



Published in final edited form as:

Magn Reson Chem. 2018 August ; 56(8): 703–715. doi:10.1002/mrc.4737.

NMReDATA, a standard to report the NMR assignment and parameters of organic compounds

Marion Pupier¹, Jean-Marc Nuzillard², Julien Wist³, Nils E. Schlörer⁴, Stefan Kuhn⁴, Mate Erdelyi⁵, Christoph Steinbeck⁶, Antony J. Williams⁷, Craig Butts⁸, Tim D.W. Claridge⁹, Bozhana Mikhova¹⁰, Wolfgang Robien¹¹, Hesam Dashti¹², Hamid R. Eghbalnia¹², Christophe Farès¹³, Christian Adam¹⁴, Pavel Kessler¹⁵, Fabrice Moriaud¹⁶, Mikhail Elyashberg¹⁷, Dimitris Argyropoulos¹⁸, Manuel Pérez¹⁹, Patrick Giraudeau^{20,21}, Roberto R. Gil²², Paul Trevorrow²³, and Damien Jeannerat¹

¹Department of Organic Chemistry, University of Geneva, 30 Quai E. Ansermet, 1211 Geneva 4, Switzerland ²Institut de Chimie Moléculaire de Reims, UMR CNRS 7312, BP 1039, 51687, Reims Cedex 2, France ³Chemistry Department, Universidad del Valle, 76001 Cali, Colombia ⁴Department of Chemistry, University of Cologne, Greinstr. 4, 50939 Köln, Germany ⁵Department of Chemistry - BMC, Uppsala University, Husargatan 3, 752 37 Uppsala, Sweden ⁶Institute for Inorganic and Analytical Chemistry, Friedrich-Schiller-University, Lessingstr. 8, 07743 Jena, Germany ⁷National Center for Computational Toxicology, Environmental Protection Agency, 109 T.W. Alexander Drive, Room D1311, Mail Drop D143-02, Research Triangle Park, NC 27711, USA ⁸School of Chemistry, Bristol University, BS8 1TS Bristol, UK ⁹Department of Chemistry, University of Oxford, Chemistry Research Laboratory, Mansfield Road, Oxford OX1 3TA, UK ¹⁰Institute of Organic Chemistry with Centre of Phytochemistry, Bulgarian Academy of Sciences, Akad. G. Bonchev Str. Bl.9, Sofia 1113, Bulgaria ¹¹University of Vienna, Department of Organic Chemistry, Währingerstr. 38, 1090 Vienna, Austria ¹²Department of Biochemistry, National Magnetic Resonance Facility at Madison (NMRFAM), 433 Babcock Drive, Madison, WI, USA ¹³Max-Planck-Institut für Kohlenforschung, Abteilung NMR, Kaiser-Wilhelm-Platz 1, 45470 Mülheim an der Ruhr, Germany ¹⁴Karlsruhe Institute of Technology, Hermann-von-Helmholtz-Platz 1, 76344 Eggenstein-Leopoldshafen, Germany ¹⁵Bruker BioSpin GmbH, Silberstreifen, 76287 Rheinstetten, Germany ¹⁶Bruker BioSpin AG, Industriestrasse 26, 8117 Fällanden, Switzerland ¹⁷Moscow Department, Advanced Chemistry Development, 6 Akademik Bakulev Street, Moscow 117513, Russian Federation ¹⁸Advanced Chemistry Development, Inc. (ACD/Labs), Venture House, Arlington Square, Downshire Way, Bracknell, Berkshire RG12 1WA, UK ¹⁹Mestrelab Research, S.L., Feliciano Barrera 9B - Bajo, ES-15706 Santiago de Compostela, Spain ²⁰EBSI Team, Chimie et Interdisciplinarité: Synthèse, Analyse, Modélisation (CEISAM) CNRS, UMR 6230, Université de Nantes, 92208, 2 rue de la Houssinière, BP 44322 Nantes,

Correspondence: Damien Jeannerat, Department of Organic Chemistry, University of Geneva, 30 Quai E. Ansermet, 1211 Geneva 4, Switzerland., damien.jeannerat@unige.ch.

Publisher's Disclaimer: DISCLAIMER

The views expressed in this paper are those of the authors and do not necessarily reflect the views or policies of the U.S. Environmental Protection Agency. Mention of trade names or commercial products does not constitute endorsement or recommendation for use.

SUPPORTING INFORMATION

Additional Supporting Information may be found online in the supporting information tab for this article.

France ²¹Institut Universitaire de France, 1 rue Descartes, 75005 Paris Cedex 05, France
²²Department of Chemistry, Carnegie Mellon University, 4400 Fifth Ave., Pittsburgh, PA 15213, USA
²³Wiley, The Atrium, Chichester PO19 8SQ, UK

Abstract

Even though NMR has found countless applications in the field of small molecule characterization, there is no standard file format available for the NMR data relevant to structure characterization of small molecules. A new format is therefore introduced to associate the NMR parameters extracted from 1D and 2D spectra of organic compounds to the proposed chemical structure. These NMR parameters, which we shall call NMReDATA (for *nuclear magnetic resonance extracted data*), include chemical shift values, signal integrals, intensities, multiplicities, scalar coupling constants, lists of 2D correlations, relaxation times, and diffusion rates. The file format is an extension of the existing Structure Data Format, which is compatible with the commonly used MOL format. The association of an NMReDATA file with the raw and spectral data from which it originates constitutes an NMR record. This format is easily readable by humans and computers and provides a simple and efficient way for disseminating results of structural chemistry investigations, allowing automatic verification of published results, and for assisting the constitution of highly needed open-source structural databases.

Keywords

Automatic structure verification; computer-assisted structure elucidation (CASE); data format; database; extracted data; NMR; NMR record; NMReDATA

1 | INTRODUCTION

Exploiting the NMR spectra of compounds consists of extracting the NMR parameters, which we call NMReDATA (for “NMR extracted data”), including chemical shift values, scalar coupling constants, and 2D correlations, and assigning them to a chemical structure. The reliability of the structure determination depends heavily on the accuracy and extent of the NMReDATA, which needs to be reported in a format compatible with a maximum number of software packages. Herein, we introduce a file format to effectively associate these NMReDATA with the assigned structure.^[1] These electronic reports will be generated by software assisting chemists in the interpretation of spectra or by using tools to create them in the case where the spectra were interpreted using “paper and pencil.” The commercial software providers Bruker, Mestrelab Research, Advanced Chemistry Development (ACD/Labs) and the non-commercial software platforms nmrshiftdb2^[2] and C6H6.org,^[3] have all committed to use the format in future releases of their products. These formatted NMReDATA will facilitate the accessibility and exchange of the partial or full structure assignment within the chemistry community. More importantly, it will constitute an easier way for the inspection and verification by reviewers because the data will be suitable for submission to scientific journals. Simply opening the NMReDATA file will permit a user to interrogate and confirm the assignments associated with the signals in the spectra. This

contrasts with the current unsatisfactory situation where NMR parameters are dispersed in peak lists, tables of correlations, and low-quality image-based spectra typically located in supplementary material, commonly in PDF format only and thereby making it difficult for both humans and computers to access and assess the data.^[4]

Because of the need to keep both the NMReDATA and the associated spectra together to allow for verification of their consistency, we have also introduced the concept of “NMR record”. An NMR record is a compressed folder including the NMReDATA file and all related 1D and 2D spectra (see Figure 1) including the raw data (i.e., the FIDs and all the acquisition and processing parameters). Such NMR records will also be optionally exported and imported by NMR software packages such as computer-assisted structure elucidation software^[5] and could be submitted with manuscripts for publication. For *Magnetic Resonance in Chemistry*, the requirement that all “spectral assignment” manuscripts submissions must be accompanied by full NMR records was announced recently and will be gradually enforced during 2018.^[6] Other journals have demonstrated interest in this initiative and may follow suit. The NMReDATA initiative intends to request the systematic deposition of NMR records with keywords and metadata about the publication, in open databases thereby making them permanently available to the community whenever the journal articles are accepted for publication. A DOI (Digital Object Identifier) must be provided to ensure stability of the data. Making spectral data available should also satisfy the “Open Data” requirements of a growing number of funding bodies.

Note that related efforts, for example, NMR-STAR^[7–10] and nmrML^[11] intend to cover the maximum number of data types and focus on the homogenization of the vendor-specific parameters or on applications such as metabolomics. In contrast, the NMReDATA initiative intends to restrict itself to a core set of NMR parameters relevant to support structure determination in the area of synthetic and natural products chemistry. This should facilitate the interoperability of the spectral data and their associated metadata. The NMReDATA are structured in two layers: the NMR parameters associated with individual spectra, and after their proper combination, the ones used for the assignment. This allows the validation of the data at three levels: (1) Check of the correspondence of the spectral NMR parameters with the actual spectra which must be either directly accessible in the NMR record, or through a direct link on freely accessible database. (2) Verification of the consistency of the NMR parameters extracted from the individual spectra. (3) Inspection of the compatibility of the NMR properties with the structure of the molecule.

