Versa DB: Assisting 13 C NMR and MS/MS Joint Data Annotation Through On-Demand Databases

Julien Cordonnier, Simon Remy, Jean-Hugues Renault, Jean-Marc Nuzillard

Chemistry—Methods

Research Article
doi.org/10.1002/cmtd.202300020

Chemistry Europe
European Chemical Societies Publishing

www.chemistrymethods.org

Versa DB: Assisting ^{13}C NMR and MS/MS Joint Data Annotation Through On-Demand Databases.

Julien Cordonnier[+, [a, b]] Simon Remy[+,*[a]] Jean-Hugues Renault,[[a]] and Jean-Marc Nuzillard[[a]]

Compound identification in complex mixtures by NMR and MS is best achieved through experimental databases (DB) mining. Experimental DB frequently show limitations regarding their completeness, availability or data quality, thus making predicted database of increasing common use. Querying large databases may lead to select unlikely structure candidates. Two approaches to dereplication are thus possible: filtering of a large DB before search or scoring of the results after a large scale search. The present work relies on the former approach. As far as we know, nmrshiftdb2 is the only open-source ^{13}NMR chemical shift predictor that can be freely operated in batch mode. CFM-ID 4.0 is one of the best-performing open-source tools for ESI-MS/MS spectra prediction. LOTUS is a freely usable and comprehensive collection of secondary metabolites. Integrating the open source database and software LOTUS, CFM-ID, and nmrshiftdb2 in a dereplication workflow requires presently programming skills, owing to the diversity of data encoding and processing procedures. A graphical user interface that integrates seamlessly chemical structure collection, spectral data prediction and database building still does not exist, as far as we know. The present work proposes a stand–alone software tool that assists the identification of mixture components in a simple way.

Introduction

Living organisms produce secondary metabolites that influence their relationships with their environment, in the widest acceptation of the term. For their study, they are withdrawn from their natural matrix by means of solvent extraction, possibly assisted by various physical and chemical processes. The resulting extracts contain most often a wide diversity of compounds, present over large ranges of concentrations, even though extraction protocols may be tuned to reach particular selectivity levels. Extract characterization is a general requirement, independently of the ultimate goal that motivated the initial extraction work. Focusing on the presence of a single or of a small number of particular, predefined compounds constitutes targeted analysis, which is not the topic of this article. Conversely, the untargeted analysis of complex mixtures of secondary metabolites is a task, also known as chemical profiling, for which this article proposes a practical assistance tool. A specific aspect of this task is that extracts from two taxonomically close organisms likely contain a non negligible proportion of common compounds, so that having carried out the thorough profiling of such an extract gives clues to the chemical content of the other one. Chemical profiling deals with mixtures of compounds, some of which were already reported as outcomes of previous studies, and other ones that must be characterized. The quick identification of known compounds, also called dereplication, constitutes an obvious optimization of the chemical profiling process.

Determining whether a compound is new or not requires an access to the knowledge accumulated by chemists over the past decades. Compound identification operates through the comparison of a set of signals from a presently available sample with those from the previously reported ones. Nuclear magnetic resonance (NMR) spectroscopy and mass spectrometry (MS) have become the main experimental techniques to determine these properties. NMR and MS data are hereafter referred to as spectral data. Both data types fulfill the requirements for the identification of the known compounds and for the structure determination of the unknown ones. Reading the literature about dereplication may lead to the idea that either one or the other technique may constitute a sufficient source of identification data. However, the association MS and NMR for dereplication is presently an active research field, motivated by a gain in efficiency produced by synergistic effect.

Comparing freshly acquired spectral data with old ones is possible only if the latter are available. There is no general–purpose, exhaustive, accurate, and publicly available database of experimental spectral data. The restriction of such a database to secondary metabolites only does not exist either. The accumulated knowledge of the relationships that exist between compound structures and spectral data opened the way to the creation of computer software dedicated to the prediction of spectral data from structures. Associating structures with

[a] J. Cordonnier,[+] Dr. S. Remy,[+] Prof. Dr. J.-H. Renault, Dr. J.-M. Nuzillard
University of Reims Champagne Ardenne, ICMR UMR7312, 51687 Reims, France
E-mail: simon.remy@univ-reims.fr

[b] J. Cordonnier[+]
University of Reims Champagne Ardenne, ESCAPE EA7510, 51097 Reims, France
E-mail: julien.cordonnier@univ-reims.fr

[+] These authors contributed equally.

