Received: 8 December 2023 | Revised: 16 February 2024 | Accepted: 17 February 2024

DOI: 10.1002/mrc.5441

RESEARCH ARTICLE

WILEY

SimpleNMR: An interactive graph network approach to aid constitutional isomer verification using standard 1D and 2D NMR experiments

Eric Hughes | **Alan M. Kenwright**

Department of Chemistry, University of Durham, Durham, UK

Correspondence
Eric Hughes, Department of Chemistry, University of Durham, Durham, UK.
Email: eric.hughes@durham.ac.uk

Abstract

Despite progress in computer automated solutions, constitutional isomer verification by NMR using one- and two-dimensional data sets is still, in the main, a manual, user-intensive activity that is challenging for a number of reasons. These include the problem of simultaneously keeping track of the information from a number of separate NMR experiments and the difficulty of another researcher subsequently verifying the assignments made without having to independently repeat the whole analysis. This paper describes a graphical interactive approach that overcomes some of these problems. By using concepts used to visualise graph networks, we have been able to represent the NMR data in a manner that highlights directly the link between the different NMR experiments and the molecule of interest. Furthermore, by making the graph networks interactive, a user can easily validate and correct the assignment and understand the decisions made in arriving at the solution. We have developed a usable proof-of-concept computer program, 'simpleNMR', written in Python to illustrate the ideas and approach.

KEYWORDS

graph network, liquid state, NMR, python small molecule analysis, software, structure verification

1 | INTRODUCTION

One- and two-dimensional NMR spectra are now, along with high-resolution mass spectrometry, among the most important routine techniques employed to confirm the structures of molecules synthesised in research laboratories, particularly in universities where the novelty of molecules often makes it impractical to confirm structures by comparison with literature data.

In synthetic organic chemistry, a typical set of NMR experiments for structure verification would probably include 1D-^{1}H, 1D-^{13}C, ^{1}H-COSY, ^{1}H-^{13}C-HSQC and ^{1}H-^{13}C-HMBC. Depending on the molecule, the data set might also include some form of multiplicity-edited experiment to help distinguish the number of directly attached protons on each carbon. This might take the form of an APT spectrum, one or more DEPT spectra, or a multiplicity-edited variant of the HSQC experiment.

Other common experiments include some variety of TOCSY experiment to help resolve issues due to signal overlap, and some form of NOE experiment to assist with determination of stereochemistry (either one-dimensional selective NOE experiments or two-dimensional NOESY or ROESY).

Advances in NMR techniques and hardware mean that such multi-experiment data sets can now be produced routinely and in relatively little time.[1] Unfortunately, advances in techniques for analysing and interpreting the data have not progressed at the same rate, with the result that data analysis can now be something of a bottleneck. In universities, the initial interpretation of the NMR data is often undertaken by research students, who may not be the most skilled or experienced practitioners of the art. Of course, their assignments should ultimately be checked by their supervisors, but this presents a further problem in that checking someone else's assignment of NMR data effectively involves redoing the assignment from scratch, and so this is not always done as often or as thoroughly as it should be, with the result that the literature contains many erroneous assignments.[2]

1.1 | Constitutional isomer verification problem

The first part of the interpretation and assignment of such an NMR data set is usually the process of constitutional isomer verification. In this process, the researcher checks that the data set is consistent with the proposed constitutional isomer (ignoring stereochemistry) and that an assignment can be made on that basis. There are a couple of points that are worth noting about this process. The first is that it differs from the process taught in most textbooks on NMR interpretation where the emphasis is on deducing the molecular structure given the molecular formula and the set of NMR spectra.[3] That procedure and the associated skills certainly have an important role when the structure is unknown, but in most cases, the synthetic chemist thinks that they know what the structure is and attempts to reconcile the spectral data set with that structure. It could be debated whether this is the 'correct' approach to take, but, regardless of the view taken on that, experience suggests that this is what actually happens. The second is that the initial process of constitutional isomer verification takes no account of stereochemistry. This may seem odd, as in many cases, the stereochemistry is an important aspect of the desired structure, but it is the case, even if the chemist is unaware that constitutional isomer verification is a separate first step, since NOE responses or characteristic ^{1}H-^{1}H coupling constants only become interpretable once the positions in the molecule of the hydrogen atoms giving rise to the relevant signals are known.

1.2 | Traditional approach

The 'traditional' method for interpreting such a data set is to identify one or more signals in the proton and/or carbon spectra that can be unambiguously assigned by virtue of their NMR parameters (chemical shift, integral and coupling). Typically, these might be something like a methyl group or a carbonyl. Obviously, this will depend on the nature of the molecule under investigation. Having made an initial assignment, the correlations in the two-dimensional spectra are used to suggest further assignments, until a complete assignment consistent with the proposed constitutional isomer is achieved. While this works reasonably well for smaller molecules, it becomes more problematic as the size of the molecule increases, for two reasons. The first is that it is very much a sequential stepwise process with each assignment being considered largely in isolation and the correctness of later assignments depends on the correctness of the earlier assignments. The second is that there is usually no record of the steps used in the chain of deduction and therefore no easy way for another researcher to evaluate the quality or 'correctness' of the assignment.

Sometimes, this process is augmented by comparing the assigned carbon shifts with predictions, usually based on databases. However, this can also be problematic, as it is not easy to evaluate the quality of the predictions and to spot those that might be erroneous or misleading. Furthermore, revising and/or changing assignments is not straightforward due to the stepwise nature of the assignment process. If it is suspected that the assignment process may have gone wrong, it is difficult to know how many steps it is necessary to backtrack, so, often, it is easier to start the whole process over again.