2 | STRUCTURE DATA FORMAT (SDF)

The NMReDATA are associated with a chemical structure using the SDF file. This file format was introduced by Molecular Design Limited to associate one or more chemical structures with diverse types of data.^[12] The structures are encoded in the commonly used MOL (.mol) format, also introduced by Molecular Design Limited, dealing with 2D and 3D structures with and without implicit hydrogen atoms. The additional data are included in the file in the form of property value pairs called “tags” (see Figure 2). SDF files can be considered as MOL files with additional metadata. This file format is the primary import and export format for chemical structure drawing software packages. It is also used by the vast

majority of publicly available chemistry databases (e.g., PubChem,^[13] ChemSpider,^[14] and ChEBI^[15]) and across the field of cheminformatics to associate chemical structures with properties, for example, solubility, affinity to a drug target, or more trivial information such as storage locations. Thus, the nmredata.sdf files will include the assigned chemical structure (black in Figure 2), the NMR data (and possibly some non-NMR data), in a compact and convenient manner (see tags in Figure 2). Structure visualization software suites (such as ChemDraw^[16] and Jmol^[17]) will open the SDF files and only display the chemical structure encoded in MOL file format, while NMR software packages will be able to access the NMRReDATA tags and to allow for their visualization, manipulation, validation, etc.

3 | CHEMICAL STRUCTURE

A key requirement of an NMR assignment is to report the structure of the compound appropriately. The soundness of a molecular structure can be judged, if not proven, by the quality of the pairing of its structural features with its extracted NMR parameters. Indeed, depending on the data quality and requirements, an NMR characterization may be reported with different levels of structural precision. In the simplest case, one may only be able (or wish) to determine the connectivity of atoms (sometimes called a “flat structure”) with no additional information about the three-dimensional structure (see Figure 3a). In some cases, the exploitation of specific NMR parameters (e.g., a scalar coupling) may provide a local structural feature (e.g., the relative configurations of a pair of carbons as shown in Figure 3b). Finally, in the most exhaustive cases, the full analysis of the coupling constants, NOE data, anisotropic parameters, or the knowledge of the absolute configuration obtained by non-NMR means, allows the unambiguous configuration of all chirality elements of the molecule to be specified as in Figure 3c. In all cases, it is important to use a representation of the structure reflecting the available knowledge in order to avoid under- or, more problematically, over-interpretation. Note that given the inability of NMR to determine absolute configuration with classical methods, one should always consider the absolute configuration of the reported structures as not being ascertained.

The choice of the degree of structural certainty to include in an NMRReDATA file also depends on the completeness of the assignment. When the protons of the methylene groups of menthol are only assigned to their directly attached carbons, a structure with implicit hydrogen atoms is appropriate (Figure 3c). But when axial and/or equatorial protons are discriminated (typically through the observation of a large *trans-diaxial* coupling constant), a structure with explicit protons is necessary to unambiguously assign the two hydrogen atoms (see Figure 3d).

In general, chemical structures can be expressed with InChI strings.^[18] But replacing the MOL structure with the InChI string can only be done if the numbering of the atom is unique and unambiguous. This can be achieved using the ALATIS code^[19] (where ALATIS stands for Atom Label Assignment Tool using InChI String). In the context of NMRReDATA, we therefore recommend the structure part of NMRReDATA to use the ordering of the atoms of ALATIS-generated structures. This allows one to safely replace the structure with an

ALATIS code, making it possible to generate pure-text assignment reports where the assignment is performed using atom indices (see Section 7).

4 | NMReDATA TAGS IN SDF FILES

NMReDATA are encoded using a set of tags (see Figure 2 and Table 1). The tag names start with the prefix “NMReDATA_” in order to avoid confusion in those cases where the SDF files also include properties from other sources. The *NMReDATA Initiative*^[1] defines, for its Version 1, the set of “tags” listed in Table 1. It is recommended to follow the order of Table 1.

4.1 | General information about the sample and the NMR experiment

A minimal set of information about the sample and the performed NMR experiments is part of the NMReDATA format. It includes the solvent composition, the temperature and, if known, the concentration of the compound. The spectrometer magnetic field is not listed here, but can be inferred from the Larmor frequency of the detected nuclei for each NMR spectrum (see below).

Other data, such as the equipment used, pulse sequences, pulse durations, and referencing are not part of the format but can be searched in the acquisition and processing parameters that are present in the associated NMR record.

We strongly recommend to include the chemical formula, the SMILES and InChI code, but only if they truly reflect the associated chemical structure. These can facilitate searches in online databases. If they are not present in NMReDATA generated by the NMR software, they should be added before they are made available in databases.

4.2 | The assignment tag

The assignment tag contains, for each atom label used in the assignment of the spectra, the chemical shift (with adequate number of decimal places^[20]) and the list of atom indices they refer to (see Figure 2). Symmetry properties, such as the degeneracy resulting from the fast rotation of methyl groups, can be expressed by associating to a label the list of indices of the atoms it refers to (i.e., the ones of the three hydrogen atoms). Similarly, the nonequivalent protons of a CH₂ will be usually designated with a single label. If necessary, for example, in cases of magnetic non-equivalence (see Figure 4) and when different coupling constants need to be reported, the two hydrogen atoms should have different labels allowing the coupling constants to be assigned to a given coupling partner. In this case, the identical chemical shift will indicate the chemical equivalence.

The redundancy of the reporting of the chemical shifts and coupling constants with respect to the description of the 1D spectra (see below) is deliberate. The compilation of the data found in a set of spectra facilitates access to these essential NMR parameters, avoiding the need to search and analyze the values found in the individual 1D and 2D spectra. Another reason is that in some spectra, the chemical shifts may be reported in an ambiguous manner reflecting the limits of the specific experiment. For example, a set of protons may overlap in a 1D ¹H spectrum and the chemical shift reported as a broad chemical shift range, while an

HSQC or a J -resolved spectrum may provide clear-cut values for each individual chemical shift. The chemical shifts in the assignment tag can be seen as the aggregation of the chemical shift measured through the set of NMR spectra of the record.

4.3 | The J-coupling tag

In some situations, the analysis of the scalar coupling constants is essential for the determination of chemical structures. By analogy to the compilation of the chemical shift, the NMRReDATA format defines a tag to list all of the assigned coupling constants. Typically, the J_{HH} and J_{CH} are used to determine the 3D structure of natural and synthetic products (see the example of menthol in Figure 5). In other cases, such as when measuring long-range J_{CH} constants or when reporting $^1J_{CH}$ including residual-dipolar coupling, the list will contain only (or mostly) $^{2,3}J_{CH}$ or $^1D_{DH}$, respectively. Obviously, other couplings such as J_{CF} and J_{HP} can also be included.

4.4 | Tags describing spectra

For each spectrum included in NMR records, a spectral tag will contain the extracted information. The name of the tag (see Table 2) encodes the type of spectrum (1D vs. 2D) and the detected isotope. For 2D correlation spectra, the isotope evolving during t_1 and the type of mixing is specified before the detected isotope.

Before listing the peaks extracted from the spectra, a header provides important characteristics of the spectrum using the keywords (see Table 3). The Larmor frequency of the detected isotope follows “Larmor=”. Decoupling or absence of decoupling (if it is assumed by default—as for HSQC, HMBC spectra) is specified using “Decoupled = *isotope*” and “NonDecoupled = *isotope*”. Specifying the type of experiment (“CorType=”, COSY, HSQC, etc.) is particularly important to allow the verification of the compatibility of the reported NMR parameters with the chemical structure (see Validation of NMRReDATA). The format of the peak list differs for 1D and 2D spectra and is discussed in the next two sections.

A pointer to the original spectrum (with FID, acquisition and processing parameters in the crude original format of the manufacturer) is mandatory to permit verification (or refinement, correction, etc.) of the NMRReDATA. The “Spectrum_Location=” is a directory path, relative to the root of the zip file, pointing to the spectrum. When the NMR record is available in a web-accessible database, the HTML link having as prefix a DOI (to insure reliability of data access) and pointing directly to the readily downloadable zip file of the record is also specified.

4.5 | Peak list of 1D spectra

For each signal extracted from 1D spectra, one line lists its parameters. The first field is the chemical shift or chemical shift range. The others, called “attributes,” provide additional information such as multiplicity, integral, assignment, etc. (see the complete list in Table 4). An example of a multiplet described using various levels of detail is given in Figure 6.

Note that not all NMR parameters will be reported depending on the ambition of the user, the quality of the spectra and the capabilities of the software package used to analyze the spectra. The type and the number of peak attributes reported in the NMReDATA will therefore reflect the level of detail of the analysis and may be quite sparse.

Note that if the software generating the NMReDATA does not allow for the assignment of the coupling partners (Figure 6h), a program which is able to identify that couplings with (nearly) identical values are found for protons that are known to be two or three bonds away could be used to assign coupling constants (Figure 6i) and construct the coupling network (Figure 5). As mentioned earlier, more sophisticated tools using spectral simulations^[21–24] could further increase the reliability of the assignment and the precision of the coupling constants, and also provide reliable spectral parameters (chemical shifts and scalar coupling constants) in strongly coupled systems.

4.6 | Peak list of 2D spectra

For each 2D spectrum, the pairs of correlated signals are listed on separate lines. Other signal properties such as the intensity (either at the exact coordinates of the chemical shifts of signals or slightly off to avoid minimal values in the middle of doublet, or antiphase patterns) are optional (see Table 5).

Even if it is still uncommon to measure coupling constants in the multiplet structure of high-resolution 2D COSY spectra,^[25] NMReDATA has the capability of incorporating the results of such an analysis (see Figure 7b for an example).