Supporting information for this article is available on the WWW under https://doi.org/10.1002/cmtd.202300020

predicted spectral data yields a predicted database, considered as a replacement for the expected but nonexistent experimental database.

A structure and spectra database, either experimental or predicted, is only useful if a compound search algorithm is available. Searching for a compound using a few tens of ^{13}C NMR chemical shift values as a key in a database containing tens of millions compounds, such as the ones of nmrpredict[1] may results in an intractable set of solutions. As well, searching from MS/MS data for a compound among the 200,000+ ones in ISDB may produce unusable results. Taking into account *a priori* constraints on the nature of the search space reduces the number of spectra comparisons and increases the quality of the search results. These constraints may be related to biological taxonomy – if looking for particular taxon – or to chemical taxonomy – if looking for a particular class of secondary metabolites –. Two approaches of dereplication are thus possible. The first one uses taxonomy (biological or chemical or both) before database search by elimination of the presumably inappropriate compounds while the second uses taxonomy assumptions as a way of scoring results after a large scale search. The present work relies on the former approach. The corresponding dereplication assistance tool includes the selection of the structure set of interest from the largest available compound collection and the prediction of the associated NMR and MS spectral data.

The first successful attempt to provide a freely usable and comprehensive collection of secondary metabolites resulted in the COlleCtion of Open Natural ProdUcTs (COCONUT) database.[2] COCONUT is a compilation of numerous publicly available structure databases. Compound selection in COCONUT can be achieved online and through a software application programming interface (API). The latter option is particularly pertinent for the creation of automated dereplication workflows. Based on the same software framework as COCONUT, the naturaL prOducTs occUrrence databaSe (LOTUS) was build from publicly available literature data about taxons and related reported compounds.[3] COCONUT and LOTUS can select compounds according to substructure, chemical class, taxon, and values of physico-chemical descriptors. NMR prediction software rely directly or indirectly on sets of carefully assigned NMR spectra of reference compounds. Direct prediction operates by looking for a targeted atom on similar atomic environments, defined by hierarchically ordered spherical environment (HOSE) codes, in reference compounds. Indirect prediction requires a modeling of chemical shifts values according to chemical descriptors either through a linear model or the training of an artificial neural network. The linear model approach as also know as the increment method and is used by the ChemDraw software. Software may improve the prediction quality by combining the results of different approaches. The only open–source predictor that can be freely operated is nmrshiftdb2, to the best of our knowledge.[4] It relies on HOSE code search in a local database whose content can be examined. Moreover, it processes series of compounds in an unattended way, so that it can be integrated without much effort in other software projects. CFM-ID 4.0 is one of the best-performing open-source tools for ESI-MS/MS spectra prediction and is available as a web server and docker images.[5]

Integrating open source database and software such as LOTUS, CFM-ID, and nmrshiftdb2 in a dereplication workflow requires presently programming skills, owing to the diversity of data encoding and processing procedures. A graphical user interface that integrates seamlessly compound selection, spectral data prediction, and database creation does not exist, at the best of our knowledge.

The present articles proposes a stand–alone software that allows secondary metabolite specialists to identify mixture components in a simple way. Selecting compounds from taxonomic data and enriching them with spectral data follows the prescription known as "the three pillars of dereplication".[6]

Materials and Methods

All calculations were carried out on a workstation DELL Precision 3650 with 64 GB of RAM memory and an Intel(R) Core(TM) i9-10900 K CPU @ 3.70GHz running Windows 11 Education version 22H2. The ACD/C+H NMR Predictors and DB 2019.1.2 software was purchased from ACD/Labs (Toronto, Ontario, Canada). VersaDB is written in python (version 3.9.7) as language.[7] The pre-requisites, installation and running procedures are described on the dedicated github repository. The graphical user interface was build using the tkinter library (v0.1.0).[8]

Structural database building

In VersaDB, compounds structures are retrieved from the LOTUS natural product occurrence DB through its API *via* the getlotus() function from the ginfo.py script. This function uses the requests library (v2.26.0) to send HTTP requests.[9] The resulting JavaScript object notation (JSON) file contains the corresponding structures and their metadata.