1.3 | simpleNMR solution

In an attempt to address some of these problems (traceability, using data in isolation, solving the problem in a sequential manner and verification by third parties), we propose a method for making and evaluating assignments based on the basic multi-experiment data set outlined above via a computer program, 'simpleNMR', that augments the chemist's approach of taking a proposed structure (constitutional isomer) and sequentially verifying that the NMR data agrees with it. In order to do this, the solution we propose relies heavily on treating the

information (NMR correlations and proposed structure) in terms of graph networks, both to represent the NMR information in the computer and to present it to the user visually and dynamically so that the user can see the NMR information as a whole. In this context, a graph network consists of 'nodes', each of which contains several pieces of related information or data. The nodes are connected by 'edges', each of which defines the relationship or interaction between two nodes. In this proposed method, the intrinsic latent networks of the NMR correlations are made apparent by overlaying them on a two-dimensional skeleton network representation of the proposed molecule (a mode of representation familiar to most chemists), providing a new approach to the assignment problem that is easily verifiable by third parties. This allows the user to work with networks of correlations that highlight anomalies in the assigned solution in an intuitive, visual manner.

The display is interactive with network information such as HMBC correlations displayed dynamically when the mouse hovers over a node. A node represents a carbon atom and the number of protons attached to it. In effect, it represents a carbon-proton fragment (CH_x) of the molecule. The value of x ranges from 0 to 3. The number of protons attached is shown by colour coding the node. The NMR correlations that link the nodes are represented by graph edges between nodes. Nodes (CH_x fragments) can be dragged over the canvas and positioned over the skeleton network image of the proposed molecule to reach an assignment. The position of the nodes can be readily changed to see how the network correlations change, which further aids assignment.

The first graph network describes the reduced structure of the proposed constitutional isomer where the nodes represent atoms and the edges represent bonds between the atoms. We work with the reduced constitutional isomer using implied, rather than explicit, hydrogens and ignore stereochemistry. This is closely related to a common way of representing organic structures that is familiar to synthetic chemists. We derive this from the chemical structure, supplied as a SMILES string, but we are only interested in the positions of the carbon atoms in the molecule. The SMILES string is used to define a molecule in RDKIT (www.rdkit.org), which is effectively a graph network. We then use that to draw a two-dimensional representation of the molecule structure from which we can determine the on-screen coordinates of the carbon atoms in the molecule. These are the reference points to which we can subsequently relate the positioning of the CH_x fragments.

The second graph network describes the most salient features of the NMR data set. The nodes correspond to CH_x fragments where $x = 0$, 1, 2 or 3. The node contains information about the chemical shifts of the carbon atom and any hydrogen atoms involved, as well as the integral and information about ^1H-^1H coupling constants if that information is available (see discussion below). Note that the information content of the HSQC experiment (one-bond coupling between carbons and hydrogens) is already represented within the nodes, so need not be considered subsequently. Other correlation experiments (COSY and HMBC) can then be used to determine interactions between the defined CH_x fragments. These interactions can be displayed as the edges in a graph, connecting the carbons of the relevant CH_x fragments, and so making it possible to relate this information to the structure defined by the first graph.

To correctly define the nodes, it is necessary to be able to identify the number of hydrogens directly bonded to each carbon. There are various methods that can be used to achieve this depending on the dispersion and complexity of the proton spectrum, and these will be discussed below. The edges, as mentioned previously, correspond to the correlations (COSY and HMBC) between the nodes, but this representation is less than fully detailed for two reasons. First, if the hydrogens of a methylene group are diastereotopic and that group shows a COSY correlation to a neighbouring group, the representation contains no information about which of the diastereotopic methylene hydrogens is/are involved in the COSY correlation. Second, if the representation shows an HMBC correlation between two CH_x fragments, A and B, where x is greater than 0 for both fragments, there is no indication whether the correlation is between the hydrogen(s) of A and the carbon of B or between the carbon of A and the hydrogen(s) of B. It may therefore seem that we are losing information by representing the data in this way, but we have found that the representation is adequate in all the cases we have tested (more than 100 molecules).

Having constructed the two graphs, we overlay the first graph on a conventional depiction of the structure derived from the smiles string provided, thereby defining the position of each carbon atom in the molecule in screen coordinates. We then seek to reconcile the second graph with it by interactively positioning the nodes representing the CH_x fragments on the structure. The nodes are colour coded to represent the value of x (0, 1, 2 or 3), and correlations are displayed. COSY correlations are permanently displayed in red, while HMBC correlations for a particular node are displayed in grey if the mouse pointer is hovered over that node. Displaying all the HMBC correlations simultaneously is generally confusing and unhelpful. At the same time, related spectroscopic information is displayed in 'pseudo' ^1H and ^{13}C spectra synthesised from information provided as input to the program.

There are a number of possible scenarios for completing the positioning of the nodes and these are discussed below, but once a satisfactory positioning of all the nodes has been achieved the result can be saved and a report generated showing the assignment of ^{1}H and ^{13}C chemical shifts to atoms in the structure. A major advantage of the method is that it is easy and quick for someone else to subsequently re-load the saved data and confirm that the positioning of the nodes makes sense both in terms of chemical shift and in terms of correlations between the nodes. It is also easy to try alternative positioning (alternative assignments) in cases where it is felt that the positioning is not optimal. Thus, the task of checking or confirming someone else's assignments becomes relatively trivial.