Note that even if this is normally not used, the NMReDATA format allows the inclusion of unassigned cross peaks; in these cases, to define a signal, one can use the corresponding chemical shift values of the signals in place of the atom labels. This allows simple peak-picking software to generate NMReDATA files, which can be updated by a subsequent assignment process. Note as well that when the NMReDATA files are used to report assigned data, unassigned signals may be ignored.

5 | VALIDATION OF NMREDATA

By providing electronically readable NMR spectral parameter sets, NMReDATA files allow the manual or automatic verification of the reported data^[26–28] and the evaluation of the soundness of the determined structure. In the following paragraph, we shall only present a broad overview of how this is in principle possible, without discussing the known difficulties associated with automated analysis. We will consider a typical set of spectra used for a structural analysis of an organic compound (often called “full analysis”) that is usually 1D ¹H and ¹³C/DEPT-135 and 2D COSY, HSQC, and HMBC spectra. Note that some of the possibilities listed below are already being introduced into commercial software such as ACD/Spectrus,^[29] CMC-se,^[30] Mnova^[31] and other noncommercial platforms such as C6H6.^[3]

Having the associated chemical structure and the description of 1D ¹H and ¹³C spectra at hand allows for the determination of whether all protons and carbons have been

documented. One of the difficulties in this process is the possibility of degenerate signals resulting from symmetry, which can be overcome by the assignment of a tag listing the equivalent atoms or the presence of multiple signals with the exact same chemical shifts. In such a way, a program could also take into account the nature of the solvent to consider whether the protons from OH or NH are expected to exchange with the ones of the solvent. We encourage that exchanging protons be named “Ex.” Note that in the example of menthol, the hydroxyl proton is not exchanging (as demonstrates the coupling with H10ax) otherwise it would have been called “Ex.” This name will indicate to an automatic verification program that a group of protons assigned to a single signal is not due to symmetry, a property which can be assessed using cheminformatic tools such as CDK.^[32]

A verification of the reported multiplicity of signals can be performed regarding the spectrum and the chemical structure. Simply considering that multiplicity should include one “d” for each spin 1/2 located 1–3 bonds from the reference nucleus and keeping in mind that “t” counts for two couplings, “q” counts for three, etc. should be enough in most cases. More sophisticated methods could take into account nearby spins $>1/2$ (such as ^2D , ^{11}B , ^{17}O) or elements with multiple isotopes (resulting to the small doublet due to the minor ^{29}Si isotope or the two doublets due to the ^{117}Sn and ^{119}Sn isotopes). The values of the coupling constants could also be taken into account. A simulation of multiplets could be compared to the experimental spectrum using the reported values of the couplings and optionally using predicted values of the coupling constants.^[33] Note that second-order effects could be anticipated by observing, from the NMRReDATA, that coupling partners present a small difference in chemical shifts. Determining the coupling constants, even in the case of strong coupling and magnetic equivalence, should be possible by spectral simulation and fitting procedures found in SPINACH^[21,22] (all kinds of spectra), ANATOLIA^[23] (currently only 1D ^1H spectra) or GISSMO.^[24]

The 2D correlations reported in NMRReDATA can also be verified by checking the presence of signals at the coordinates of the chemical shifts of the correlated spins. The consistency of the reported data with the structure can also be tested by simply verifying that the number of bonds between the correlated spins is compatible with experiments based on scalar coupling (1-bond for HSQC, 2-, 3-, (4-)bonds for COSY and HMBC data). More sophisticated tools such as Logic for Structure Determination^[34] can confirm that the structure is compatible with the NMRReDATA, and when the structure is not uniquely determined, Logic for Structure Determination provides the user with the list of the alternative structures fully compatible with the spectral data.

Quality factors based on the comparison of the experimental chemical shifts and coupling constants with predicted ones^[2,33,35,36] could be easily calculated from the NMRReDATA.

6 | AMBIGUITIES

Ambiguities are commonplace in NMR assignment. In some cases, ambiguous assignment has no consequence on the determined structure. For example, the inability to assign two methoxy groups (OMe) does not mean that the structure does not have two OMe, but only that their assignments could not be determined. At a different level of the assignment, the

lack of resolution of an HSQC spectrum may not allow unambiguous assignment of a pair of carbons (resulting to the quite common “these carbons are interchangeable”) but still permit the determination of a structure. These two types of ambiguities were introduced into the NMReDATA format. Because dealing with ambiguities complicates the visualization of the data, we introduced a tag called “NMReDATA_LEVEL” to indicate when ambiguities are reported in the NMReDATA. This allows software tools that are not able to deal with ambiguities to reject such data and facilitate the quick development of programs dealing with the most common nonproblematic cases.

6.1 | Ambiguities in the assignment of spins

The first type of ambiguity occurs at the level of the assignment when a single chemical shift is associated with a list of atoms in the structure. It consists of listing a set of signals/labels for which a set of different assignments is possible (see Table 6). Another assignment ambiguity case occurs, for example, when a methoxy group (with ^1H and ^{13}C labeled “CH3a” and “Ca,” respectively) is not distinguished from a second methoxy group because of a lack of signal in the HMBC spectrum. One way to deal with NMReDATA with this type of ambiguities consists of duplicating (or multiplying when ambiguities include more than two possibilities) the table of references to the atom indices. These types of ambiguities can be dealt with by considering a set of possible atom numbering tables (instead of a unique one).

6.2 | Ambiguities in the assignment of a peak

Ambiguities can also be specified at the level of the assignment of a peak to a 1D or 2D spectrum, for example, if an HMBC signal is too broad in the indirect dimension to point to only one carbon (see the fourth entry in Table 7). In such a case, possible candidate signals are listed in parentheses. One consequence of the presence of such ambiguities is that the validation tool should use a logical “or” on the list of possible assignments.

7 | PURE TEXT DESCRIPTIONS

In many cases such as reports and PhD theses a text document is used to report spectral assignments and including the content of NMReDATA files may not be possible or desirable. We, therefore, support a textual form of the core information of .nmredata.sdf files for cases without ambiguities. It consists in a single line of text composed of the content of the second column of the Table 8. (See Supporting Information for an example).

After a header announcing “NMReDATA” and its version number, the chemical structure is encoded using the ALATIS format. This ensures that all information about the structure is retained and that the sequence of atoms is unambiguous. This allows the replacement of assignment tag labels with indices of the atom in the ALATIS format within double quotes. Then follows the compilation of chemical shifts and (when present) the assignment of the couplings. Afterwards, for each 1D spectrum, the chemical shifts are listed with the most relevant attributes in parentheses and the DOI and path pointing to the specific spectrum within the record. Finally, for each 2D spectrum, the list of correlated signals is given by direct reference to the atoms indices.

Note the absence of italics, Greek letters, or subscript or superscripts, facilitating the extraction of the text and making it possible to compress the text as QR codes, for example.

8 | ANALYSIS OF MIXTURES

Because the NMR record includes the spectra of a given sample, it may, by nature, contain multiple components (mixtures, slowly exchanging isomers, etc.). In cases where more than one compound is associated with an NMR record, one should simply generate one NMReDATA file per assigned compound (see Figure 8).

9 | CONCLUSION

Despite the wide application of NMR for the study of small molecules, there is no standardized procedure for the exchange of structure related NMR parameters between software suites. As a consequence, retrieving different types of experimental information from reports and publications has been a challenge in the NMR community. Herein, we addressed this challenge by developing a data structure for NMReDATA Records (Version 1.0) in a concise human- and machine-readable format. It results from the collective work by a wide range of stakeholders including experts in structure elucidation, software developers, and journal editors. The format will foster the generation and the exchange of NMR data, and facilitate their validation by chemists, NMR experts, authors and reviewers of publications, and, finally, all users of the published data. The urgent need of an electronic data format in combination with an open-access repository of NMR spectral data is further supported by the large number of wrong structures found in the literature^[27,28] and the difficulty in revising them.

After integration of the new format into NMR software, the generation of NMR records including the NMReDATA files will take no additional time. A “save as ...” option or a simple copy/paste is all that will be necessary to save or transfer NMR records to dedicated NMR databases such as cheminfo.org. When deposited for review, they will initially be embargoed and only accessible to the reviewers using a coded URL included in the manuscript. When the article is published, the NMR records will become open to the public and relevant records could be merged into databases such as C6H6.org,^[3] nmrshiftdb2 (which will become part of the previous),^[2,37] BMRB,^[10] ChemSpider,^[14,38] Metabolights,^[39] and SDBS.^[40]

The NMReDATA initiative collaborates with the Biological Magnetic Resonance Bank^[10] to develop interfaces between the NMReDATA and the NMR-STAR^[7–10] data formats. These interfaces are designed to guarantee seamless access to reference metabolites from BMRB in NMReDATA data format, and also to readily deposit published NMReDATA data into BMRB.

Version 1 of the format should be seen as a compromise, which will evolve in order to integrate new features as motivated developers join the initiative. A future release could address the question of the format of the raw spectra, and additional requirements for the Open Data initiative. Currently, NMR records can include any format generated by

spectrometer manufacturers—only the NMRReDATA per se are standard—but accepting a single standard for the spectra would further simplify the utilization of NMR records.