The user can make queries according to taxonomical (key: T) and/or chemical (key: C) ontology and/or chemical formula (key: F) criteria and/or LOTUS ID (key: LTS), and according to the choice of specific natural product DB included in LOTUS. A single selected criterion is listed as follow: "key : databasename : criterionlevel : nameof criterion", where each element is necessary to browse the resulting JSON text and to obtain the corresponding structural DB. Depend ing on the query mode, if more than one criterion is selected, the criteria could be added as a list of individual queries or combined. In the first case the getlotusadd() function, search for each of the critera individually and merged the result into a single file after removal of duplicates. In the second case, the getlotusor() function finds compounds matching the taxonomical (key: T) criteria and then filter the results only retaining entries that meet the remaining criteria (keys C, F or LTS). The resulting list of structures associated with their LOTUS ID are stored into two similar files, cfmidinput.txt and structuraldb.txt to be used as input for CFM-ID or as structural DB in other platform (Sirius,[10] NAP[11]) respectively. The associated metadata from LOTUS (chemical and physical properties, taxonomy and chemontology informations) are saved in a file called cfmid input.tsv.

The plotly library (v5.4.0) was used to draw the interactive sunburst representing the chemical class distribution of the chemical structure set.[12] The classification is based on NPClassifier ontology provided by LOTUS, and the visualization is inspired by the CANOPUS-treemap chart.[13]

Research Article
doi.org/10.1002/cmtd.202300020
Chemistry Europe
European Chemical Societies Publishing

^{13}C chemical shift and MS/MS spectra prediction

The ^{13}C chemical shift prediction function of VersaDB was adapted from Kuhn and Nuzillard.[14] First, 2D coordinates of each compound listed as SMILES in the cfmidinput.txt input file are calculated using the RDKit library version 2022.03.5 run in the conda environment with python version 3.10.6. The l2sdf.py script writes the corre sponding structure descriptions into the cfmidinput2D.sdf file, in the Structure-Data format (SDF). The latter is used as an input by the predictSdf.bat script, a wrapper for the java-based nmrshiftdb2 packages.[1] The ^{13}C NMR chemical shifts of all carbon atoms in all structures are thus predicted, written in the cfmidinput2Dnmr.txt file. Carbon atoms in molecules and molecules in file are then reordered through molsort.py to produce cfmidinput2Dnmr sorted.txt. Then, nmrtags.py allows to merge cfmidinput2D.sdf and cfmidinput2Dnmrsorted .txt files *via* the RDKit library, to create LOTUSDBpredict.sdf output file. This output file is the first file readable by ACD/Labs or MarvinView software containing ^{13}C predicted chemical shifts. The fakeACD.py script defines a trans formation function that is used by sdfrw.py in order to add C NMR SHIFTS in "ACD/Labs C NMR Predictor" style tags to an input SDF file (LOTUSDBpredict.sdf), using ^{13}C NMR chemical shift values calculated by nmrshiftdb2. The resulting SDF file (fakeacdLOTUSDB predict.sdf) can be imported in the database module. Finally, the script called tagged.py allows to merge *via* RDKit library the fake acdLOTUSDBpredict.sdf with cfmidinput.tsv input files. This creates the ^{13}CNMRDatabase.sdf file containing the structures, their ^{13}C nmrshiftdb2-predicted chemical shifts in ACD/Labs C NMR Predictor style and all the compound metadata collected from LOTUS upon request for their biological, chemical taxonomy or physical–chemical properties. A script called createfolder.py gathers all the output files in a project folder named [datetime]NMRDB.

The MS/MS spectra prediction is performed through the CFM-ID 4.2.6.0 docker image.[5] MS/MS spectra are predicted at three collision energies using the CFM-ID pre–trained models. The present version of VersaDB was developed for Windows and thus requires the Windows Subsystem version 2 for Linux to be installed on the computer in addition to the Docker Desktop application. The predictMSMSspectra() function creates batches of 200 struc tures from cfmidinput.txt using the boltons python library (v21.0.0)[15] and generates a temporary command file sent to the docker image of CFM-ID for each structure batch. The function then runs each of the CFM-ID command file and writes the output spectra in a temporary folder called specmgf. Prediction calcula tions are parallelized using the GNU Parallel package.[16]

Finally, the predicted spectra are concatenated into a single file called predictspectra.mgf. In addition to this file, an errorlog file is created, containing all the molecules IDs for which no spectra were predicted. The annotatepredictMSMSspectra() function modifies the header of the spectra .mgf files by adding or substracting the mass of one proton to the precursor mass according to the prediction mode (+/- 1.007825) and adding metadata from Lotus DB contained in the cfmidinput.tsv file. It generate the final custom *in silico* mass spectral database file, MSMSspectraDatabase.mgf, and the annotationGNPSformat.tsv file containing the metadata needed to publish the DB on the GNPS platform. Finally, all the file are grouped into a single folder ([datatime]MSMSDB).

Spectral matching with third–party programs

The use of VersaDB was illustrated with the retrieval of the structure of dehydroabietic acid. Its experimental MS/MS spectrum and ^{13}C chemical shifts were downloaded from the GNPS database and directly copied from the original literature, respectively.[17,18] The experimental MS/MS data file was opened with the "import spectra list" module and the predicted MS/MS spectra DB was imported using the import user database module of the MetGem application.[19] Spectral matching in MetGem was performed using the parameters shown in the Supplementary Information file, in Fig. 1.

Results and Discussion

VersaDB software allows its user to predict locally ^{13}C NMR chemical shifts and MS/MS spectral libraries on demand for a subset of natural products selected according to biological or chemical taxonomic criteria. A view of the VersaDB graphical user interface is reproduced in Figure 1. The user first selects the biological taxonomic classification(s) to be displayed in the left upper zone of the interface. Then, the three left panels allows the user to select the family, the genus, or the species of an organism (frames 1 and 2). The series of three central panels offers the possibility to select the chemical class of the compounds of interest (frame 3). Natural products can also be retrieved by their chemical formula or LOTUS ID (frame 4). The compound selection method is validated by clicking on the "add to the input" button once a criterion is chosen. The retained criteria are displayed in frame 5. The user can either select the intersection (add) or the union (combined) of the list of structure corresponding to each criteria to create a structural database suitable for the object of study (*e.g.* a plant genus). Before predicting spectral data, an interactive graphical repre-sentation of the DB chemical class distribution can be drawn and opened in any web browser (frame 6). Finally, frame 7 is dedicated to the prediction of the spectral data. The resulting

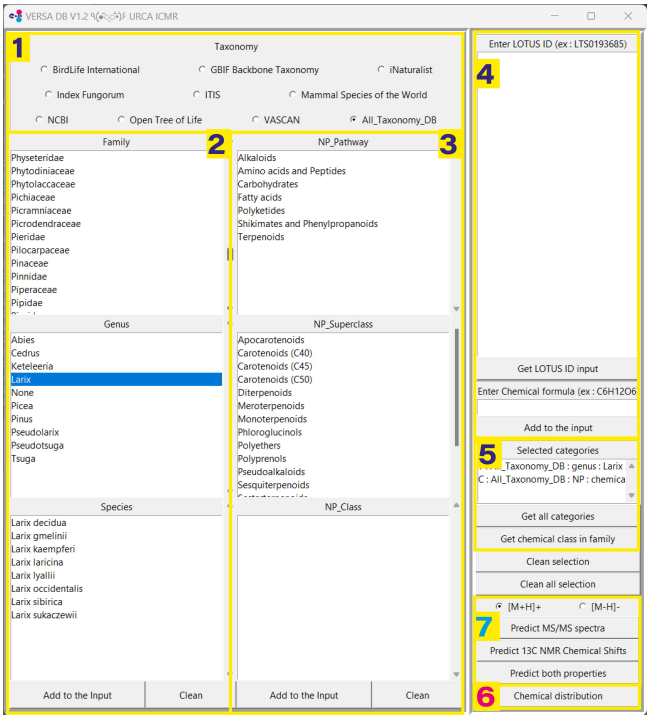


Figure 1. Graphical user interface of VersaDB.