We should emphasise that the program uses the correlations revealed by the two-dimensional NMR experiments to construct a network relative to at least one initial assignment provided by the user. In this respect, it simply provides a much easier graphical method of achieving what is the usual workflow undertaken by the synthetic chemist. When the positioning of nodes has been satisfactorily completed, it provides confirmation that the pattern of correlations and the initial assignment(s) are consistent with the postulated constitutional isomer. This is not formally the same as structure verification since, in principle, multiple constitutional isomers may be consistent with the correlation network. However, in practice, the number of possible structures is limited by knowledge of the synthetic pathway used, and perusal of the literature shows that the level of verification that the program provides is often regarded, in the context of other supporting data, as sufficient to establish the identity of the compound. However, if there is a need to provide more rigorous analysis of the NMR data, there are a number of other computer programs aimed at structure elucidation and/or structure verification using NMR data,[4] and, in particular, the LSD software developed by Nuzillard and Plainchont[5] allows for a more detailed correspondence between NMR spectral features and structure, albeit at the cost of considerably increased complexity for the user.

2 | RESULTS AND DISCUSSION

We have implemented the ideas outlined above in a computer program called simpleNMR that can be downloaded online (from https://github.com/EricHughesABC/simpleNMR). The program is written mostly in python but also makes use of some third-party java code to deal with the carbon chemical shift prediction module (see below). The program is a proof of concept and is still under active development, but it is fully functioning, and the Github repository provides executable files for both Windows and MacOs for simple installation. The discussion in this paper will therefore emphasise the concepts in play rather than practical details relating to the program. More detailed information about how to use the program can be found on the Github pages.

2.1 | NMR data entry

The first part of the process in applying the new method is to extract the relevant information from the NMR spectra. At the current state of development, we choose to use Mnova software[6] to do the data processing and use the output from that as input to the program. The reasons for this were partly availability (the software is widely available to researchers at Durham) and partly because it provides a 'vendor neutral' platform for data processing so we could handle and evaluate raw NMR data from a range of sources without having to develop separate data input protocols.

The core experiment of the method is the HSQC, since that is the experiment that defines the direct bonding of hydrogens to carbons and so allows us to identify the CH_x fragments used in the program and to relate the correlations of those fragments to the positions of their constituent carbon atoms. The presence of an HSQC experiment is therefore essential for the program to work. Other spectral information is also necessary for the program to reach a satisfactory complete assignment, but exactly what information is needed may vary with the molecule in question, so the program will attempt to process information derived from any of the following experiments: 1D-^{1}H, 1D-^{13}C, 1D-^{1}H-PureShift, ^{1}H-^{1}H COSY and ^{1}H-^{13}C HMBC. Note that it is not necessary for all of those experiments to be present. The program will attempt to work with whatever subset it is presented with as long as the HSQC is present. Having said that, it is, of course, obvious that the more useful data the program is provided with, the easier it should be to determine a correct assignment.

Since we are aiming to combine information from different experiments, it is vital that the referencing is consistent between the various experiments. We are also aiming to extract the chemical shifts of peaks and correlations as accurately as possible, so it is important that the spectra are processed in such a way as to maximise the available digital resolution. Practical suggestions for how to do this and how to 'peak pick' the chemical shifts of the correlations in the 2D experiments in a consistent manner are provided in the supporting information.

Having picked the relevant information in Mnova, we need to transfer that information to the program. In the first instance, we chose to do this by constructing an excel file that could subsequently be read by the program. The excel file contains several worksheets, one for each NMR experiment plus a worksheet to hold the SMILES string representing the proposed structure. Apart from the worksheet containing the SMILES string, they are produced simply by copying the information in the relevant peak/integral tables in Mnova and pasting that into the excel sheet. Again, examples and guidance on how to do this are provided in the supporting information. Note that, except for the HSQC data, it is not necessary to 'peak pick' every peak in the 2D spectra. For example, in COSY spectra, the diagonal peaks contain no information about correlations, and pairs of peaks that are symmetrical about the diagonal merely duplicate information, so it is possible to get all the information required by only picking a selected subset of the peaks.

In the following discussion, we will assume that in addition to the HSQC data, we also have 1D-^{1}H, 1D-^{13}C, ^{1}H-^{1}H COSY and ^{1}H-^{13}C HMBC spectra, as this is a common combination of experiments. When approaching constitutional isomer verification using such a data set, there are a number of valuable, simple, preliminary steps that should be observed before even attempting an assignment, none of which involve any detailed interpretation of the spectra, given the fact that spectroscopy is not very good at proving structures but is excellent at disproving them. So before attempting an assignment, we should at least check the following. Does the 1D-^{13}C spectrum contain the expected number of signals? Does the number of carbon signals involved in correlations in the HSQC spectrum correspond to the number of carbons with directly bonded hydrogens in the structure? Do the integrals in the 1D-^{1}H spectrum make sense? If the answer to any of those questions is no, then there is no point in attempting an assignment, at least until the problems revealed by those questions have been resolved so that the answer to all three is yes.