We believe that, in the future, all users of NMR data will have access to software packages allowing them to easily check chemical structures for consistency with all experimental data used for structure verification or elucidation.^[5] This shall significantly reduce the number of wrong structures published in chemical journals.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

ACKNOWLEDGEMENTS

D. J. and M. P. thank the State of Geneva and the Swiss NSF (Grant 200021_147069 and 206021_128746) for funding. H. D. and H. R. E are supported by the U.S. National Institutes of Health (NIH) grant P41GM103399 and P41GM111135 (NIGMS). N. E. S gratefully acknowledges funding by the Deutsche Forschungsgemeinschaft (DFG, IDNMR project, Grant SCHL 580/3–1). J. M. is thanked for useful discussions and for comments on the manuscript. E. S. and A. M. are acknowledged for their valuable comments in the final stage of review.

Funding information

Swiss NSF, Grant/Award Number: 200021_147069 and 206021_128746 206021_128746; Deutsche Forschungsgemeinschaft, Grant/Award Number: 580/3–1; U.S. National Institutes of Health (NIH), Grant/Award Numbers: P41GM111135 and P41GM103399; State of Geneva, Grant/Award Number: 200021_147069

REFERENCES

- [1]. <http://www.nmredata.org>.
- [2]. Kuhn S, Schlörer NE, Magn. Reson. Chem 2015, 53, 582 10.1002/mrc.4263 [PubMed: 25998807]
- [3]. Patiny L, Zasso M, Kostro D, Bernal A, Castillo AM, Bolaños A, Asencio MA, Pellet N, Todd M, Schlörer N, Kuhn S, Holmes E, Javor S, Wist J, Magn. Reson. Chem 2017 10.1002/mrc.4669
- [4]. Jeannerat D, Magn. Reson. Chem 2017, 55, 7 10.1002/mrc.4527 [PubMed: 27642110]
- [5]. Elyashberg ME, Williams AJ, Blinov K, Contemporary Computer-Assisted Approaches to Molecular Structure Elucidation, RSC Publishing, Cambridge 2012.
- [6]. Gil R, Martin G, Magn. Reson. Chem 2017, 55, 1057 10.1002/mrc.4631 [PubMed: 28699648]
- [7]. Spadaccini N, Hall SR, J. Chem. Inf. Model 2012, 52, 1901 10.1021/ci300074v [PubMed: 22725659]
- [8]. Hall SR, Spadaccini N, J. Chem. Inf. Model 1994, 34, 505 10.1021/ci00019a005
- [9]. Hall SR, J. Chem. Inf. Model 1991, 31, 326 10.1021/ci00002a020
- [10]. Ulrich EL, Akutsu H, Doreleijers JF, Harano Y, Ioannidis YE, Lin J, Livny M, Mading S, Maziuk D, Miller Z, Nakatani E, Schulte CF, Tolmie DE, Kent Wenger R, Yao H, Markley JL, Nucleic Acids Res 2008, 36, D402 10.1093/nar/gkm957 [PubMed: 17984079]
- [11]. Schober D, Jacob D, Wilson M, Cruz JA, Marcu A, Grant JR, Moing A, Deborde C, de Figueiredo LF, Haug K, Rocca-Serra P, Easton J, Ebbels TMD, Hao J, Ludwig C, Gunther UL, Rosato A, Klein MS, Lewis IA, Luchinat C, Jones AR, Grauslys A, Larralde M, Yokochi M, Kobayashi N, Porzel A, Griffin JL, Viant MR, Wishart DS, Steinbeck C, Salek RM, Neumann S, Anal. Chem 2018, 90, 649 10.1021/acs.analchem.7b02795 [PubMed: 29035042]
- [12]. Dalby A, Nourse JG, Hounshell WD, Gushurst AKI, Grier DL, Leland BA, Laufer J, J. Chem. Inf. Model 1992, 32, 244 10.1021/ci00007a012
- [13]. Kim S, Thiessen PA, Bolton EE, Chen J, Fu G, Gindulyte A, Han L, He J, He S, Shoemaker BA, Wang J, Yu B, Zhang J, Bryant SH, Nucleic Acids Res 2016, 44, D1202 10.1093/nar/gkv951 [PubMed: 26400175]
- [14]. <http://www.chemspider.com>.

- [15]. Hastings J, de Matos P, Dekker A, Ennis M, Harsha B, Kale N, Muthukrishnan V, Owen G, Turner S, Williams M, Steinbeck C, *Nucleic Acids Res* 2013, 41, D456 10.1093/nar/gks1146 [PubMed: 23180789]
- [16]. <http://www.cambridgesoft.com/software/overview.aspx>.
- [17]. http://wiki.jmol.org/index.php/Main_Page.
- [18]. Heller SR, McNaught A, Pletnev I, Stein S, Tchekhovskoi D, *Aust. J. Chem* 2015, 7, 23 10.1186/s13321-015-0068-4
- [19]. Dashti H, Westler WM, Markley JL, Eghbalian HR, *Sci. Data* 2017, 4, 170073 10.1038/sdata.2017.73 [PubMed: 28534867]
- [20]. Pauli GF, Chen SN, Lankin DC, Bisson J, Case RJ, Chadwick LR, Godecke T, Inui T, Kronic A, Jaki BU, McAlpine JB, Mo S, Napolitano JG, Orjala J, Lehtivarjo J, Korhonen SP, Niemitz M, *J. Nat. Prod* 2014, 77, 1473 10.1021/np5002384 [PubMed: 24895010]
- [21]. Hogben HJ, Krzystyniak M, Charnock GT, Hore PJ, Kuprov I, *J. Magn. Reson., Ser. A* 2011, 208, 179 10.1016/j.jmr.2010.11.008
- [22]. Kuprov I, *Magn. Reson. Chem* 2017 10.1002/mrc.4660
- [23]. Cheshkov DA, Sheberstov KF, Sinitsyn DO, Chertkov VA, *Magn. Reson. Chem* 2017 10.1002/mrc.4689
- [24]. Dashti H, Westler WM, Tonelli M, Wedell JR, Markley JL, Eghbalian HR, *Anal. Chem* 2017, 89, 12201 10.1021/acs.analchem.7b02884 [PubMed: 29058410]
- [25]. Jeannerat D, *Magn. Reson. Chem* 2000, 38, 156 10.1002/(SICI)1097-458X(200003)38:3%3C156::AID-MRC610%3E3.0.CO;2-R
- [26]. Elyashberg M, Williams AJ, Blinov K, *Nat. Prod. Rep* 2010, 27, 1296 10.1039/c002332a [PubMed: 20480119]
- [27]. Robien W, in *Progress in the Chemistry of Organic Natural Products*, Vol. 105, 2017, (Eds: Kinghorn AD, Falk H, Gibbons S, Kobayashi JI), Springer 2017 137 10.1007/978-3-319-49712-9_3
- [28]. Nicolaou KC, Snyder SA, *Angew. Chem. Int. Ed* 2005, 44, 1012 10.1002/anie.200460864
- [29]. Advanced Chemistry Development, Inc, Toronto.
- [30]. Biospin Bruker, Germany.
- [31]. Mestrelab.
- [32]. Willighagen EL, Mayfield JW, Alvarsson J, Berg A, Carlsson L, Jeliazkova N, Kuhn S, Pluskal T, Rojas-Cherto M, Spjuth O, Torrance G, Evelo CT, Guha R, Steinbeck C, *Aust. J. Chem* 2017, 9, 33 10.1186/s13321-017-0220-4
- [33]. Navarro-Vázquez A, Santamaría-Fernández R, Sardina FJ, *Magn. Reson. Chem* 2017 10.1002/mrc.4667
- [34]. Nuzillard JM, Plainchont B, *Magn. Reson. Chem* 2017 10.1002/mrc.4612
- [35]. Smith SG, Goodman JM, *J. Am. Chem. Soc* 2010, 132, 12946 10.1021/ja105035r [PubMed: 20795713]
- [36]. Grimblat N, Zanardi MM, Sarotti AM, *J. Org. Chem* 2015, 80, 12526 10.1021/acs.joc.5b02396 [PubMed: 26580165]
- [37]. <http://www.nmrshiftdb.org>.
- [38]. Pence HE, Williams A, *Chem J. Education* 2010, 87, 1123 10.1021/ed100697w
- [39]. <http://www.ebi.ac.uk/metabolights>.
- [40]. http://sdbs.db.aist.go.jp/sdbs/cgi-bin/cre_index.cgi.

