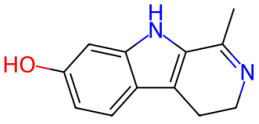
gelcohol **28**



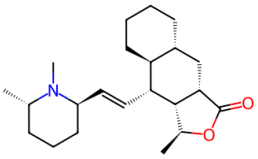
gibberellic acid **29**



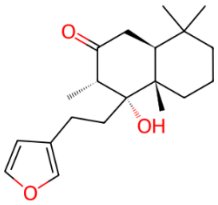
ginkgotoxin **30**



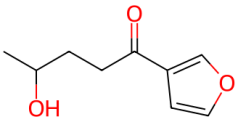
harmalol **31**



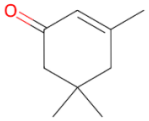
himbacine **32**



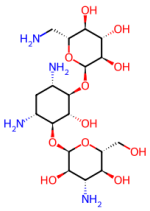
hispanolone **33**



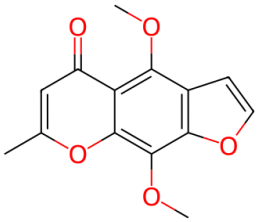
4-ipomeanol **34**



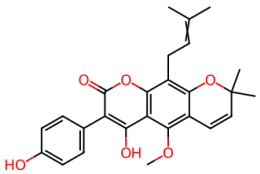
isophorone **35**



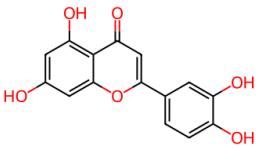
kanamycin **36**



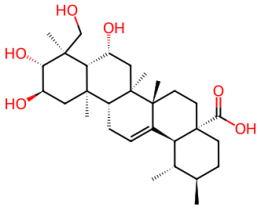
khellin **37**



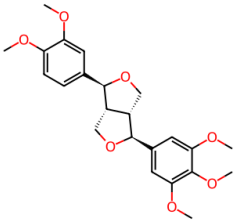
lonchocarpic acid **38**



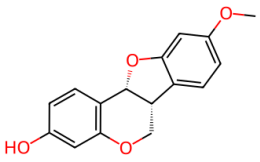
luteolin **39**



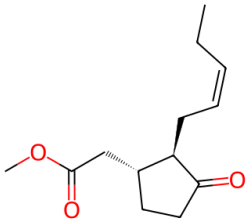
madecassic acid **40**



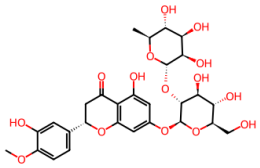
magnolin **41**



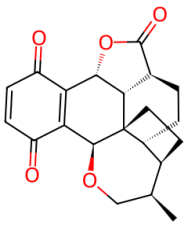
medicarpin **42**



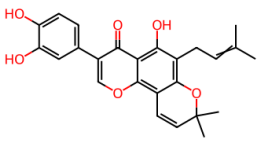
methyl jasmonate **43**



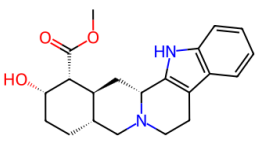
neohesperidine **44**



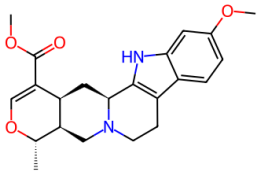
pleurotin **45**



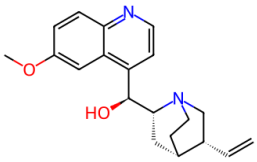
pomiferin **46**



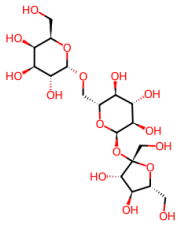
pseudoyohimbine **47**



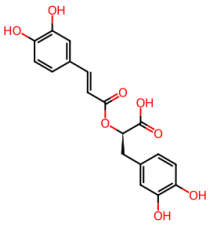
pubescine **48**



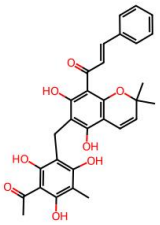
quinidine **49**



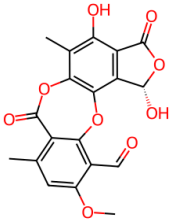
raffinose **50**



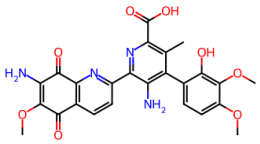
rosmarinic acid **51**



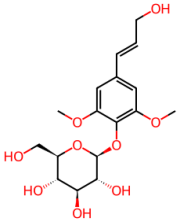
rottlerin **52**



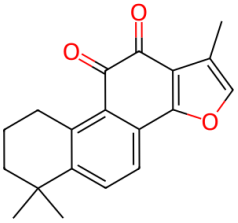
stictic acid **53**



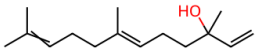
streptonigrin **54**



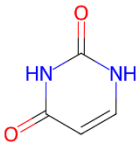
syringin **55**



tanshinone IIA **56**



trans-nerolidol **57**



uracil **58**

Table 1. Results of compound retrieval from experimental ^{13}C NMR chemical shifts

Compound name	Rank ACD/Labs	Rank nmrshiftdb2
abscisic acid **1**	1*	1
alpha-santonin **2**	1	1
artemisinin **3**	1	1
ascochitine **4**	2**	6
aspergillic acid **5**	1	1
atropine **6**	1	4
bakuchiol **7**	1	1
beta-caryophyllene **8**	2	3
beta-lapachone **9**	1*	1
betulin **10**	3	12
bicuculline **11**	1	1
boldine **12**	2	2
camptothecin **13**	1	2
capsaicin **14**	2	2
cardamonin **15**	6	4
chicoric acid **16**	2	2
chrysophanol **17**	1	1
cytidine **18**	—	—
daidzein **19**	1	1
(-)-dehydrocostuslactone **20**	1	1
(-)-eburnamonine **21**	1	1
ellipticine **22**	1	1
ergosterol **23**	2	10
esculetin **24**	1	1
eucalyptol **25**	1	1
fastigillin B **26**	1*	3
formononetin **27**	1	1
gelcohol **28**	1	1
gibberellic acid **29**	1*	13
ginkgotoxin **30**	1	1
harmalol **31**	1**	4
himbacine **32**	1*	6
hispanolone **33**	2	2
4-ipomeanol **34**	—	—
isophorone **35**	1	1
kanamycin **36**	5*	10
khellin **37**	1	1
lonchocarpic acid **38**	1	1
luteolin **39**	1	1
madecassic acid **40**	3	3
magnolin **41**	10	4
medicarpin **42**	3	5
methyl jasmonate **43**	1	2
neohesperidine **44**	4	15
pleurotin **45**	1	11
pomiferin **46**	1	1
pseudoyohimbine **47**	1	1

pubescine **48**	1*	4
quinidine **49**	2	1
raffinose **50**	21	1
rosmarinic acid **51**	1	1
rottlerin **52**	1	1
stictic acid **53**	1	1
streptonigrin **54**	1**	1
syringin **55**	3	1
tanshinone IIA **56**	1	1
trans-nerolidol **57**	2	1
uracil **58**	—**	2

* and ** respectively indicate the widening of the chemical shift tolerance to 6 and 8 ppm.