Before going on to look at the operation of the program itself, there are some points that are worth making about experiment selection and data processing in relation to the program. One of the key features of the program is that it deals with CH_x fragments where $x = 0$, 1, 2 or 3, so, obviously, the ability to correctly determine x is vital. One way to facilitate this is by using a multiplicity-edited version of the HSQC experiment, such that CH and CH_3 correlations have one phase (say positive) while CH_2 correlations have the opposite phase (say negative). If such an experiment is provided, then it is easy for the program to identify quaternary carbons ($x = 0$) since they appear in the 1D-^{13}C spectrum but not in the HSQC, and to identify CH_2 groups (negative intensity in the HSQC correlation). Of the remaining carbon signals, any with a chemical shift greater than 67 ppm can be confidently assigned as CH signals since no methyl groups come higher than 67 ppm.[7] Then, it only remains to distinguish between CH and CH_3 signals below 67 ppm. When dispersion is good and there is little or no peak overlap in the 1D-^{1}H spectrum, this can be done straightforwardly from the integrals in the proton spectrum. However, if overlap in the proton spectrum is significant, as it often is in larger molecules, then additional experiments to facilitate the necessary distinction may be considered. These include other varieties of multiplicity-edited HSQC to show only CH correlations[8] or quadruple-quantum HSQC to show only CH_3 correlations.[9] In the vast majority of examples we have looked at so far, we have not found it necessary to include additional experiments to make this last distinction, but it remains an active area of investigation as we seek to determine the optimum set of experiments to use with the program. When the program encounters two HSQC correlations with the same carbon chemical shift and with the integral of the corresponding proton signals each equal to one hydrogen, it assumes that they correspond to a diastereotopic methylene pair only if a multiplicity-edited version of the HSQC experiment has been used (the preferred situation) by utilising the phase of the correlations. If the HSQC is not a multiplicity-edited version, then the user needs to manually identify the correlations as belonging to a diastereotopic methylene pair by setting the intensities to negative values in the HSQC correlation table. The ^{2}J COSY correlation between such a diastereotopic methylene pair is redundant information and is not utilised by the program.

The second point concerns data processing and, in particular, the positioning of integral regions in the 1D-^{1}H spectrum. Where peak overlap occurs, there may not be an obvious point at which to split the integral in the overlapped region based solely on the appearance of the proton spectrum. But if an overlapped region with, for example, a total integral corresponding to two hydrogens shows HSQC correlations to two separate carbons and the proton shifts of the HSQC correlations are significantly different, then there is a case to be made for splitting the integral in the middle of the overlapped region to give two integrals each corresponding to one hydrogen. The selected regions may not correspond exactly to individual proton resonances, but the information content is correct as far as the program is concerned.

It is worth pointing out that not all cases of overlap can be resolved in this methodology, any more than they can using conventional assignment techniques. To take an extreme example, if the sample in question is a

saturated long-chain fatty acid the majority of the CH_2 groups in the molecule will have effectively identical chemical shifts in both the proton and carbon spectra, so assignment corresponding to each individual CH_2 group is impossible.

The question of overlap also impacts on which peaks should be picked in the COSY and HMBC spectra. Consider an overlapped peak in the 1D-^{1}H spectrum with an integral corresponding to five hydrogens, and the HSQC spectrum shows correlations from that peak to four separate carbon resonances, which we can allocate as 1 × CH_2 and 3 × CH's. There is no point in picking a correlation to that five-hydrogen overlapped peak in the COSY spectrum since it will be impossible to tell which CH_x fragment it relates to. The same applies to correlations in the HMBC involving that signal. The only sensible thing to do in that case is to not pick those correlations, since picking them and passing that data to the program will result in ambiguous and/or misleading correlations being displayed on screen. This may seem problematic since we are effectively leaving out correlations, but the fact is that in most cases the number of observed correlations (COSY and HMBC combined) is far greater than is necessary to make a well supported assignment, so we can afford to leave out any that are ambiguous. Indeed, in the case of HMBC, our guideline should be to only pick correlations to proton signals that are not overlapped or at least only to proton signals that have HSQC correlations that are clearly resolved in the proton dimension.

The problem of peak overlap is possibly more significant for COSY experiments than for HMBC, but that may be alleviated by making use of the increased dispersion available in heteronuclear experiments such as the recently described HSQC-Clip-COSY.[10] The benefits of such experiments in this context are the subject of ongoing evaluation.

2.2 | Interacting with the program

Having selected the correlations that are unambiguous, and constructed the required excel file, we can use the program to position the CH_x fragments on the structure and confirm their positions using the correlations. There are three options offered when the program opens a new excel file (new problem) and the most basic of these is the use 'random' positions for the CH_x fragments. This displays the fragments with those connected by COSY correlations adjacent to each other and the remaining fragments distributed in a semi-random fashion as a starting point for the assignment. In order to demonstrate how the program works, we have included a small number of examples in the program that can be downloaded from Github and will use some of those examples to illustrate the way(s) in which the program can work. The first example we have chosen is 2-ethyl-1-indanone. Figure 1 shows the program interface using 'random' positioning. The large window to the left shows the structure with the CH_x fragments displayed in 'random' positions but with COSY correlations shown in red. The large window to the right shows representations of the 1D proton and carbon spectra synthesised from the data provided in the excel file. These are linked by the program with the CH_x fragments displayed in the window on the left so that hovering the mouse pointer over a CH_x fragment in the window on the left causes, inter alia, the corresponding signals in the spectra in the window on the right to be highlighted, and vice-versa. The far-right section of the right-hand window also displays information of likely (common) functional environments corresponding to the highlighted chemical shifts, that is, the sort of information typically found in simple correlation charts. The CH_x fragments are numbered in order of decreasing carbon chemical shift, so in this molecule, Fragment 1 corresponds to the carbon signal at 208.9 ppm, and Fragment 11 corresponds to the carbon signal at 11.6 ppm. The CH_x fragments in the window on the left are also colour coded according to the value of x, so quaternary fragments ($x = 0$) are coloured blue, methyne fragments ($x = 1$) are coloured orange, methylene fragments ($x = 2$) are coloured green and methyl fragments ($x = 3$) are coloured purple.