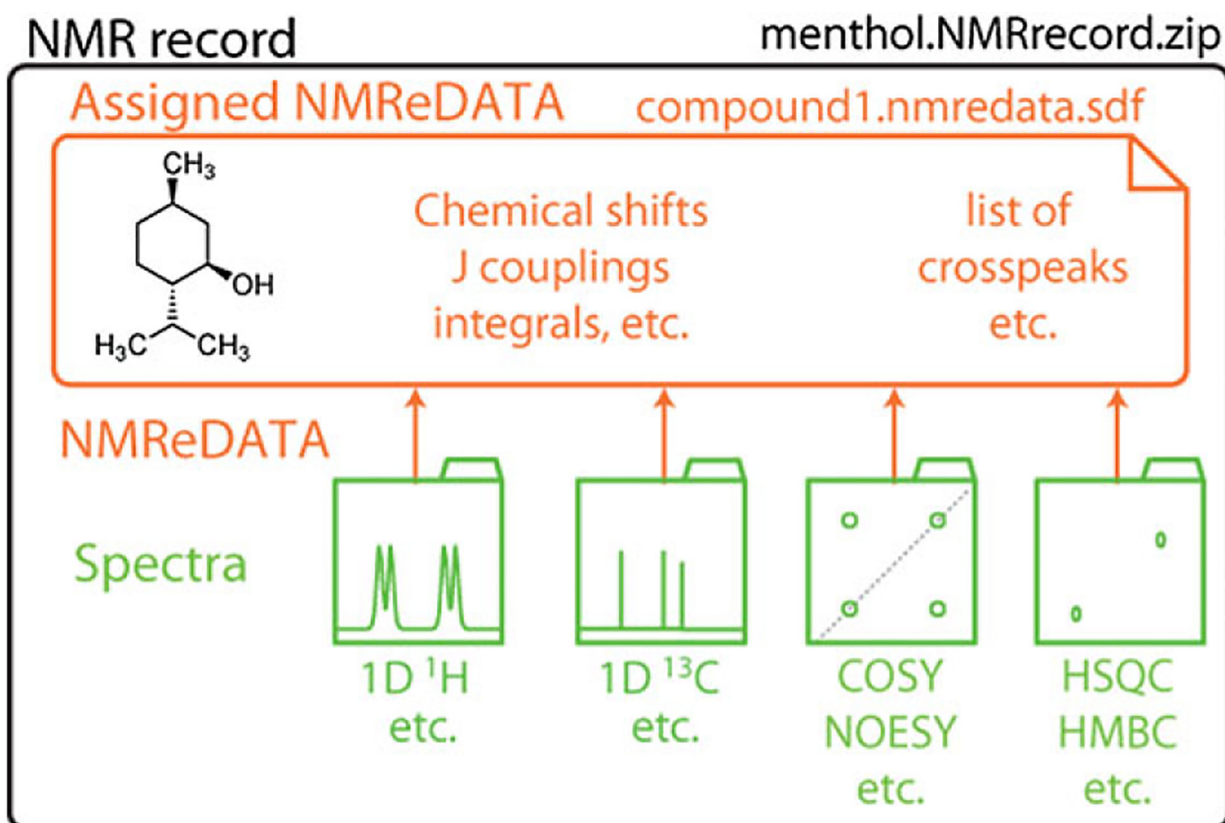
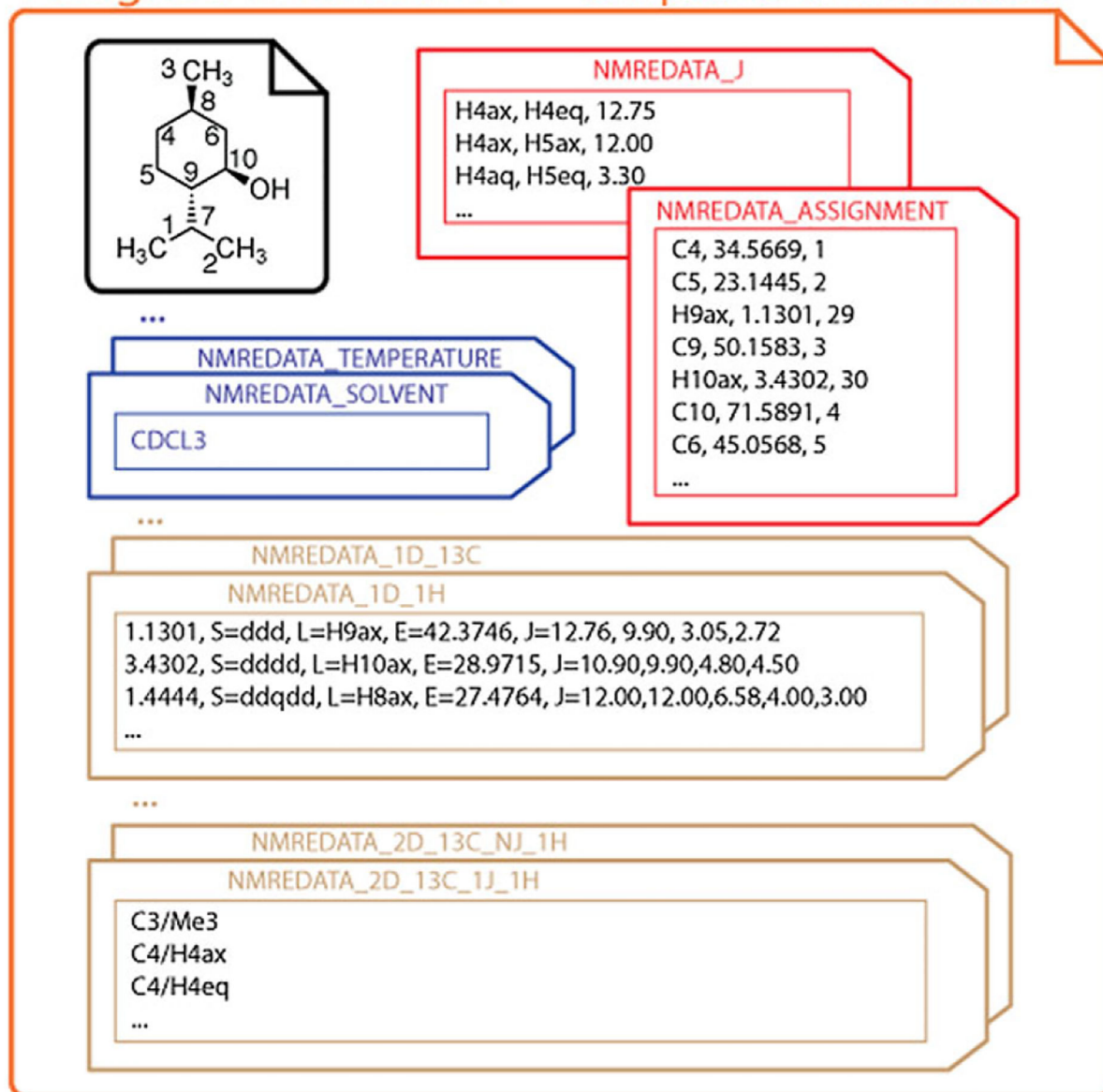


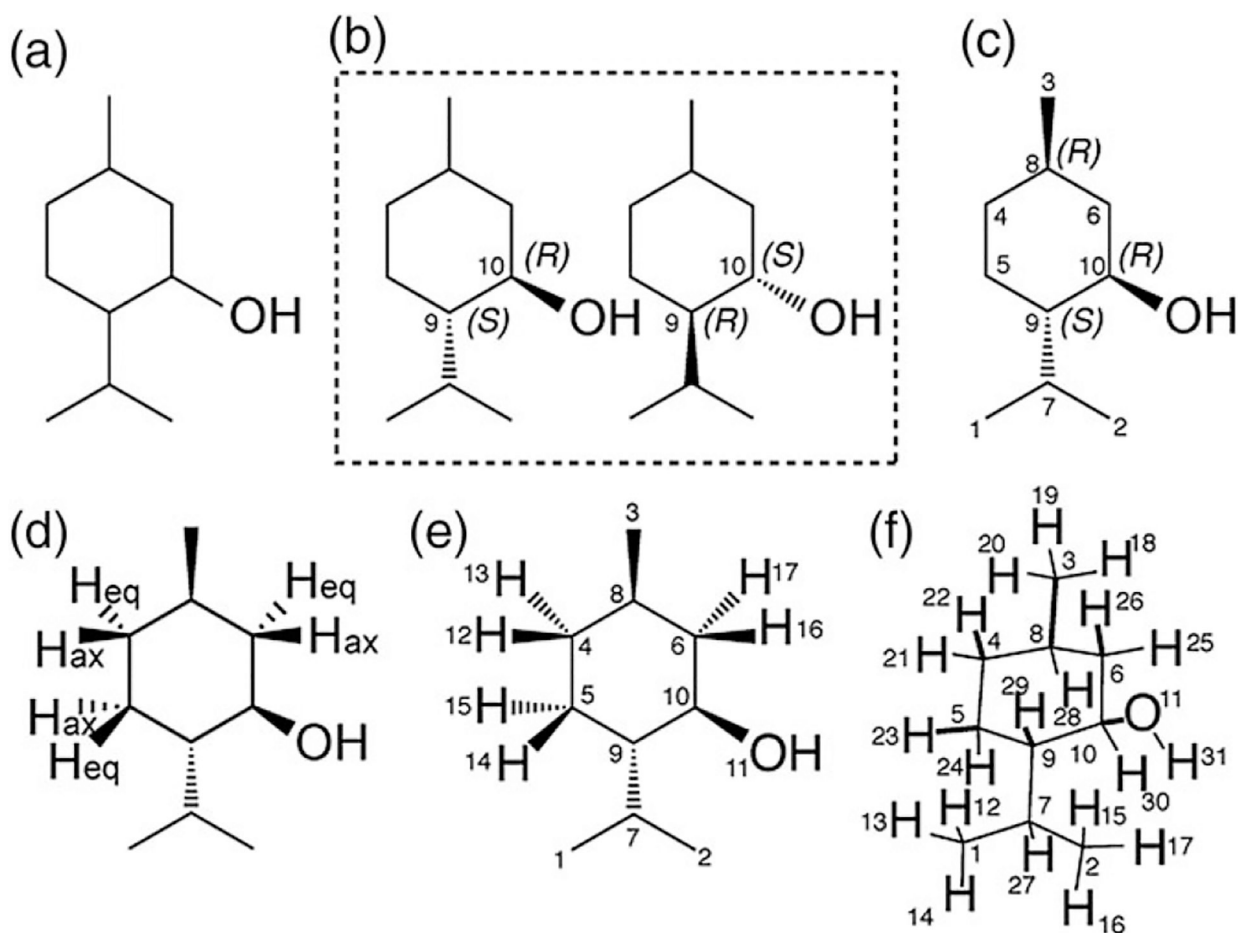
FIGURE 1.