DB with structures and spectra included is compatible with many software and dereplication platforms such as Sirius and the NAP workflow from GNPS.[10,11] The MS/MS spectral DB is suitable for spectral matching on GNPS or MetGem[17,19] and the ^{13}C chemical shift DB is designed to be compatible with ACD/Labs software and can be used in the framework of the CARAMEL dereplication workflow.[20]

In the example presented hereafter, VersaDB was used to generate a DB of compounds previously described in tree species belonging to the *Larix* genus, with emphasis on terpenoids. Larch, *Larix decidua*, is one of the twelve species of this genus. The present example was intended to guide the users to integrate VersaDB into their dereplication workflow and did not aim at benchmarking the prediction tools. The query for terpenoids from the *Larix* genus returned 376 compounds stored in a structural DB ("structuraldb.txt"). As shown by Figure 2, diterpenoids are highly represented in this structural DB. The MS/MS spectra (3 collision energies) were predicted in 7 minutes while the ^{13}C NMR chemical shifts were predicted in less than 1 minute. The spectral data of a compound could not be predicted because the structural information return by LOTUS contained 2 molecules (LOTUS ID : LTS0043877).

The experimental MS/MS spectra of dehydroabietic acid, a compound known to be produced by *Larix* trees, and the predicted MS/MS DB were imported in the MetGem software (See SI. 1 and example folder on the github repository). Dehydroabietic acid was annotated with a cosine score of 0.42. This moderate spectral similarity may be explained by a too extensive fragmentation pattern in the predicted data and/or too mild fragmentation conditions during the acquisition of the experimental data. On the other hand the predicted ^{13}C chemical shift DB was imported in the ACD/Labs software, and the dehydroabietic acid was annotated through the "search by chemical shift" module in first position according to Hit Quality Index value (See SI. 2). Those results illustrated the use of VersaDB for the quick annotation of a natural product. This annotation was performed through the comparison of experimental spectral data with taxonomically restricted database of predicted MS/MS and NMR spectral data.

The current version of the VersaDB software is an attempt to interface spectroscopic prediction programs in order to simplify database building and handling between dereplication tools without resorting on advanced programming skills. Future major development should include GUI redesign and compatibility improvement through a web interface. Resorting on user–supplied structural DB is currently possible by adding the corresponding files in the LotusDBinput folder although importation may be made easier through a dedicated module. Finally, as querying LOTUS DB through the web returns data that were not updated since their initial release, it might be more relevant to access compound structures directly from the wikidata API.[21]

Conclusions

This article reports the development of a python-based graphical user interface designed for the creation of local custom databases of natural products selected according to taxonomic criteria. The structural database and the corresponding predicted spectral databases can be further used to perform

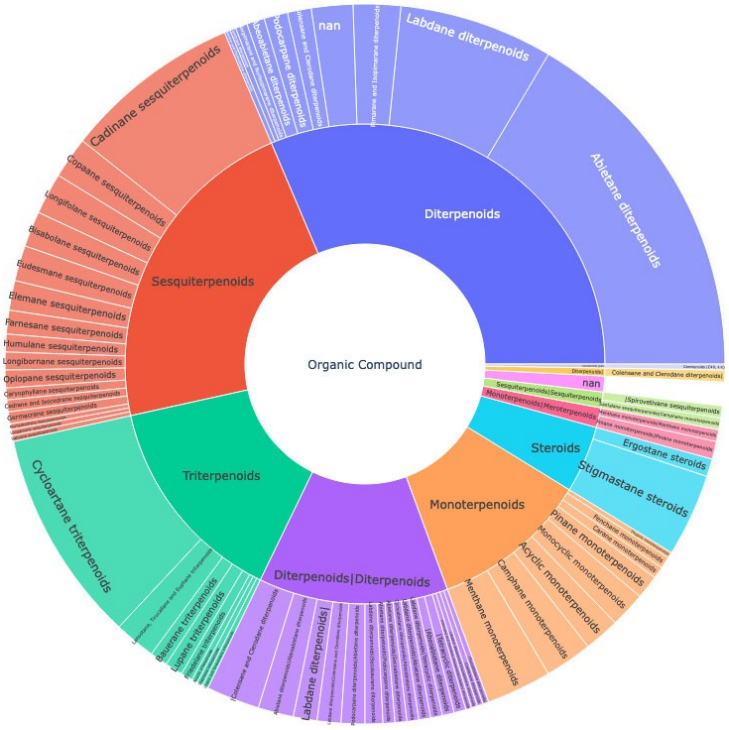


Figure 2. Chemical class distribution of compounds listed as terpenoids from *Larix* species retrieved from Lotus DB.

dereplication of complex mixtures through dedicated platforms and software such as Sirius, GNPS, MetGem and ACD/Labs. The use of VersaDB is illustrated by the annotation of a natural product for which experimental NMR and MS/MS spectra were searched and found in the purposely created custom databases.

Acknowledgements

JC thank the Région Grand-Est for its personal financial support.

Conflict of Interests

The authors have no conflicts of interest to declare

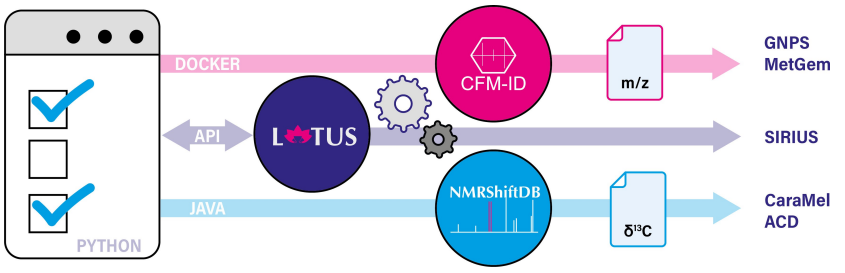
Data Availability Statement

The data that support this study are openly available at https://github.com/simremy/versadb_tk.

Keywords: Dereplication · Mass spectrometry · Natural products · Nuclear Magnetic Resonance · Software

[1] C. Steinbeck, S. Kuhn, *Phytochemistry* **2004**, *65*, 2711.
[2] M. Sorokina, P. Merseburger, K. Rajan, M. A. Yirik, C. Steinbeck, *Journal of Cheminformatics* **2021**, *13*, 2.
[3] A. Rutz, M. Sorokina, J. Galgonek, D. Mietchen, E. Willighagen, A. Gaudry, J. G. Graham, R. Stephan, R. Page, J. Vondrášek, C. Steinbeck, G. F. Pauli, J.-L. Wolfender, J. Bisson, P.-M. Allard, *eLife* **2022**, *11*.
[4] S. Kuhn, S. R. Johnson, *ACS Omega* **2019**, *4*, 7323.
[5] F. Wang, J. Liigand, S. Tian, D. Arndt, R. Greiner, D. S. Wishart, *Anal. Chem.* **2021**, *93*, 11692.
[6] M. Lianza, R. Leroy, C. M. Rodrigues, N. Borie, C. Sayagh, S. Remy, S. Kuhn, J.-H. Renault, J.-M. Nuzillard, *Molecules* **2021**, *26*, 637.
[7] G. Van Rossum The Python Library Reference, release 3.9.7, can be found under https://www.python.org/downloads/release/python-397/ **2020**.Lundh1999}
[8] Lundh. Fredrik, An introduction to tkinter, can be found under https://github.com/python/cpython/blob/3.11/Lib/tkinter/ init .py **1999**.
[9] R. V. Chandra, B. S. Varanasi, Python requests essentials, can be found under https://pypi.org/project/requests/ **2015**.
[10] K. Dührkop, M. Fleischauer, M. Ludwig, A. A. Aksenov, A. V. Melnik, M. Meusel, P. C. Dorrestein, J. Rousu, S. Böcker, *Nat. Methods* **2019**, *16*, 299.
[11] R. R. da Silva, M. Wang, L.-F. Nothias, J. J. J. van der Hooft, A. M. Caraballo-Rodríguez, E. Fox, M. J. Balunas, J. L. Klassen, N. P. Lopes, P. C. Dorrestein, *PLoS Comput. Biol.* **2018**, *14*, e1006089.
[12] P. T. Inc., Plotly Open Source Graphing Library for Python, can be found under https://plotly.com/python/ **2022**.
[13] K. Dührkop, L.-F. Nothias, M. Fleischauer, R. Reher, M. Ludwig, M. A. Hoffmann, D. Petras, W. H. Gerwick, J. Rousu, P. C. Dorrestein, S. Böcker, *Nat. Biotechnol.* **2021**, *39*, 462.
[14] S. Kuhn, J. Nuzillard, *Chemistry–Methods* **2022**.
[15] M. Hashemi, Boltons, can be found under https://github.com/mahmoud/boltons **2021**.
[16] O. Tange, GNU Parallel 20201222 ('Vaccine'), can be found under https://www.gnu.org/software/parallel/ **2020**.
[17] M. Wang, J. J. Carver, V. V. Phelan, L. M. Sanchez, N. Garg, Y. Peng, D. D. Nguyen, J. Watrous, C. A. Kapono, T. Luzzatto-Knaan, C. Porto, A. Bouslimani, A. V. Melnik, M. J. Meehan, W.-T. Liu, M. Crüsemann, P. D. Boudreau, E. Esquenazi, M. SandovalCalderón, R. D. Kersten, L. A. Pace, R. A. Quinn, K. R. Duncan, C.-C. Hsu, D. J. Floros, R. G. Gavilan, K. Kleigrewe, T. Northen, R. J. Dutton, D. Parrot, E. E. Carlson, B. Aigle, C. F. Michelsen, L. Jelsbak, C. Sohlenkamp, P. Pevzner, A. Edlund, J. McLean, J. Piel, B. T. Murphy, L. Gerwick, C.-C. Liaw, Y.-L. Yang, H.-U. Humpf, M. Maansson, R. A. Keyzers, A. C. Sims, A. R. Johnson, A. M. Sidebottom, B. E. Sedio, A. Klitgaard, C. B. Larson, C. A. B. P, D. TorresMendoza, D. J. Gonzalez, D. B. Silva, L. M. Marques, D. P. Demarque, E. Pociute, E. C. O'Neill, E. Briand, E. J. N. Helfrich, E. A. Granatosky, E. Glukhov, F. Ryffel, H. Houson, H. Mohimani, J. J. Kharbush, Y. Zeng, J. A. Vorholt, K. L. Kurita, P. Charusanti, K. L. McPhail, K. F. Nielsen, L. Vuong, M. Elfeki, M. F. Traxler, N. Engene, N. Koyama, O. B. Vining, R. Baric, R. R. Silva, S. J. Mascuch, S. Tomasi, S. Jenkins, V. Macherla, T. Hoffman, V. Agarwal, P. G. Williams, J. Dai, R. Neupane, J. Gurr, A. M. C. Rodríguez, A. Lamsa, C. Zhang, K. Dorrestein, B. M. Duggan, J. Almaliti, P.-M. Allard, P. Phapale, L.-F. Nothias, T. Alexandrov, M. Litaudon, J.-L. Wolfender, J. E. Kyle, T. O. Metz, T. Peryea, D.-T. Nguyen, D. VanLeer, P. Shinn, A. Jadhav, R. Müller, K. M. Waters, W. Shi, X. Liu, L. Zhang, R. Knight, P. R. Jensen, B. Palsson, K. Pogliano, R. G. Linington, M. Gutiérrez, N. P. Lopes, W. H. Gerwick, B. S. Moore, P. C. Dorrestein, N. Bandeira, *Nat. Biotechnol.* **2016**, *34*, 828.
[18] M. S. Costa, A. Rego, V. Ramos, T. B. Afonso, S. Freitas, M. Preto, V. Lopes, V. Vasconcelos, C. Magalhães, P. N. Leão, *Sci. Rep.* **2016**, *6*, 23436.
[19] F. Olivon, N. Elie, G. Grelier, F. Roussi, M. Litaudon, D. Touboul, *Anal. Chem.* **2018**, *90*, 13900.
[20] J. Hubert, J. M. Nuzillard, S. Purson, M. Hamzaoui, N. Borie, R. Reynaud, J. H. Renault, *Anal. Chem.* **2014**, *86*, 2955.
[21] D. Vrandečić, M. Krötzsch, *Commun. ACM* **2014**, *57*.

Manuscript received: April 5, 2023
Version of record online: ■■, ■■



J. Cordonnier, Dr. S. Remy, Prof. Dr. J.-H. Renault, Dr. J.-M. Nuzillard*

1 – 6

Versa DB: Assisting ^{13}C NMR and MS/MS Joint Data Annotation Through On-Demand Databases.

These authors contributed equally.- VersaDB is python-based program that first create a DB of natural products according to taxonomical criteria then predict MS/MS spectra and ^{13}C NMR chemical shift to produce spectral DB ready for integration in dereplication workflows.