From this position, the first requirement, as in the 'traditional' method outlined above, is to make a first assignment. In the case of this molecule, there is only one methyl group so that makes an ideal first assignment. Having dragged the methyl (purple) CH_x fragment to the appropriate position on the structure, we can easily position the fragments that are linked by COSY correlations, paying attention to respect the correct values of x in case any of the COSY correlations extend over more than three bonds. Having done that, we can check that the HMBC correlations make sense by hovering the mouse pointer over each assigned fragment sequentially and checking that there are no strangely long (greater than three bonds) or unassigned HMBC correlations. Note that since the CH_x fragments contain within themselves the information about the direct C-H bond, a two-bond HMBC correlation will only link nearest neighbour carbon atoms in the structure, while a three-bond HMBC will correspond to next nearest neighbour carbon atoms in the structure. Figure 2 shows the position after the steps outlined above with the mouse pointer hovering over the methylene group (C10) closest to the methyl group (C11).

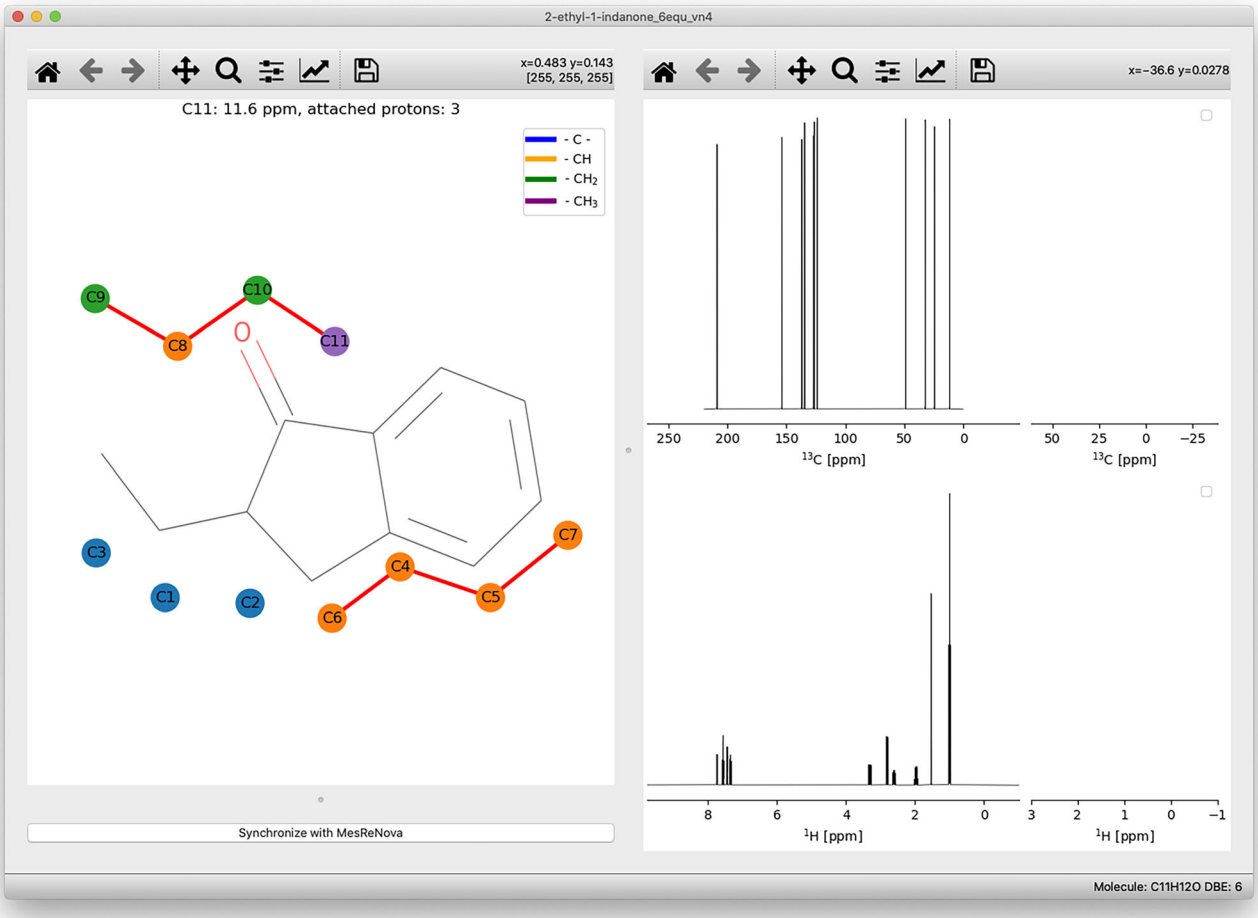

FIGURE 1 The program interface showing 'random' positioning of the CH_x fragments for 2-ethyl-1-indanone. This figure is a screen shot showing the layout of the whole screen. Subsequent figures will only show the structural elements in the left-hand window to aid clarity. A video showing the same process is provided in Movie S1.