An “NMR record” consists of a compressed folder in .zip format and includes a set of NMR spectra with the time domain FIDs, the acquisition and processing parameters (in green), and the NMReDATA file (.nmredata.sdf file in orange) containing a chemical structure and the extracted NMR data

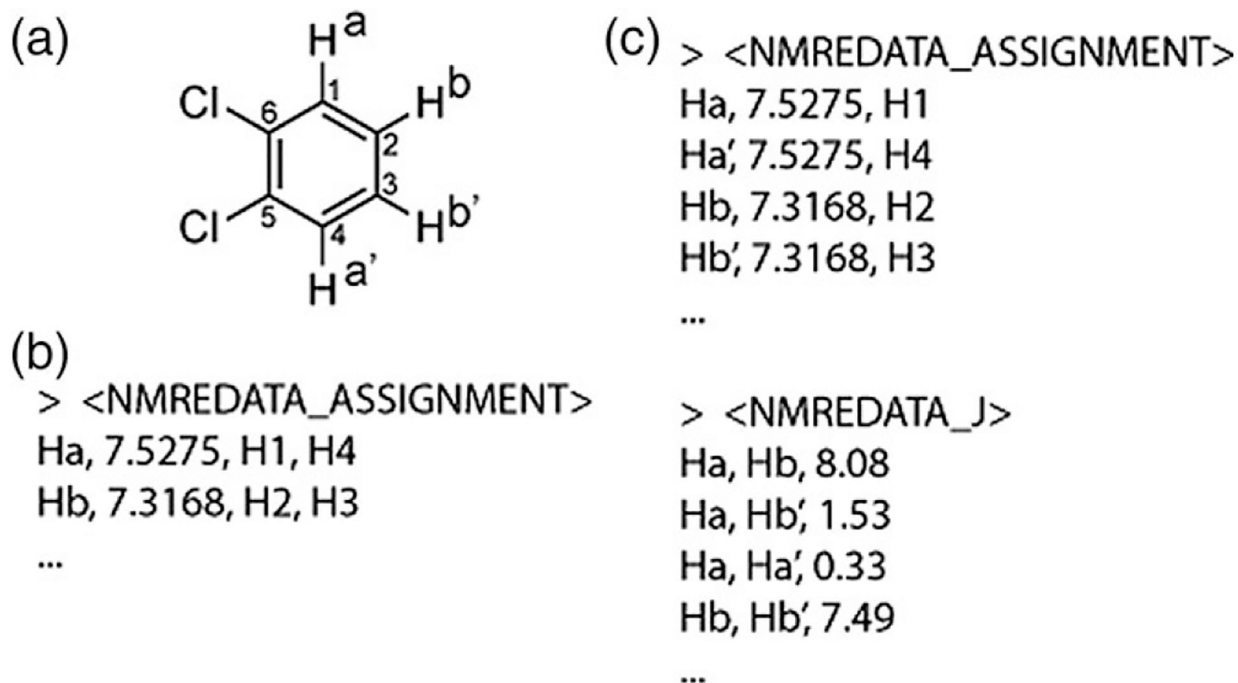
Assigned NMR eDATA compound1.nmredata.sdf

**FIGURE 2.**

Pictorial representation of the content of an SDF file included in NMR records. It consists of a structure (in black) and the NMR extracted data (NMReDATA tags). The property value pairs comprise information on the sample (blue tags), compilations of the chemical shift, and scalar coupling information (red tags) and descriptions of 1D and 2D spectra (brown tags)

**FIGURE 3.**

Different levels of interpretation of the NMR spectra of menthol. (a) The absolute configuration of the carbons is unknown. (b) The relative configuration of carbons 9 and 10 is defined based on the observation of a characteristic large *trans*-diaxial coupling ($J > 10$ Hz). (c) The chirality is fully determined. (d) Structure with explicit hydrogen atoms depicted for unambiguous assignment of the pairs of nonequivalent protons of the methylene groups. (e) Representation and ALATIS numbering of nonequivalent atoms. (f) True 3D model structure with all hydrogen atoms, including equivalent ones, with ALATIS numbering. InChI = 1S/C10H20O/c1-7(2)9-5-4-8(3)6-1ü(9)II/h7-11H,4-6H2,1-3H3/t8-,9+,10-/m1/s1 as reported by ALATIS. Note that even if the chirality of menthol is fully determined, the assignments of carbons 1 and 2 are usually not unambiguously assigned because it has no consequence on the determination of the structure

**FIGURE 4.**

Examples of NMREDATA_ASSIGNMENT tags for the description of the ^1H spectrum of orthodichlorobenzene (a). For simplicity only hydrogen atoms were listed in the figure. In general, the list includes reference to carbon, fluorine, and phosphorus atoms if their spectra are part of the NMR record. (b) When ignoring the complex structures of the AA'BB' system, the chemically equivalent protons can be described using a single label ("a" and "b"). (c) When the coupling constants were extracted using optimal fitting with spectral simulations, distinct labels for the equivalent protons (a, a', b, b') are necessary to report all coupling constants using an NMREDATA_J tag (see next paragraph). Note that because hydrogen atoms were implicit in structure (a), "H" was added and the indices of the atoms are the ones of the directly attached heavy atoms

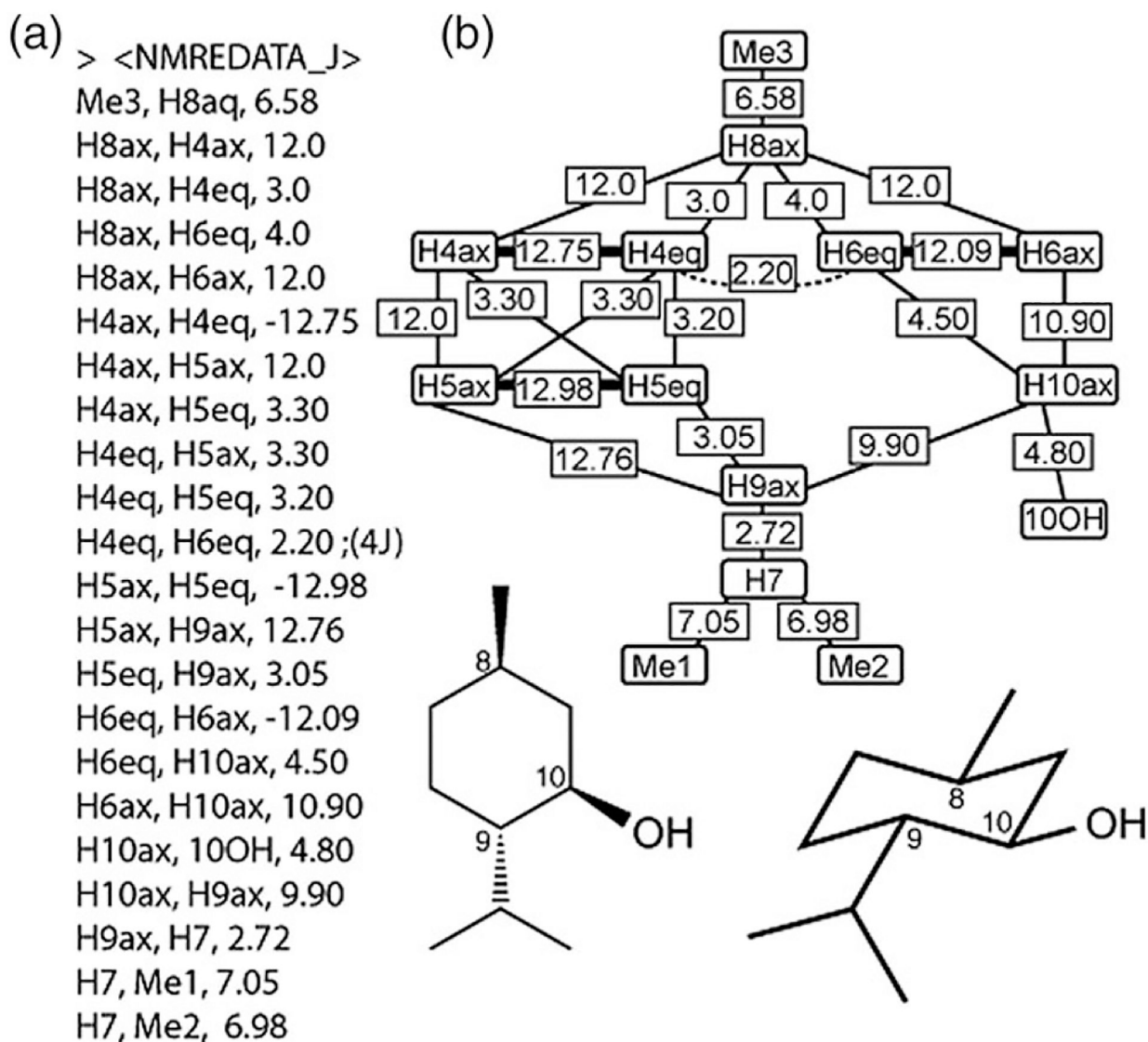
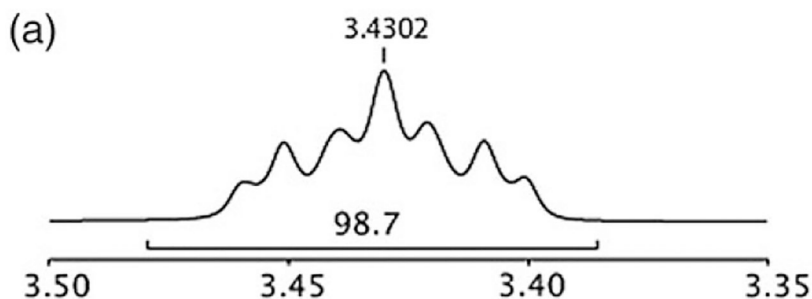


FIGURE 5.