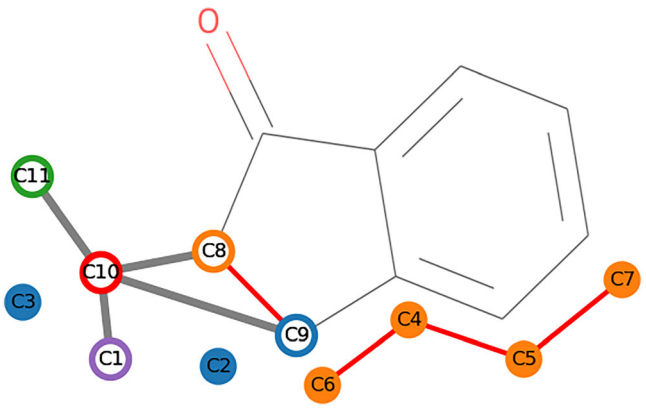

FIGURE 2 Situation after initial placement of COSY linked fragment of 2-ethyl-1-indanone with the mouse pointer hovering over the methylene group adjacent to the methyl group

The HMBC correlations are shown as grey lines, and the CH_x fragments that they correlate with are now shown as white circles with variously coloured borders/rims. The corresponding signals are highlighted in the pseudo 1D ^{13}C and ^{1}H spectra in the same colours as the rims. In this case, there are four HMBC correlations displayed. Three are to CH_x fragments that have already been positioned using the COSY correlations, and the HMBC correlations are all found the be reasonable (to neighbour or next nearest neighbour fragments, corresponding to two- or three-bond HMBC correlations). The fourth HMBC correlation is to a quaternary carbon (C1) that has not yet been assigned (purple rim) and the corresponding highlighted peak in the carbon spectrum is at more than 200 ppm. This must be the carbonyl carbon, both by HMBC proximity and by virtue of its characteristic chemical shift. We can therefore position that with confidence. Hovering the mouse pointer over that fragment (C1) once it has been positioned reveals an HMBC correlation to an unassigned CH fragment (C7) at the end of a COSY chain, which must be the aromatic CH closest to the carbonyl (Figure 3). Using the COSY

correlations, we can therefore easily position the remaining CH fragments in the aromatic ring, which leaves only the two quaternary carbons in the aromatic ring to assign, and this can be done using the usual rule of thumb that, in aromatic rings, three-bond HMBC correlations are much more common than two-bond and/or four-bond correlations. Figure 4 shows the structure with all the CH_x fragments positioned and the mouse pointer hovering over one of the aromatic quaternary carbons (C2), showing the characteristic three-bond HMBC correlations.

Once a satisfactory placement of all the CH_x fragments has been found, it can be saved and a report of the shifts generated (see below). If the problem is subsequently revisited, it can be opened using the saved coordinates rather than the 'random' positioning described above. This is an important feature because it makes it very easy for another researcher to not only see the assignment that has been made but also, by hovering the mouse pointer over the various fragments, to see the justification for that assignment. So the assignment can be rapidly checked and verified independently by, for example, a supervisor. In the longer term, it might be possible to deposit such information with publications so that referees could easily and rapidly evaluate the justification for NMR assignments.

2.3 | Using ^{13}C databases

The workflow outlined above still requires some basic knowledge of characteristic NMR shifts and increasingly these days researchers make use of databases and predictive software to suggest characteristic shifts for particular chemical environments, particularly for ^{13}C spectra where the relatively large chemical shift range favours this approach. It is not always easy, however, to evaluate whether the predicted shifts are correct, even when supporting statistics are provided. The simpleNMR program allows the possibility of rapidly checking assignments by evaluating whether the correlations from the 2D experiments make sense and therefore allows us to use predictive software as a starting point, knowing that misassignments are likely to be readily apparent. We have included in the program a copy of the ^{13}C chemical shift prediction software from NMRShiftDB2[11] that generates predicted ^{13}C shifts for the structure described in the SMILES string provided in the excel file. The program then tries to reconcile the predicted shifts with the closest

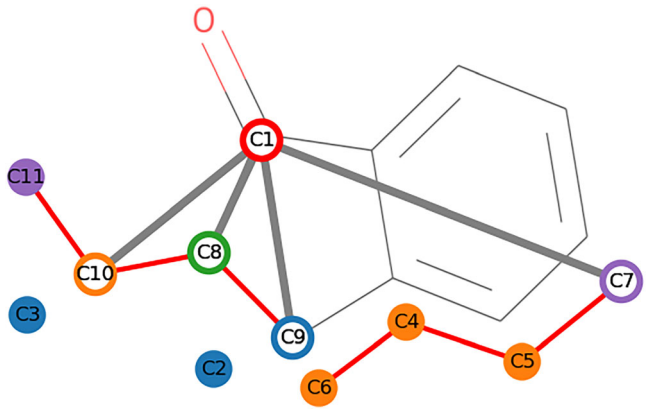

FIGURE 3 HMBC correlations to the carbonyl carbon

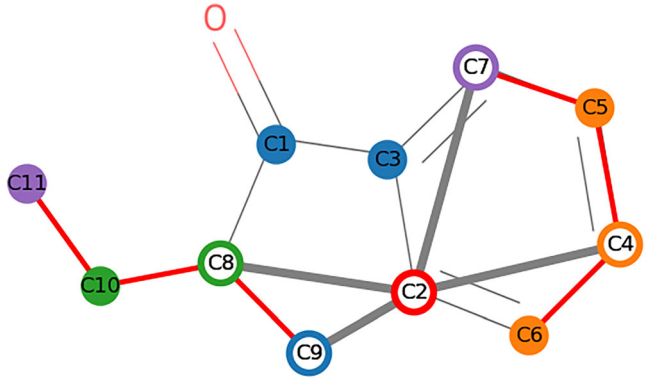

FIGURE 4 Positioning of CH_x fragments completed, with the mouse pointer hovering over one of the quaternary aromatic carbons (C2) showing the characteristic three-bond HMBC correlations in the aromatic ring

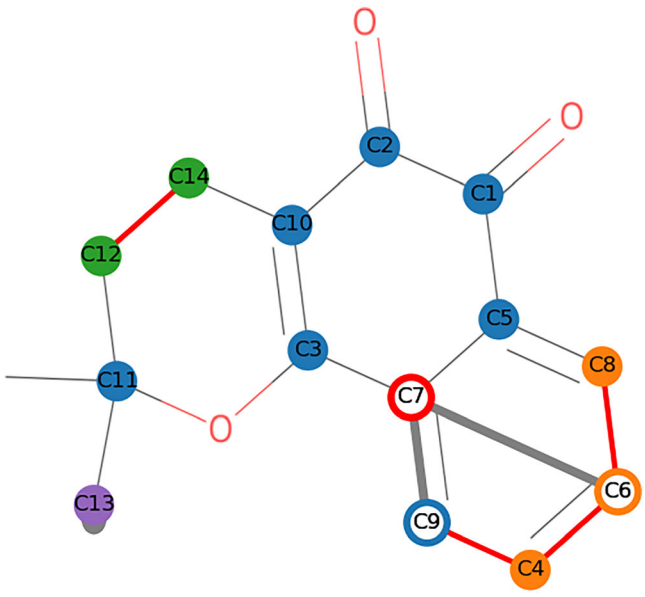

FIGURE 5 Assignments for beta-Lapachone based on ^{13}C chemical shift predictions using the NMRShiftDB2 software. The mouse cursor is hovering over the quaternary aromatic carbon, C7, and showing (probably erroneous) two- and four-bond HMBC correlations. This can be readily rectified by interchanging C7 with the adjacent C5. A movie of this process is provided in Movie S2.

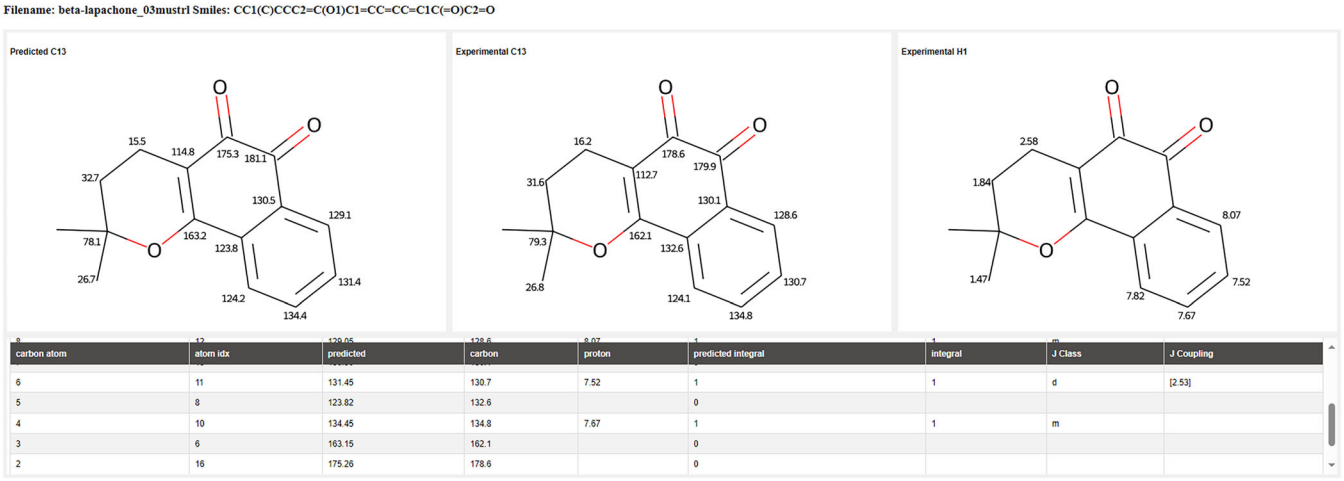

FIGURE 6 Report for the (corrected) assignment of beta-Lapachone, showing the significant discrepancy between the predicted and observed ^{13}C chemical shift for carbon atom 5, one of the quaternary aromatic carbons

experimental values in each of the classes of CH_x fragment and positions the fragments over the structure on that basis. This is actually the default setting for opening a new problem. If we use this approach with the example used above (2-ethyl-1-indanone), the predictions are sufficiently close to the experimental values that the assignment is correct. But this is not always the case.

As an example, we look at the assignment offered for beta-lapachone. Figure 5 shows the assignment based on the predicted ^{13}C chemical shifts. Most of them are correct and can easily be verified, but the assignments of the quaternary carbons in the aromatic ring are unlikely to be correct because, as shown, they result in two- and four-bond HMBC correlations, rather than the expected three-bond correlations. This can be readily corrected by interchanging the assignments of the quaternary carbons in the aromatic ring (C5 and C7 in this case). Once we have done that and achieved a satisfactory positioning of the fragments, we can save the assignment and generate the report. An example report for this molecule (with the corrected assignment) is shown in Figure 6. The report shows both the predicted ^{13}C shift and the experimental ^{13}C shift, so that a rapid evaluation of the prediction can be made. In this case, it is clear that the prediction for one of the quaternary aromatic carbons differs from the observed shift by a significant amount (about 7 ppm), but all of the other predictions are in reasonable agreement with the observed shifts. So the use of predicted ^{13}C shifts often provides a useful starting point for an assignment, but needs checking and in some cases correcting.

Chemical shift prediction is an area of active development, and some may feel that other prediction software would perform better. The NMRShiftDB2 software was chosen in this context because it is open source and freely available. It may be the case that other prediction software would perform better in some circumstances (certain classes of compound) and could be accessed by the program in the future, but the fundamental point is that it is often difficult to evaluate the quality of predictions, so the simpleNMR program offers a ready way to evaluate predictions by comparing assignments based on the predictions with the observed correlations.

3 | CONCLUSIONS

We have presented a method to aid constitutional isomer verification by evaluating a combined data set consisting, typically, of 1D-^{1}H, 1D-^{13}C, ^{1}H-^{1}H COSY and ^{1}H-^{13}C HMBC spectra. The method works by seeking to identify CH_x fragments where $x = 0$, 1, 2 or 3, and then using a graphical interface to position those fragments on a representation of the molecular structure while displaying COSY and/or HMBC correlations between the fragments. We have written a computer program, 'simpleNMR', to implement these ideas, that is available to download online (https://github.com/EricHughesABC/simpleNMR). The advantages of the method are (1) that it deals in concepts that are familiar to synthetic chemists (CH_x fragments and molecular structures), (2) that it makes it easy to take full account of all of the available correlation information simultaneously, (3) that the assignments made (positioning of CH_x fragments) can be stored and recalled making it easy for others to check and validate the assignment and (4) predictions of ^{13}C chemical shifts can be used to suggest an initial assignment which can then be easily checked and corrected if the suggested assignments are not correct.

It is also worth pointing out that the underlying structures used in the program (network graphs) are widely used in machine learning applications, raising the possibility of applying that technology to help in refining initial assignments based on ^{13}C chemical shift predictions.[12,13]

It is not easy to adequately demonstrate the functioning of the program in a journal article, so interested readers are encouraged to download a copy of the program (executables for PC and Mac are available in addition to the source code) and explore it for themselves. The examples discussed in this article are built in to the program.

ACKNOWLEDGEMENTS

We would like to thank Catherine Heffernan and Juan Aguilar-Malavia of the Durham University NMR service for their helpful encouragement and patience.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

AUTHOR CONTRIBUTIONS

Eric Hughes started the project and programmed simpleNMR. Alan Kenwright contributed key design features to the program, tested and produced the majority of the example data and wrote the first draft of the paper.

PEER REVIEW

The peer review history for this article is available at https://www.webofscience.com/api/gateway/wos/peer-review/10.1002/mrc.5441.

ORCID

Eric Hughes https://orcid.org/0000-0003-1264-5108
Alan M. Kenwright https://orcid.org/0000-0002-7169-7103

REFERENCES

[1] E. Kupče, T. D. W. Claridge, *Angew Chem. Int. Edit.* **2017**, *56*, 11779.
[2] W. Robien, *Molecules* **2021**, *26*, 3413.
[3] a) N. E. Jacobsen, *NMR Data Interpretation Explained: Understanding 1D and 2D NMR Spectra of Organic Compounds and Natural Products*, Wiley, New York **2016**. b)L. D. Field, H. L. Li, A. M. Magill, *Organic Structures From 2D NMR Spectra*, Wiley, Sydney **2015**. c)A. Randazzo, *Guide to NMR Spectral Interpretation: A Problem Based Approach to Determining the Structure of Small Organic Molecules*, Loghia Di Amoresano Claudia, Rome **2018**.
[4] J. Nuzillard, *eMagRes* **2014**, *3*(4), 287. https://doi.org/10.1002/9780470034590.emrstm1384
[5] J. Nuzillard, B. Plainchont, *Magn. Reson. Chem.* **2018**, *56*, 458.
[6] MestReNova, version 14.3, Mestrelab Research, Santiago de Compostela, Spain, www.mestrelab.com, **2023**.
[7] H.-O. Kalinowski, S. Berger, S. Braun, *Carbon-13 NMR Spectroscopy*, Vol. 221, John Wiley & Sons, Chichester **1988** 367.
[8] T. Parella, J. Belloc, F. Sanchez-Ferrando, A. Virgili, *Magn. Reson. Chem.* **1998**, *36*, 715.
[9] T. Diercks, M. Schwaiger, H. Kessler, *J. Magn. Reson.* **1998**, *130*, 335.
[10] T. Gyongyosi, I. Timari, D. Sinnaeve, B. Luy, K. E. Kover, *Anal. Chem.* **2021**, *96*, 3096.
[11] S. Kuhn, N. E. Schlörer, *Magn. Reson. Chem.* **2015**, *53*, 582.
[12] Y. Guan, S. V. Shree Sowndarya, L. C. Gallegos, P. C. St, P. C. John, R. S. Paton, *Chem. Sci.* **2021**, *12*(36), 12012.
[13] X. Xue, H. Sun, M. Yang, X. Liu, H. Y. Hu, Y. Deng, X. Wang, *Anal. Chem.* **2023**, *95*(37), 13733.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: E. Hughes, A. M. Kenwright, *Magn Reson Chem* **2024**, 1. https://doi.org/10.1002/mrc.5441