(a) Example of a J -coupling tag. (b) Corresponding ^1H - ^1H coupling network. The all-equatorial substitution of the cyclohexane ring of menthol can be deduced from the six characteristically large coupling constants (>8 Hz) of pairs of axial protons. Thick, normal, and dotted lines represent 2J , 3J , and 4J , respectively, originating from the spectral tag (see following sections). If known, the signs of the coupling constants should be specified, otherwise the absolute values are listed. Optionally, the number of bonds between the pair of spins can be specified after “nb=”. (e.g., “Me3, H8aq, 6.58, nb=3”)



(b) > <NMREDATA_1D_1H>
Larmor=500.13

.....

(c) 3.475-3.385, E=98.7

(d) 3.475-3.385, E=98.7, N=1

(e) 3.4302, S=m, E=98.7, N=1

(f) 3.4302, I=1.705E5

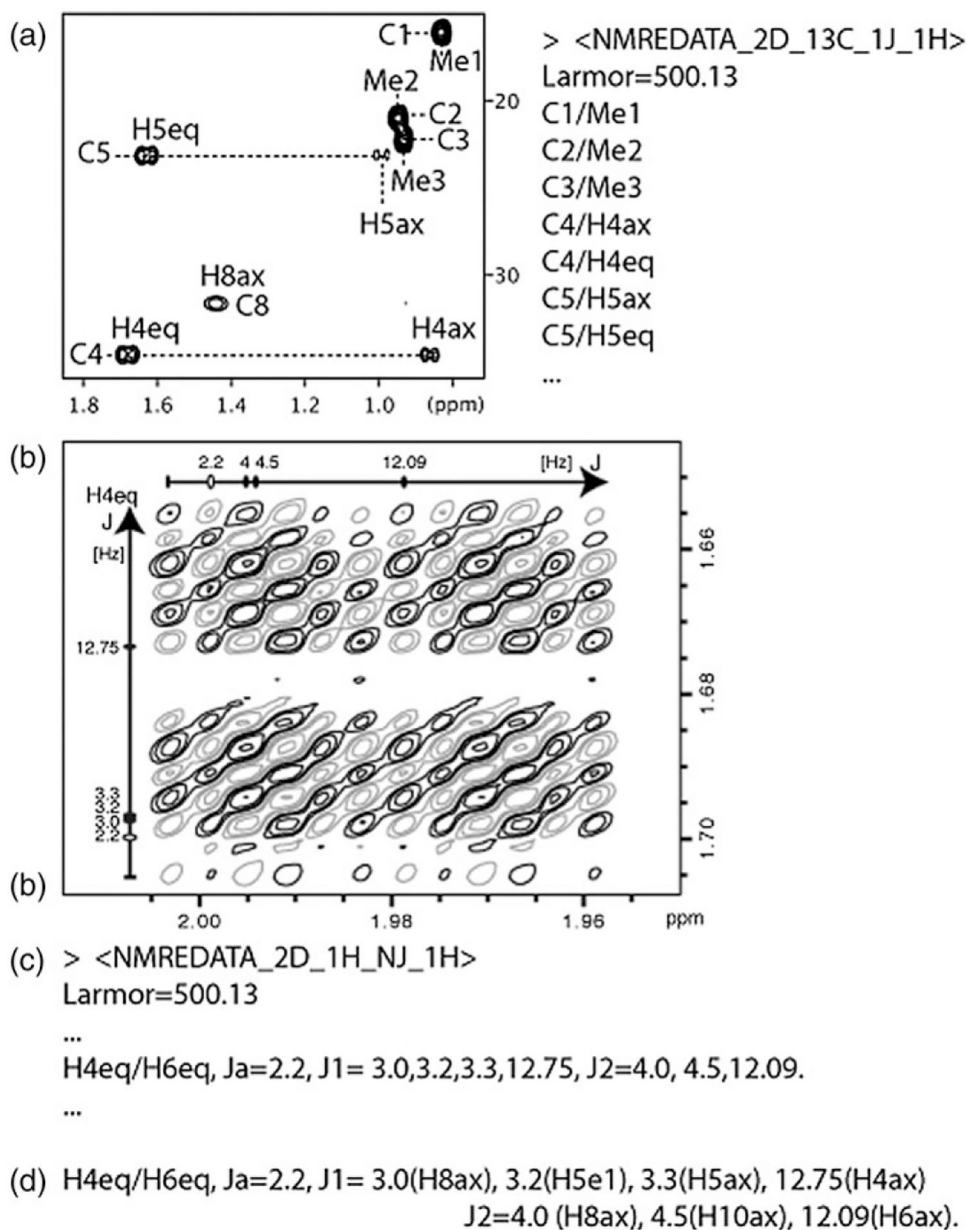
(g) 3.4302, S=dddd, L=H10ax

(h) 3.4302, S=dddd, L=H10ax, J=10.90, 9.90, 4.50, 4.80

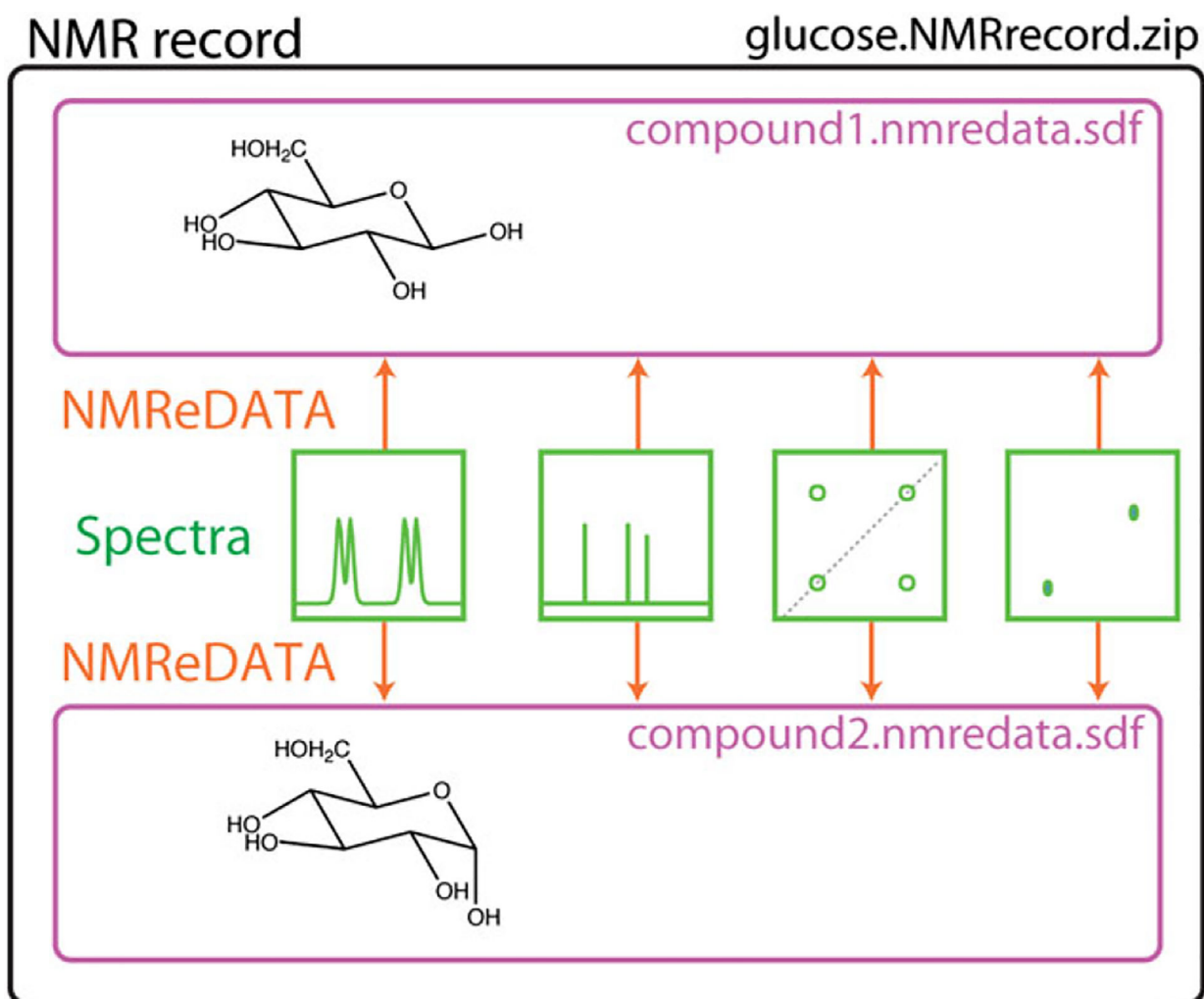
(i) 3.4302, S=dddd, L=H10ax, J=10.90 (H6ax), 9.90(H9ax), 4.50(H6eq), 4.80(10OH)

FIGURE 6.

Examples of descriptions of the same 1D ^1H signal. (a) Multiplet of proton H10 of menthol. (b) Tag name and spectral characteristics of the 1D ^1H spectrum. (c) Signal defined as a chemical shift range with the crude integral. (d) Normalization of the integral to one proton. (e) Signal identified with a single chemical shift with an “m” multiplicity. (f) Signal with only its peak intensity, typical for 1D ^1H decoupled ^{13}C spectra. (g) Signal identified as a “dddd” and assigned to the proton H10ax. (h) The coupling constants measured in the multiplet are listed and assigned in (i)

**FIGURE 7.**

Examples of NMRReDATA tag of 2D spectra of menthol. (a) Assignment of region of interest of a 2D HSQC spectrum. (b) Cross peak between the protons H4eq and H6eq of a high-resolution DQF-COSY spectrum illustrating how coupling constants can be included in the NMRReDATA without (c) or with assignment (d)

**FIGURE 8.**

Structure of the NMR record of a sample containing the two slowly exchanging isomers of glucose. Two NMRReDATA files (`compound1.nmredata.sdf` and `compound2.nmredata.sdf`) corresponding to the two anomers of glucose are included in the NMR records. They both refer to the same set of spectra

TABLE 1

NMReDATA tags names and content (Version 1.1)¹

Name	Content
NMReDATA_VERSION	1.1
NMReDATA_SOLVENT	Solvent, or mixture of solvent with their proportions, etc.
NMReDATA_ID	DOI of the NMR record
Doi = ...	URL pointing to the directly accessible zipped record
Record = ...	Location of the nmredata.sdf file relative to the root of the NMR record
Path = ...	Other references can be added by the software packages generating the files and the databases storing them.
...	
NMReDATA_TEMPERATURE	Temperature of the sample during the NMR acquisition (in Kelvin).
NMReDATA_CONCENTRATION (optional)	Concentration in mM
NMReDATA_PH (optional)	pH
NMReDATA_LEVEL	Specify the complexity of the content of the NMReDATA files ²
NMReDATA_FORMULA	Chemical formula of the compound
NMReDATA_INCHI	InChI string of the chemical structure
NMReDATA_ALATIS	Standard InChI string of the chemical structure as generated by ALATIS
NMReDATA_ASSIGNMENT	Associate the labels used for the assignment to a chemical shift and the atom numbers in the chemical structure
NMReDATA_J (only present if couplings were assigned)	Compile assigned scalar couplings
NMReDATA_ID_Isotope	

Name	Content
NMREDATA_2D_Isotope (F1)_Code of the <i>mixing_Isotope</i> (F2)	List the peaks extracted from 1D spectra (see Table 4)
	List peaks extracted from 2D spectra (see Table 5)

¹ Refer to the website of the initiative for more details. [1]

² A nonzero value indicates the presence of ambiguous assignment (see section “Ambiguities”).

TABLE 2

Nomenclature of spectral tag names

Tag name	Type of spectrum
NMREDATA_ID_1H	1D ¹ H spectrum
NMREDATA_ID_13C	1D ¹³ C spectrum
NMREDATA_ID_13C#2	Second 1D ¹³ C spectrum (for example, for aDEPT spectrum)
NMREDATA_2D_13C_1J_1H	2D spectrum with ¹³ C evolution during <i>t_r</i> , ¹ H evolution during Detection and <i>J</i> mixing (HSQC, or HMQC experiments)
NMREDATA_2D_1H_NJ_1H	2D COSY spectrum

¹ More details can be found on the website of the NMRReDATA initiative. [1]

TABLE 3

Spectral properties (1D and 2D spectra)

Keyword	Spectrum property
Larmor = <i>value in MHz</i>	Larmor frequency of the detected isotope
CorType = <i>type</i>	Type (1H, 13C, DEPT135, COSY, HSQC, etc.)
Decoupled = <i>isotope</i>	List of decoupled isotopes ¹
Nondecoupled = <i>isotope</i>	Denies implicit decoupling ¹
Spectrum_Location =	<i>Path to the spectrum relative to the root of the NMR record</i>
Pulseprogram =	<i>Name of the pulse program</i>

¹ For HSQC and HMQC experiments proton and carbon decoupling is implicit in F1 and F2, respectively. For HMBBC experiments proton decoupling is implicit

TABLE 4

Peak attributes of 1D spectra

Key letter	Peak attributes	Examples
-	Chemical shift	4.1738 - or -
-	Chemical shift range	4.1512–4.1978
S=	Multiplicity	s, m, d, t, dd, etc.
J=	Scalar coupling constants	9.90 9.90 (H9ax)
N=	Number of nuclei	1, 2, 3, etc.
L=	Assignment (label)	H10ax
E=	Integral (crude)	99.87
I=	Signal intensity (relative)	1.67E5
W=	Width at half height (Hz)	1.3
T1=	T ₁ relaxation time (s)	1.54
T2=	T ₂ relaxation time (s)	1.24
Diff=	Diffusion coefficient (m ² /s)	3.1E-9

TABLE 5

Peak attributes of 2D spectra

Key letter	Peak attributes	Examples
-	Assignment of the peak F1/F2	C10/H10
E=	Volume (crude)	7.34E6
I=	Signal intensity (relative)	1.67E5
W1=	Width at half height in F1 (Hz)	2.3
W2=	Width at half height in F2 (Hz)	1.3
Ja=	Active scalar coupling (Hz)	10.9
J1=	Passive scalar coupling in F1 (Hz)	5.60
J2=	Passive scalar coupling in F2 (Hz)	7.34
		9.90(H9ax)
T1=	T ₁ relaxation time (s)	1.54
T2=	T ₂ relaxation time (s)	1.24
Diff=	Diffusion coefficient (m ² /s)	3.1E-9

TABLE 6

Ambiguities in the assignment tag

	Example	Meaning
1	Interchangeable = Ha, Hb	The spins Ha and Hb may be interchanged
2	Interchangeable = (CH3a, Ca), (CH3b, Cb)	The spins of CH3a together with Ca, may be interchanged with those of the (CH3b, Cb)

TABLE 7

Ambiguities in spectral tags

Example	Meaning
1 1.2324, L = Ha	Unambiguous assignment of 1D signal
2 1.2324, L = (Ha Hb)	The signal is assigned to either Ha or Hb.
3 Ha/C1	Unambiguous assignment of a 2D signal
4 Ha/(C1 C2)	The signal is assigned to C1 and/or C2 in the F1 dimension
5 (Ha Hb)/(C1 C2)	The signal is assigned to one or more of the four possible combinations Ha/C1, Ha/C2, Hb/C1 and Hb/C2

TABLE 8

Pure text translation of nmredata.sdf files

	nmredata.sdf file	Pure text equivalent
1		NMReDATA(VI)
2	MOL block	Dot = <i>doi</i> ; record = <i>wr</i> ⁵ ; path = <i>path of sdf file</i>
3	ASSIGNMENT tag	ALATIS InChI code of the molecule
4	Ha, 1.3453, 10	ChemShifts: <10> ⁷ = 1.3453 ¹
5	Ha, 1.3453, H10	⁷ = 1.3453 (for implicit hydrogen atoms) ¹
6	Hb, 1.5442, 4, 5, 6	<4,5,6> ⁷ = 1.5442 ¹
7	J tag	J:
8	Ha, Hb, 7.50	<10,4,5,6> ⁷ = 7.50 ¹
9	ID_ isotope tag	ID(Type ⁶ , Larmor frequency ² , solvent ³ , temperature ⁴):
10	1.3453, S = <i>text1</i> , N = <i>text2</i> , J = <i>text3</i> , L = <i>text4</i> , ...	
11		1.3453 (<i>text1</i> , < <i>text4</i> > ⁷ , <i>text2</i> H, J = <i>text3</i>) ¹
12		Path = <i>path</i> ⁵
13	2D_ iso1_ mix_ iso2 tag	2D(CorType ⁸ , iso1, iso2, Larmor frequency ² , solvent ³ , temperature ⁴):
14	text1/text2	"text1"/"text2" ¹
15		Path = <i>path</i> ⁵

¹ A comma followed by a space is added as separator between entries. A period followed by a space is added after the last entry.

² From the "Larmor = "property of 1D and 2D spectra.

³ From the NMReDATA_SOLVENT tag.

⁴ From the NMReDATA_TEMPERATURE tag.

⁵ From the "Spectrum_Location" properties of 1D and 2D tags.

⁶ Either the isotope of the spectrum (1H, 13C; without superscripts) or the name of the pulse sequence (DEPT135, DEPT90, etc.).

⁷ In the absence of a comma to list multiple atoms, the "<" and ">" delimiters are not mandatory.

From the "CorType" = "of property of 2D tags."

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript