JCIM JOURNAL OF CHEMICAL INFORMATION AND MODELING

ACS AUTHORCHOICE

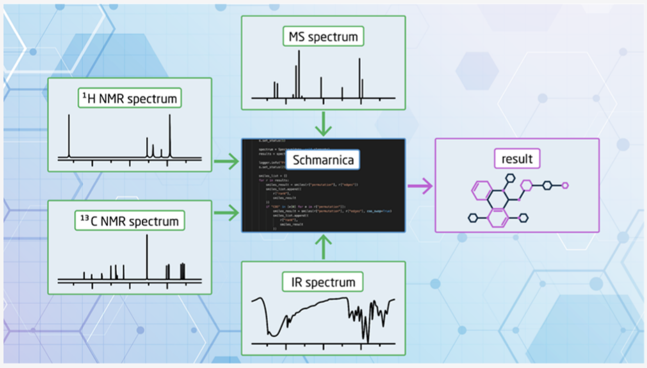
Database Independent Automated Structure Elucidation of Organic Molecules Based on IR, ^{1}H NMR, ^{13}C NMR, and MS Data

Matevž Pesek,* Andraž Juvan, Jure Jakoš, Janez Košmrlj, Matija Marolt, and Martin Gazvoda*

Cite This: *J. Chem. Inf. Model.* 2021, 61, 756−763 Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: Herein, we report a computational algorithm that follows a spectroscopist-driven elucidation process of the structure of an organic molecule based on IR, ^{1}H and ^{13}C NMR, and MS tabular data. The algorithm is independent from database searching and is based on a bottom-up approach, building the molecular structure from small structural fragments visible in spectra. It employs an analytical combinatorial approach with a graph search technique to determine the connectivity of structural fragments that is based on the analysis of the NMR spectra, to connect the identified structural fragments into a molecular structure. After the process is completed, the interface lists the compound candidates, which are visualized by the WolframAlpha computational knowledge engine within the interface. The candidates are ranked according to the predefined rules for analyzing the spectral data. The developed elucidator has a user-friendly web interface and is publicly available (http://schmarnica.si).



1. INTRODUCTION

The idea of using computers to solve chemical structures from experimental spectroscopic data dates from the 1960s.[1,2] Over the past decades, by combining the techniques of chemistry and computer science, a number of structure elucidation systems have been developed,[2−17] accompanied by reports of constant improvements of known and developments of novel impressive algorithms.[18−33] The goal of such expert systems is to determine unknown molecular structures from experimental data with minimal human intervention. Although CASE (Computer Aided Structure Elucidation) programs are beginning to provide very good results in structure elucidation, they are still not entirely automated and usually a number of 2D in addition to the 1D NMR spectra must be provided.[29] A recently reported system for fully automatic processing and assignment of ^{1}H and ^{13}C NMR spectra, that can further elucidate and determine the relative stereochemistry of complex molecules[34] indicates that the development of structure-elucidation systems is still a lively and evolving field.

Most of the reported CASE systems take a database-oriented approach to structural elucidation and therefore heavily rely on databases containing chemical structures and spectra.[2,18,31−33] The absence of structural motifs in databases and mismatching because of experimental differences of the recorded spectra are potential drawbacks of these approaches. To move away from the typical database-oriented computer-assisted elucidation systems, our aim is to develop an algorithm that will be independent of database-searching and would not rely on predefined molecular formulas. We envisage developing an algorithm that will, based on the provided IR and NMR data, first identify a set of small structural fragments and then bind them together based on NMR data, thus building a structure in a bottom-up fashion. The proposed process of structure elucidation mimics a spectroscopist-driven approach of resolving the chemical structure from spectroscopic data.

In empiric elucidation of the structure of an unknown organic molecule, a spectroscopist must combine at least two spectroscopic methods: NMR and MS, NMR or IR, or even all three of them (IR, NMR, and MS) to derive to the correct result because IR, NMR, and MS each give only partial information on the structure of the molecule. While IR reveals some functional groups that are difficult to identify by other methods, it is impossible to determine the connection of structural fragments without NMR, whereas MS is indispensable in verifying the correctness of the proposed molecular structure and for elucidating the missing structural fragments that cannot be determined by IR or NMR. Therefore, we designed an

Received: November 16, 2020
Published: December 30, 2020

ACS Publications

https://dx.doi.org/10.1021/acs.jcim.0c01332

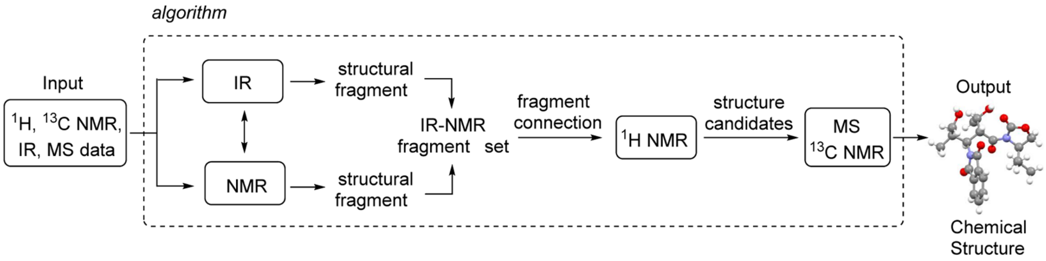


Figure 1. Flowchart of the proposed structure elucidator.

algorithm that relies on a combination of all three methods to determine the correct structure. It produces the most fitting compound given the input data from ^{1}H and ^{13}C NMR, IR, and MS spectra and elucidates the chemical structure of a molecule similarly to a trained spectroscopist, mimicking the human-driven process of structure elucidation: (i) starting by identifying functional groups by IR and proton- and carbon-containing structural fragments by ^{1}H and ^{13}C NMR spectra, (ii) connecting the identified structural fragments together by relying on the ^{1}H NMR spectrum and (iii) completing (rechecking) the elucidated molecular structure by MS and ^{13}C NMR spectrum. The flowchart of the developed algorithm is presented in Figure 1.

2. METHODS

2.1. Description of the Algorithm. The input of the algorithm consists of tabulated IR, ^{1}H NMR, and MS data; in addition, the ^{13}C NMR data can also be added, although the algorithm can perform the elucidation process without ^{13}C NMR (vide infra). The ^{1}H NMR spectrum contains proton resonances which are described by three values: the chemical shift, the integral, and the splitting pattern. The chemical shift (denoted as shift in Table 1) is the resonant frequency of a proton relative to a TMS standard and is expressed in parts per million (ppm). Roughly, it provides information about the chemical surrounding to which the proton is bound. The integral (count) gives the relative number of protons present at each resonance, while the splitting pattern (splitting) provides detailed insights into the connectivity pattern of neighboring protons in a molecule. For the first-order ^{1}H NMR spectra, the splitting pattern to n chemically equivalent neighboring protons splits the proton resonance into a $n + 1$ multiplet with intensity ratios following the Pascal's triangle. In the IR spectrum, some functional groups give rise to characteristic absorption bands described by their intensity, position (frequency, in cm^{-1}), and appearance. Although the bands can be intense or weak and broad or narrow, the algorithm specifically considers only broad absorption bands (broad in Table 1). The algorithm also makes use of the molecular mass (mass) of the compound. ^{13}C NMR data can be included if available. ^{13}C NMR allows the identification of nonequivalent carbon atoms in an organic molecule and shows a single peak for each chemically nonequivalent carbon atom. The chemical shift (denoted as shift in Table 1), analogous to ^{1}H NMR, is the resonant frequency of a carbon relative to a TMS standard or residual solvent and is expressed in parts per million (ppm).

Table 1. Definition of the Input of the Algorithm from ^{1}H NMR, IR, and MS, along with Optional ^{13}C NMR and Sources

data source	data
^{1}H NMR	shift: value from the spectrum [ppm]
	count: integral [#]
	splitting: number of peaks—1 (H neighbors)
IR	frequency: position of a peak in the IR spectrum [cm^{-1}]
	broad: value is true if the absorption band peak is broad
MS	mass of the molecule [g/mol]
^{13}C NMR (optional)	shift: value from the spectrum [ppm]

The algorithm operates with the predefined chemical shift (^{1}H NMR) and frequency (IR) ranges that are associated to the specific structural fragments—a structural fragment may correspond to only a few atoms, a functional group or an even larger structural part of the molecule. Based on these, the algorithm searches for a set of candidate compounds by (i) identifying potential matches of fragments in the input spectra, (ii) connecting the identified fragments in both spectra into joint entities, and (iii) identifying the number of fragments in the analyzed molecule and filling in the elements, which are not visible in the IR and ^{1}H NMR spectra. Finally, the algorithm joins the fragments into candidate compounds and ranks them by evaluating a number of rules, yielding a list of candidate compounds ranked by a relevance score, which gives higher rank to more likely matches. Tables with IR, ^{1}H NMR, and ^{13}C NMR frequency and chemical shift ranges for functional groups and proton- and carbon-based structural fragments,[35] along with IR, ^{1}H NMR, ^{13}C NMR, and MS data of 70 compounds from the literature were employed for algorithm development and testing of its performance.[36]

We describe more details on the individual steps of the algorithm in the following paragraphs. We illustrate them on a very simple example of methanol, for which the input IR spectrum contains two peaks at IR$_1$ = 3347 (broad) and IR$_2$ = 2945 (narrow) [cm^{-1}], ^{1}H NMR two peaks at NMR$_1$ = {3.66 ppm, H-count = 1, coupling = singlet} and NMR$_2$ = {3.43 ppm, H-count = 3, coupling = singlet}, one ^{13}C NMR peak at 50.1 ppm and has the molecular mass of 32.03.

The functioning of the developed algorithm is presented in Figure 2. The algorithm performs elucidation in five steps, which are color-coded in Figure 2, that is, step 1 (green), step 2 (red), step 3 (blue), step 4 (yellow), and step 5 (violet).

2.1.1. Step 1: Identification of IR and NMR Structural Fragments. The algorithm first deduces the structural fragments (functional groups) from their positions (frequency) in the IR spectrum. Based on the predefined frequency ranges, the algorithm assigns the input peaks to individual fragments. Because the frequency ranges for some fragments can overlap, a peak may be assigned to several structural fragments. In our example, the list of possible IR fragments related to the two peaks is CH_3, CH_2, CH, OH, and NH. In the ^{1}H NMR spectrum, the algorithm determines the proton-containing structural fragments and their neighboring protons/groups

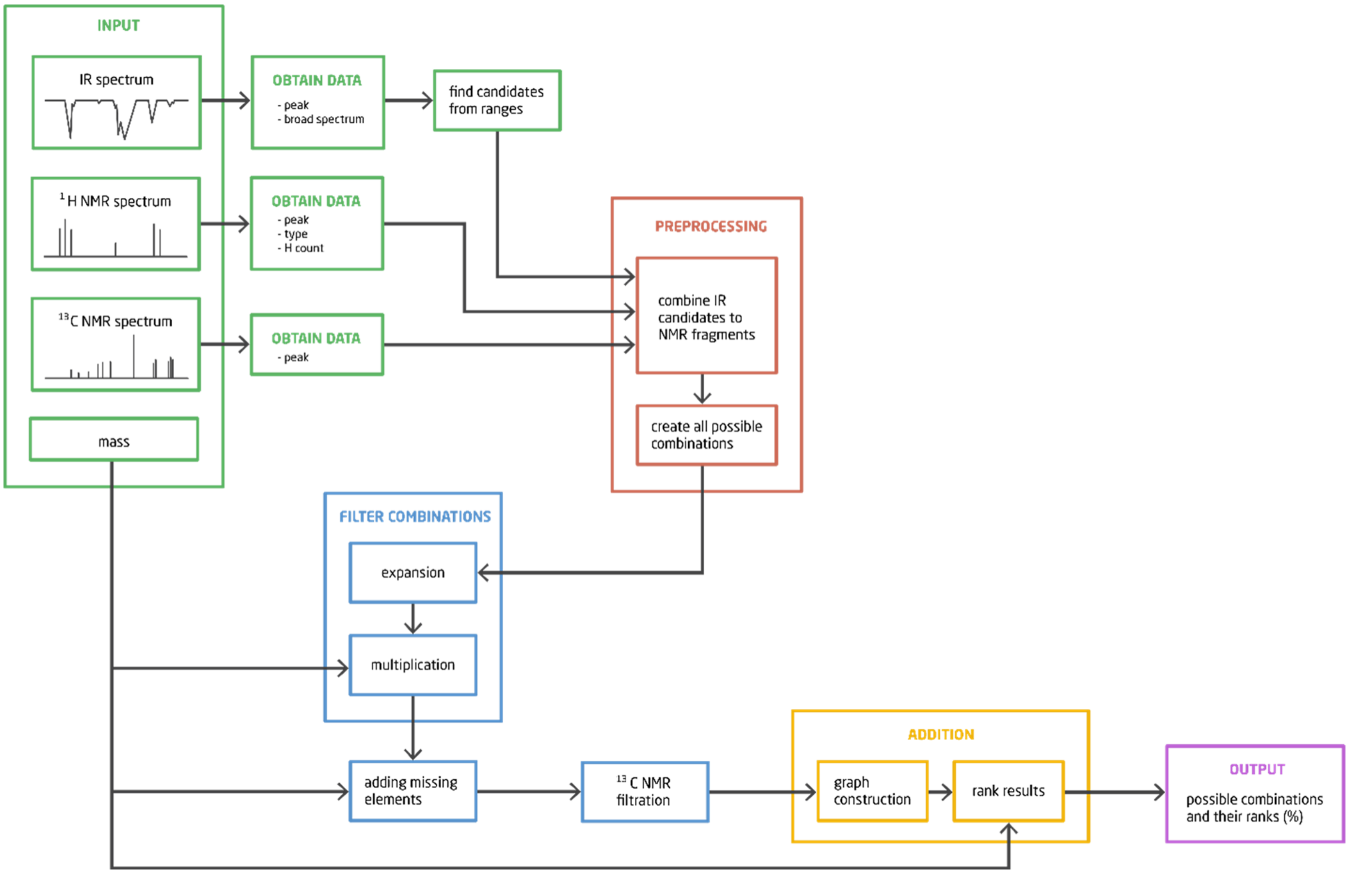


Figure 2. Schematic presentation of the elucidation algorithm. Each color denotes one of the five steps.

based on the chemical shift and splitting patterns. The result of this step is a set of separately identified IR and ^{1}H NMR structural fragments. For the ^{13}C NMR spectrum, the algorithm assigns possible identifications to the fragments, based on the reported ^{13}C NMR peaks. This process is done independently from the ^{1}H NMR and IR identification. The possible carbon atom candidates are later used in step 4 for additional validation of the candidate combinations.

2.1.2. Step 2: Grouping IR and NMR Structural Fragments. The IR spectrum only provides information on existence of structural fragments in the compound, but not their exact counts. On the other hand, the ^{1}H NMR spectrum provides information on the hydrogen count of proton-containing structural fragments. In step 2, the algorithm relates the IR and ^{1}H NMR structural fragments from step 1 into group(s) of structural fragments based on compatibility of the numbers of their hydrogen atoms. The fragments may be grouped in different ways, for example, one fragment deduced from IR can be attached to multiple NMR peaks and vice versa. Therefore, in step 2, the algorithm generates a full set of possible IR and ^{1}H NMR fragment combinations, which are denoted as IR−NMR fragments. In our example, NMR$_1$ (1 H atom) can be matched by CH, OH, and NH, while NMR$_2$ (3 H atoms) by CH$_3$, as well as CH, OH, and NH (if they appear three times). The algorithm thus maps all of these combinations into a set of matched IR−NMR fragments.

2.1.3. Step 3: Identifying Compound Candidates. In step 3, the IR−NMR fragments are augmented in different ways to match the input molecular mass, thus producing groups of structural fragments, each group representing constituent parts of a candidate compound.

2.1.3.1. Expansion. Proton resonances in the ^{1}H NMR spectrum may reflect multiple structural fragments in the molecule. For example, an NMR proton resonance with the hydrogen count of 6 may represent two CH$_3$ fragments, 3 CH$_2$ fragments, or 6 CH fragments. The algorithm thus examines all IR−NMR fragments and expands them with all possible combinations of fragments that match the hydrogen count in the input data. In our example, the algorithm expands the combinations of NMR$_2$ with CH, OH, and NH by creating three copies of each fragment in the combination to match the hydrogen count of three.

2.1.3.2. Multiplication. Because of isomorphism and the relative nature of ^{1}H NMR spectra, the count of individual IR−NMR fragments in the candidate compound may be incorrect (too low). Given the molecular mass, the algorithm checks if the counts should be augmented to match the mass and, in these cases, multiplies the number of fragments in all candidate compounds. In our simple example, no multiplication is needed.

2.1.3.3. Insertion. Some functional groups (e.g., ether) and commonly encountered halogen atoms (e.g., Cl, Br, and I) do not have specific signals in the IR and ^{1}H NMR spectra. The algorithm therefore inserts these elements into the candidate compounds to match the molecule mass.

2.1.4. Step 4: Checking for Validity. In the fourth step, each candidate compound is checked for validity. First, the candidate compound's mass is compared to the input mass, and the candidate is removed if the difference is too large. Then, the Erdős−Gallai algorithm[37] is used to test if a connected graph can

Table 2. Rules for Computation of the Relevance Score of a Candidate Compound

rule	description	penalty on the relevance score
graph contains unconnected edges (rule 1)	one or more bonds are not connected between fragments (for compounds with less than 5 unconnected edges)	−0.5% per unconnected edge
sum of difference in ppm (rule 2)	difference in NMR ppm values between connected fragments within the compound. the larger the ppm difference, the lower the relevance.	−0.5% per 1 ppm difference
difference in mass between constructed compound and the fragments' mass sum (rule 3)	the larger the difference in masses, the lower the relevance.	−0.5% per 1 g/mol difference

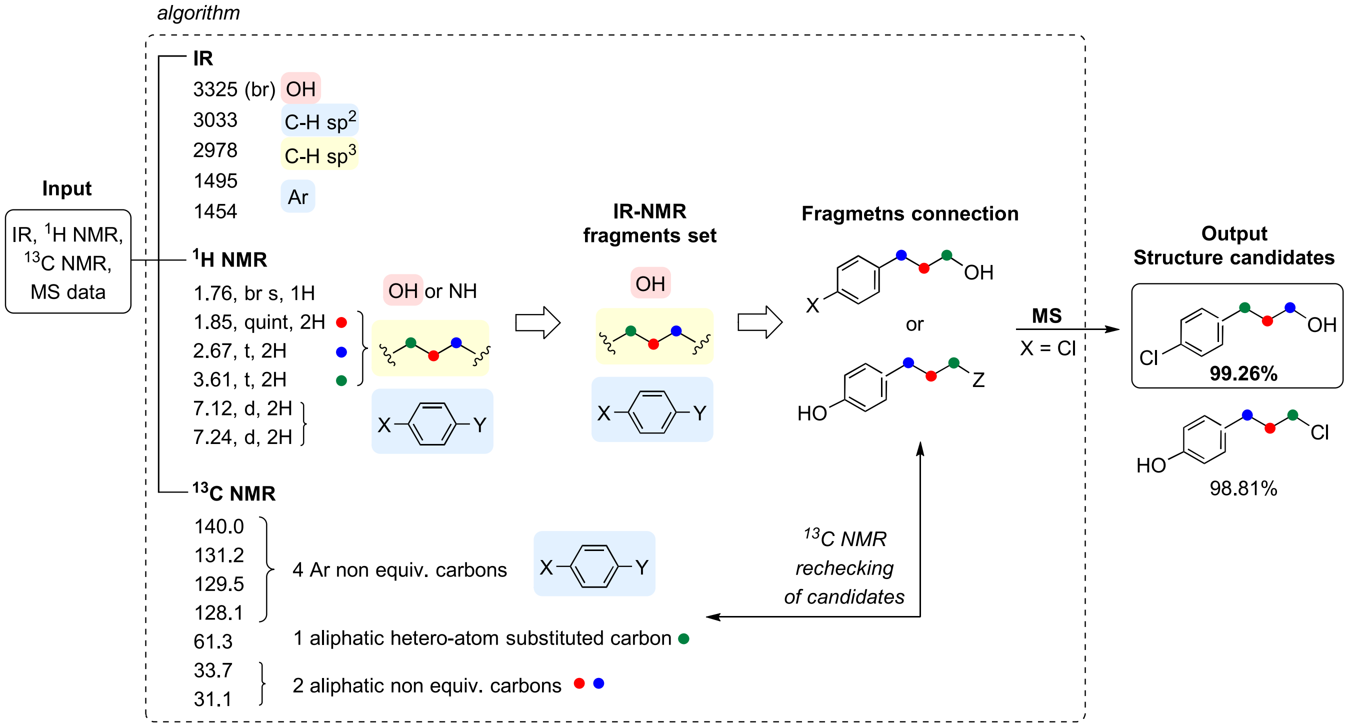


Figure 3. Simplified elucidation process with the proposed algorithm for 3-(4-chlorophenyl)propan-1-ol (1).

be constructed from the compound's fragments, given each fragment's number of connections. The compound is discarded, if there are more than five unconnected edges left after construction. The constructed compounds with less than five unconnected edges are still considered but penalized (see Table 2). In the investigated example of methanol, only a single candidate compound CH_3−OH is valid according to all three criteria.

In addition, the algorithm optionally evaluates the combinations against the ^{13}C NMR data, if given. Each carbon fragment in a candidate compound is checked, comparing the reported chemical shift in the ^{13}C NMR spectrum with predefined ^{13}C NMR regions for the investigated structural fragment. If a reported shift belongs to multiple potential fragments, all options are considered as correct. A candidate composition is therefore valid, if all identified fragments are also compliant with their ^{13}C NMR predefined range positions. In the investigated example of methanol, the CH_3 fragment value of 50.1 ppm falls into the predefined region of aliphatic carbon atoms with electronegative substituent, in this case, aliphatic alcohol. For the one remaining compound candidate, the ^{13}C NMR input data confirm the presence of the CH_3 fragment and confirms the candidate.

2.1.5. Step 5: Creating and Ranking the Compounds. Step 4 results in several candidate compounds, each consisting of a number of unconnected fragments. In step 5, the algorithm observes the fragments of a candidate compound as vertices of a graph, for which the edges (bonds between elements) must be determined. The algorithm connects the edges in compliance with each fragment's neighbor count and neighbors' sum of hydrogen atoms. In the process, some of the candidates are removed—these include compounds where all the fragments cannot be connected because of limitations in connections between the elements; others lack the elements, which were not visible in any of the input spectra and were not added in step 3. Each candidate compound can produce any number of valid graphs (multiple possible edge combinations); therefore, the algorithm evaluates each one using a set of rules, yielding a relevance score for each candidate compound. The rules are shown in Table 2. Initially, the algorithm assigns a relevance score of 100% to each candidate compound. Based on the rules, the algorithm lowers the score of the candidates accordingly.

The output of the algorithm is a list of compound candidates with the corresponding relevance scores. If a compound can be connected, does not contain unconnected edges, and the fragment set produces a connected graph with a mass, similar to the measured mass of the compound, it will receive a high relevance score. If it fails on one or more rules, the relevance score is lowered accordingly.

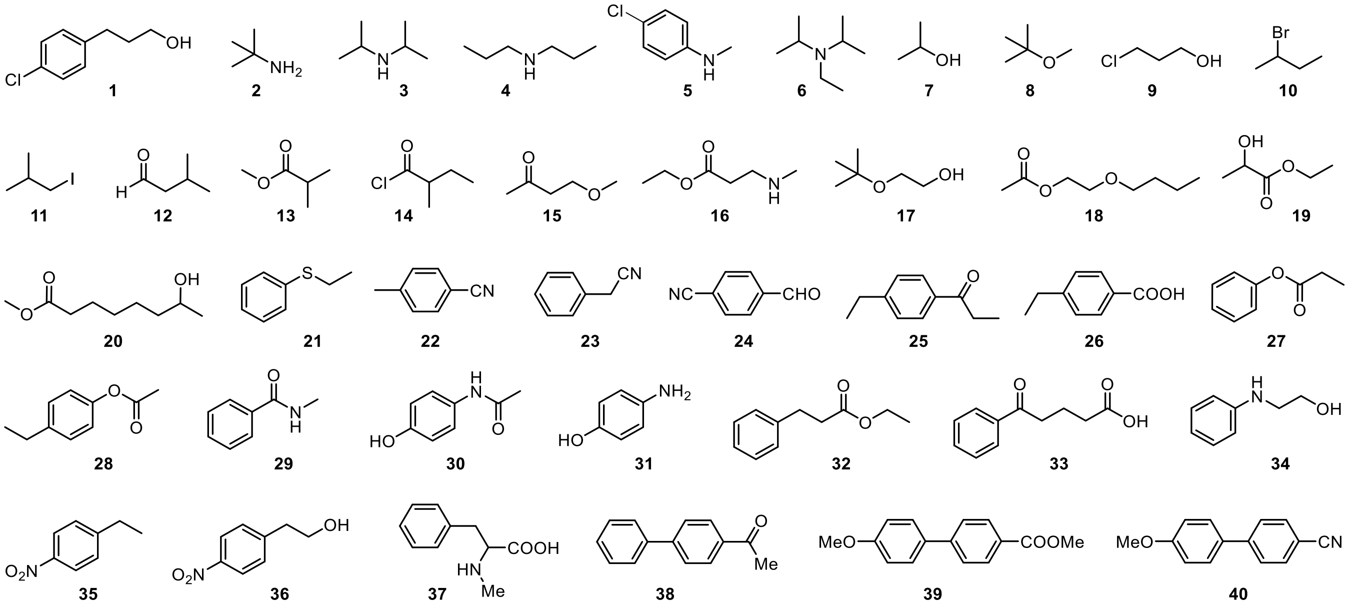


Figure 4. Selected examples of resolved structures with IR, ^{1}H NMR, ^{13}C NMR, and MS data from the literature.

2.2. Computational Methods.

The presented elucidation approach was developed in Python 3,[38] with the web service for online elucidation using the Django[39] framework with JavaScript support on the front-end. The web service for online elucidation also includes the Pysmiles library,[40] which translates the elucidated compounds into SMILES representations. The Pysmiles library also employs the NetworkX library,[41] which enables the manipulation and creation of graphs. Using the online service WolframAlpha,[42] the web service visualizes the elucidated end result compounds, which the algorithm provides as the output.

The algorithm employs a combinatorial approach to combine the identified fragments into compounds. Several time-optimizing approaches were also used to minimize the number of possible combinations. Using the Erdős–Gallai theorem,[37] we removed the possible compound combinations. This theorem provides a necessary and sufficient condition for a finite sequence of chemical bonds to establish whether the combination is potentially possible. Using backtracking,[43] we recursively built graphs from the compound combinations. Finally, we developed a simple decision model,[44] which ranks the built compounds given a set of rules.

3. RESULTS AND DISCUSSION

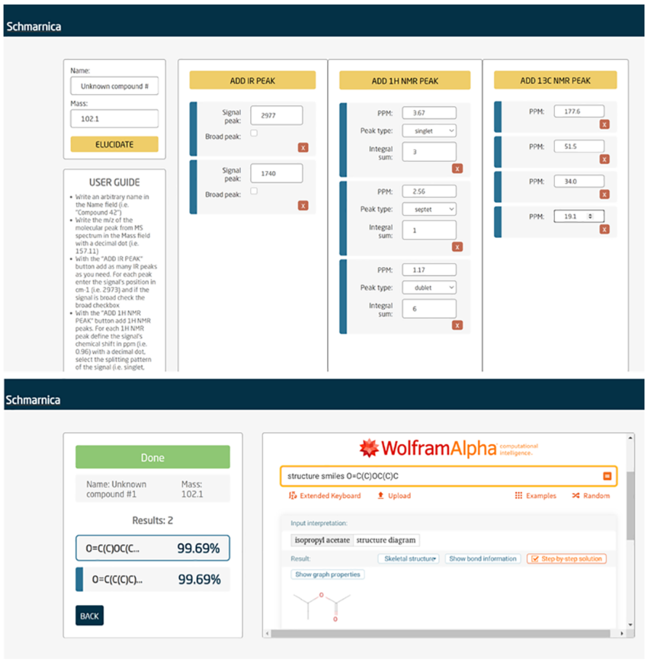
In the following, we describe a simplified elucidation process with the proposed algorithm on 3-(4-chlorophenyl)propan-1-ol (**1**, Figure 3). The input tabular data for IR, ^{1}H NMR, ^{13}C NMR, and MS spectra were adopted from the literature: IR: 3325 (broad), 1495, 1454, 1060, 1029, 968, 754 cm^{-1};[45] ^{1}H NMR: 1.76 (br s, 1 H), 1.78–1.92 (m, 2H), 2.67 (t, 2H), 3.61 (t, 2H), 7.12 (d, 2H), 7.24 (d, 2H);[45] ^{13}C NMR: 140.0, 131.2, 129.5, 128.1, 61.3, 33.7, 31.1;[45] and MS: 170 (the nominal mass that corresponds to the molecular ion peak M$^+$ of [$C_9H_{11}{}^{35}ClO$]$^+$).[46] The peak values of 3033 and 2978 cm^{-1} for C(sp^2)-H and C(sp^3)-H bond stretching were added to the IR spectra because the algorithm fails to provide elucidated structure if the present structural fragments of the compound are not assigned (or missing) in the IR spectrum (vide infra). Additionally, multiplet proton resonance at δ 1.78–1.92 ppm was assigned as 1.85 (quint-like, 2H) (vide infra).

The algorithm started the elucidation process (step 1) by analyzing the IR spectrum. A broad peak at 3325 cm^{-1} was assigned to an OH group and peaks at 3033 and 2978 cm^{-1} to the structural fragments with C(sp^2)-H and C(sp^3)-H bonds, respectively. The 1495 and 1454 cm^{-1} peaks implied the presence of an aromatic ring (C═C stretching). Analysis of the ^{1}H NMR spectrum revealed the presence of one proton bound to a heteroatom because of the broad singlet resonance at 1.76 ppm with the integral value of 1, which may correspond to an OH or NH moiety. A quintet-like resonance with integral of 2 indicated a CH_2 fragment having two pairs of chemically equivalent neighboring protons (see below). The two resonances with the shifts in the aliphatic region of the NMR spectrum, that is, at 2.67 and 3.61 ppm, with integral 2 and a triplet splitting pattern (2 neighbors) suggested the $-CH_2-CH_2-CH_2-$ hydrocarbon chain. The two doublets with integrals 2 in the aromatic region of the NMR spectrum indicated a para-substituted aromatic ring. Analysis of the ^{13}C NMR spectrum revealed the presence of four nonequivalent aromatic carbon atoms resonating at 140.0, 131.2, 129.5, and 128.1 ppm, two nonequivalent aliphatic carbon atoms connected in a hydrocarbon chain resonating at 33.7 and 31.1 ppm, and one heteroatom-substituted aliphatic carbon resonating at 61.3 ppm. In step 2, these elements were combined, into IR–NMR fragments: an OH group, a $-CH_2-CH_2-CH_2-$ hydrocarbon chain, and a para-substituted aromatic ring. In step 3, the difference of the sum of calculated mass of fragments on the IR–NMR fragments list and the experimental input mass value differed by 35 which corresponds to the mass number of chlorine-35, which was added to the candidate compound. The process was completed (steps 4 and 5) by constructing the compound graph and attaching the chlorine atom at the aromatic ring para relative to the propyl alcohol substituent. In this way, the algorithm derived to the correct structure. The phenol structure candidate, also possible with respect to the IR–NMR fragments set list, received a lower relevance score because of the ^{1}H NMR chemical shift δ 1.76 ppm of the OH group. In step 4, each carbon fragment in the candidate compounds was checked with respect to the predefined ranges and reported values of ^{13}C NMR data (vide supra). All carbon atoms in both structure candidates in Figure 3 were possible with respect to reported ^{13}C NMR input values.

Selected examples of structures resolved from their corresponding literature IR, ^{1}H NMR, ^{13}C NMR, and MS tabulated data by the proposed algorithm are presented in Figure 4. The algorithm resolves the structures of various primary, secondary, tertiary, and aromatic amines **2**−**6**, **16**, **31**, **34**, **37**, alcohols **7**, **9**, **17**, **19**, **20**, **30**, **31**, **34**, **36**, ethers **8**, **15**, **17**, **18**, halogenated compounds **5**, **10**, **11**, aldehydes **12**, **24**, esters **13**, **16**, **18**−**20**, **27**, **28**, **32**, **39**, acetyl chloride **14**, ketones **15**, **25**, **33**, **38**, sulfide **21**, nitriles **22**−**24**, **40**, carboxylic acids **26**, **33**, **37**, amides **29**, **30**, nitro compounds **35**, **36**, and biphenyl derivatives **38**−**40**. The literature spectroscopic data of compounds in Figure 4 is collected in the Supporting Information.[36] The algorithm recognizes various functional groups and differentiates between structural isomers of organic molecules.

When elucidating structures of organic molecules, we solve a problem that is related to the class of inverse problems, which are most frequently ill-posed and usually do not have a unique solution. Therefore, the correct structure of the investigated molecule was not always the first on the list of the compound candidates with the highest relevance score. For example, for compound **33**, the highest-ranking candidate was 2-oxo-5-phenylpentanoic acid (score 99.41%), whereas the correct structure of the investigated compound, 5-oxo-5-phenylpentanoic acid (**33**), was ranked second (score 99.05%). Close inspection of the ^{1}H NMR spectrum [δ 11.10 (br s, 1H), 7.50 (m, 5H), 3.07 (t, 2H), 2.50 (t, 2H), 2.10 (quint-like, 2H)][47] revealed that terminal methylene groups of the $-CH_2-CH_2-CH_2-$ chain with resonances at δ 3.07 (t, 2H) and 2.50 (t, 2H) ppm, have chemical shifts that could imply an attachment on an aromatic ring, carbonyl, as well as a carboxylic group. Therefore, both of the above described compound candidates are probable. However, in investigated cases the correct structure of the molecule was usually first or second, in few cases third, result on the list of compound candidates and always had a relevance score above 99%.

The elucidator has a user-friendly web interface and is publicly available (http://schmarnica.si). Users can input numerical values of IR, ^{1}H NMR, and MS spectral data, along with optional input of ^{13}C NMR data (Figure 5), of the investigated compound into designated fields and run the elucidation process. The elucidation process can run with or without ^{13}C NMR data, although ^{13}C NMR data improve the process and result of elucidation. The process incorporates the described combinatorial search for potential fragments based on the two spectra, and a graph search algorithm, which evaluates the connectivity of the potential fragments from each possible combination. After the process is finished, the interface lists the compound candidates, which are visualized by the WolframAlpha computational knowledge engine within the interface. The candidates are ranked according to the predefined rules for analyzing the spectral data (vide supra), and the ranking score is displayed next to each candidate. If the algorithm successfully resolves the structure from the spectral data (vide infra), the molecular candidate with the highest-percent matching (for the tested examples) almost always corresponded with the correct structure (vide supra). The processing time, in which elucidator derives the list of structure candidates, depends of the number of structural fragments present in the investigated molecule and their connectivity. On the test data, the algorithm usually derived a list of structure candidates in a few seconds to up to 5 min. In cases of compounds with several structural fragments (e.g., compound **20**) or elements that can be identified only by MS (e.g., Cl, Br, I atoms, and ethers), the time of calculation can significantly increase because of the exponentially increased number of possible combinations.



Name: Unknown compound #

Mass: 102.1

USER GUIDE

- Write an arbitrary name in the Name field (i.e. "Compound 42")
- Write the m/z of the molecular peak from MS spectrum in the Mass field with a decimal dot (i.e. 157.11)
- With the "ADD IR PEAK" button add as many IR peaks as you need. For each peak enter the signal's position in cm-1 (i.e. 2973) and if the signal is broad check the broad checkbox
- With the "ADD 1H NMR PEAK" button add 1H NMR peaks. For each 1H NMR peak define the signal's chemical shift in ppm (i.e. 0.96) with a decimal dot, select the splitting pattern of the signal (i.e. singlet,

Signal peak: 2977 Broad peak: □

Signal peak: 1740 Broad peak: □

PPM: 3.67 Peak type: singlet Integral sum: 3

PPM: 2.56 Peak type: septet Integral sum: 1

PPM: 1.17 Peak type: dublet Integral sum: 6

PPM: 177.6

PPM: 51.5

PPM: 34.0

PPM: 19.1

Name: Unknown compound #1 Mass: 102.1

Results: 2

O=C(C)OC(C...	99.69%
O=C(C(C)C)...	99.69%

structure smiles O=C(C)OC(C)C

Input interpretation: isopropyl acetate structure diagram

Result:

Figure 5. Snapshots from the Schmarnica user interface (http://schmarnica.si/, accessed Nov 13, 2020): IR, ^{1}H NMR, and MS data input, and optional input of ^{13}C NMR data (above), along with presentation of the elucidated structure by WolframAlpha (below) for the case of isopropyl acetate. Computational complexity and developed optimizations.

An evaluation of different optimization and data techniques was used to determine the efficiency of the proposed model. First, we evaluated the time complexity of the proposed approach from its baseline (no optimizations) to the final version. In order to evaluate the impact of the individual optimization technique on the process, we only evaluated the complexity and not the classification performance of the approach. The aggregated results are shown in Table 3.

The initial time needed to elucidate 40 compounds was 341.04 s. With weight checking and removal of potentially incorrect combinations, this time was significantly reduced to 21.66 s. Further optimizations, which additionally excluded the impossible compound combinations, reduced the calculation time by another 50% to 11.45 s. Inclusion of the ^{13}C NMR data did not significantly affect the time complexity. However, the number of elucidation candidates was reduced by about 20%. This part of the process decreased the final number of candidates by removing the incorrect candidates from the results, while not negatively affecting the classification performance by potentially removing the correct results from the candidate list.

Considering all optimizations, the average number of candidates per compound significantly decreased by considering the weight of the combinations, followed by the connectivity methods and tree realization checks. Additionally, the ^{13}C NMR data, which we added as an optional input to the proposed approach, reduced the number of candidates.

It is important to note that in its current form the algorithm can resolve relatively simple organic molecules (Figure 4). It currently processes only first-order ^{1}H NMR spectra and fails to

Table 3. Analysis of Impact of the Cumulative Optimization Techniques on the Algorithm's Performance

optimization type	sum of time spent for testing[a,b]	average time for elucidation[b]	average number of candidates per test[c]
none (baseline)	341.04	8.53	45.2
weight of compound candidate	21.66	0.54	3.3
Erdős−Gallai connectivity check	12.88	0.32	2.1
tree realization check	11.45	0.29	1.9
^{13}C NMR−additional data	11.38	0.28	1.6

[a]Reported test results for 40 tested compounds. [b]Reported in seconds. [c]Test refers to an average number of candidates on a single compound in a set of 40 tested compounds.

resolve the unknown structure if the IR spectrum does not provide the information on the functional groups that are present in the molecule and should be seen in the IR spectrum. For instance, if the −NH− group is present in the molecule, but the IR spectrum for some reason (hidden/superimposed/not assigned signal) does not provide the corresponding peak value for this group, that is, ca. 3300 cm^{-1}, the algorithm fails to resolve the structure. In some cases, the spectral data of ^{1}H NMR were adjusted to fit the first-order NMR data. For example, in the above presented elucidation of compound **1**, the splitting pattern of proton resonance at δ 1.78−1.92 ppm was assigned as a quintet-like (quint-like) at δ 1.85 ppm (vide supra). We decided to use this term as it corresponds well to what one can actually observe in the spectrum without knowing the structure of the compound. Depending on the resolution of the spectrum, the quintet-like splitting pattern commonly appears for the central methylene protons in $X-CH_2-CH_2-CH_2-Y$ hydrocarbon chain that are coupled to the nonequivalent neighboring pairs of $X-CH_2$ and CH_2-Y protons with similar coupling constants. In the literature, the resonance for this type of central methylene protons is correctly reported as a multiplet or triplet of triplets; however, at the current stage, the algorithm cannot process more complex splitting patterns (e.g., dd, dt, tt, etc.) or multiplets. The exception is the phenyl group, C_6H_5- (Ph−) that is defined as a multiplet resonance with integral 5 in the region around 7 ppm. The algorithm can also process para-substituted phenyl ring as it frequently resembles two doublet resonances with an integral ratio of 2:2 in the aromatic region of the spectra. Therefore, only compounds with mono- and para-substituted phenyl rings can be currently processed by the algorithm. One of the primary goals in further developments will be upgrading the algorithm to resolve more complex NMR data as well as to resolve structures from partly incomplete spectral data (vide supra). The algorithm proposed herein serves as a ground for further developments, which will increase its capacity of resolving more complex molecular structures.

4. CONCLUSIONS

The proposed algorithm is the first step in the development of a user-friendly and database-independent chemical structure elucidator that would mimic a spectroscopist-driven process of resolving the molecular structure from spectral data, that is, building the molecular structure form small predefined fragments. It recognizes various functional groups and differentiates between structural isomers of organic molecules. In its current form, the algorithm can resolve rather simple organic structures and will thus serve as a basis for further developments. The elucidator is publicly available through a web-interface, which can be used to elucidate and visualize unknown compounds. For interested researchers, source code of the elucidator is also publicly available.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jcim.0c01332.

Literature spectroscopic data of compounds employed for algorithm development and testing of its performance (PDF)

■ AUTHOR INFORMATION

Corresponding Authors

Matevž Pesek − *Faculty of Computer and Information Science, University of Ljubljana, SI-1000 Ljubljana, Slovenia*; Email: matevz.pesek@fri.uni-lj.si

Martin Gazvoda − *Faculty of Chemistry and Chemical Technology, University of Ljubljana, SI-1000 Ljubljana, Slovenia*; orcid.org/0000-0003-3421-0682; Email: martin.gazvoda@fkkt.uni-lj.si

Authors

Andraž Juvan − *Faculty of Computer and Information Science, University of Ljubljana, SI-1000 Ljubljana, Slovenia*

Jure Jakoš − *Faculty of Chemistry and Chemical Technology, University of Ljubljana, SI-1000 Ljubljana, Slovenia*

Janez Košmrlj − *Faculty of Chemistry and Chemical Technology, University of Ljubljana, SI-1000 Ljubljana, Slovenia*; orcid.org/0000-0002-3533-0419

Matija Marolt − *Faculty of Computer and Information Science, University of Ljubljana, SI-1000 Ljubljana, Slovenia*

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jcim.0c01332

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

The authors acknowledge the financial support from the Slovenian Research Agency (Research Core Funding grant P1-0230 and projects J1-8147, J1-9166, and L1-9191).

■ REFERENCES

(1) Lindsay, R. K.; Buchanan, B. G.; Feigenbaum, E. A.; Lederberg, J. DENDRAL: a Case Study of the First Expert System for Scientific Hypothesis Formation. *Artif. Intell.* **1993**, *61*, 209−261.

(2) *Contemporary Computer-Assisted Approaches to Molecular Structure Elucidation*; Williams, A. J., Blinov, K., Elyashberg, M. E., Eds.; The Royal Society of Chemistry: Cambridge, U.K., 2012.

(3) *Modern NMR Approaches to the Structure Elucidation of Natural Products*; Williams, A. J., Blinov, K., Rovnyak, D., Eds.; The Royal Society of Chemistry: Cambridge, U.K., 2016.

(4) Lederberg, J.; Sutherland, G. L.; Buchanan, B. G.; Feigenbaum, E. A.; Robertson, A. V.; Duffield, A. M.; Djerassi, C. Applications of Artificial Intelligence for Chemical Inference I. The Number of Possible

Organic Compounds: Acyclic Structures Containing Carbon, Hydrogen, Oxygen, and Nitrogen. *J. Am. Chem. Soc.* **1969**, *91*, 2973−2976.

(5) Funatsu, K.; Sasaki, S.-i. Recent Advances in the Automated Structure Elucidation System, CHEMICS. Utilization of Two-Dimensional NMR Spectral Information and Development of Peripheral Functions for Examination of Candidates. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 190−204.

(6) Bremser, W. Structure Elucidation and Artificial Intelligence. *Angew. Chem., Int. Ed. Engl.* **1988**, *27*, 247−260.

(7) Will, M.; Fachinger, W.; Richert, J. R. Fully Automated Structure Elucidation - A Spectroscopist's Dream Comes True†. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 221−227.

(8) Hong, H.; Xin, X. ESSESA: An Expert System for Structure Elucidation from Spectra. 5. Substructure Constraints from Analysis of First-Order - Spectra. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 1259−1266.

(9) Huixiao, H.; Yinling, H.; Xinquan, X.; Yufeng, S. ESSESA: An Expert System for Structure Elucidation from Spectra. 6. Substructure Constraints from Analysis of l3C-NMR Spectra. *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 979−1000.

(10) *ACD Labs*, https://www.acdlabs.com/ (accessed Nov 13, 2020).

(11) *Bio-Rad Laboratories Incorporated*, http://www.bio-rad.com (accessed Nov 13, 2020).

(12) Masui, H.; Hong, H. Spec2D: A Structure Elucidation System Based on 1H NMR and H−H COSY Spectra in Organic Chemistry. *J. Chem. Inf. Model.* **2006**, *46*, 775−787.

(13) Elyashberg, M. E.; Karasev, Y. Z.; Martirosian, E. R.; Thiele, H.; Somberg, H. Expert Systems as a Tool for the Molecular Structure Elucidation by Spectral Methods. Strategies of Solution to the Problems. *Anal. Chim. Acta* **1997**, *348*, 443−463.

(14) Elyashberg, M. E.; Blinov, K. A.; Martirosian, E. R. A New Approach to Computer-Aided Molecular Structure Elucidation: The Expert System Structure Elucidator. *Lab. Autom. Inf. Manag.* **1999**, *34*, 15−30.

(15) Blinov, K. A.; Elyashberg, M. E.; Molodtsov, S. G.; Williams, A. J.; Martirosian, E. R. An Expert System for Automated Structure Elucidation Utilizing ^{1}H−^{1}H, ^{13}C−^{1}H and ^{15}N−^{1}H 2D NMR Correlations. *Fresenius' J. Anal. Chem.* **2001**, *369*, 709−714.

(16) Peironcely, J. E.; Rojas-Chertó, M.; Fichera, D.; Reijmers, T.; Coulier, L.; Faulon, J.-L.; Hankemeier, T. OMG: Open Molecule Generator. *J. Cheminf.* **2012**, *4*, 21.

(17) *MOLGEN Molecular Structure Generation*, http://www.molgen.de (accessed Nov 13, 2020).

(18) Steinbeck, C. Recent Developments in Automated Structure Elucidation of Natural Products. *Nat. Prod. Rep.* **2004**, *21*, 512−518.

(19) Dunkel, R.; Wu, X. Identification of Organic Molecules from a Structure Database Using Proton and Carbon NMR Analysis Results. *J. Magn. Reson.* **2007**, *188*, 97−110.

(20) Koichi, S.; Arisaka, M.; Koshino, H.; Aoki, A.; Iwata, S.; Uno, T.; Satoh, H. Chemical Structure Elucidation from ^{13}C NMR Chemical Shifts: Efficient Data Processing Using Bipartite Matching and Maximal Clique Algorithms. *J. Chem. Inf. Model.* **2014**, *54*, 1027−1035.

(21) Elyashberg, M. Identification and Structure Elucidation by NMR Spectroscopy. *Trends Anal. Chem.* **2015**, *69*, 88−97.

(22) Koichi, S.; Koshino, H.; Satoh, H. Handling of Highly Symmetric Molecules for Chemical Structure Elucidation in a CAST/CNMR System. *J. Comput. Chem. Jpn.* **2016**, *14*, 193−195.

(23) Dias, D.; Jones, O.; Beale, D.; Boughton, B.; Benheim, D.; Kouremenos, K.; Wolfender, J.-L.; Wishart, D. Current and Future Perspectives on the Structural Identification of Small Molecules in Biological Systems. *Metabolites* **2016**, *6*, 46.

(24) Mohamed, A.; Nguyen, C. H.; Mamitsuka, H. Current Status and Prospects of Computational Resources for Natural Product Dereplication: a Review. *Briefings Bioinf.* **2016**, *17*, 309−321.

(25) Jindalertudomdee, J.; Hayashida, M.; Zhao, Y.; Akutsu, T. Enumeration Method for Tree-like Chemical Compounds with Benzene Rings and Naphthalene Rings by Breadth-first Search Order. *BMC Bioinf.* **2016**, *17*, 113.

(26) Li, D.-W.; Wang, C.; Brüschweiler, R. Maximal Clique Method for the Automated Analysis of NMR TOCSY Spectra of Complex Mixtures. *J. Biomol. NMR* **2017**, *68*, 195−202.

(27) Buevich, A. V.; Elyashberg, M. E. Enhancing Computer Assisted Structure Elucidation with DFT Analysis of J-Couplings. *Magn. Reson. Chem.* **2020**, *58*, 594−606.

(28) Pereira, F.; Aires-de-Sousa, J. Computational Methodologies in the Exploration of Marine Natural Product Leads. Mar. *Drugs* **2018**, *16*, 236.

(29) Burns, D. C.; Mazzola, E. P.; Reynolds, W. F. The Role of Computer-Assisted Structure Elucidation (CASE) Programs in the Structure Elucidation of Complex Natural Products. *Nat. Prod. Rep.* **2019**, *36*, 919−933.

(30) Leelananda, S. P.; Lindert, S. Using NMR Chemical Shifts and Cryo-EM Density Restraints in Iterative Rosetta-MD Protein Structure Refinement. *J. Chem. Inf. Model.* **2020**, *60*, 2522−2532.

(31) Valli, M.; Russo, H. M.; Pilon, A. C.; Pinto, M. E. F.; Dias, N. B.; Freire, R. T.; Castro-Gamboa, I.; Bolzani, V. S. Computational methods for NMR and MS for Structure Elucidation I: Software for Basic NMR. *Phys. Sci. Rev.* **2019**, *4*, 20180108.

(32) Valli, M.; Russo, H. M.; Pilon, A. C.; Pinto, M. E. F.; Dias, N. B.; Freire, R. T.; Castro-Gamboa, I.; Bolzani, V. S. Computational Methods for NMR and MS for Structure Elucidation II: Database Resources and Advanced Methods. *Phys. Sci. Rev.* **2019**, *4*, 20180167.

(33) Bitchagno, G. T. M.; Tanemossu, S. A. F. Computational Methods for NMR and MS for Structure Elucidation III: More Advanced Approaches. *Phys. Sci. Rev.* **2019**, *4*, 20180109.

(34) Howarth, A.; Ermanis, K.; Goodman, J. M. DP4-AI Automated NMR Data Analysis: Straight from Spectrometer to Structure. *Chem. Sci.* **2020**, *11*, 4351−4359.

(35) For table of ^{1}H NMR and ^{13}C NMR chemical shift ranges see: *A Complete Introduction to Modern NMR Spectroscopy*; Macomber, R. S., Ed.; John Wiley and Sons: New York, 1998. For table of IR frequency ranges see: *Introduction to Spectroscopy*, 4th ed.; Pavia, D. L., Lampman, G. M., Kriz, G. S., Vyvyan, J. R., Eds.; Brooks/Cole: Belmont, USA, 2009.

(36) See Supporting Information.

(37) Erdős, P.; Gallai, T. Gráfok Elő́irt Fokszámú Pontokkal. *Mat. Lapok* **1960**, *11*, 264−274.

(38) *Python*, https://www.python.org/download/releases/3.8/ (accessed Nov 13, 2020).

(39) *Django*, https://www.djangoproject.com/ (accessed Nov 13, 2020).

(40) *Pysmiles*, version 1.0.1, https://pypi.org/project/pysmiles/ (accessed Nov 13, 2020).

(41) *NetworkX*, version 1.0.1, https://networkx.github.io/ (accessed Nov 13, 2020).

(42) *WolframAlpha*, version 1.0.1, https://www.wolframalpha.com/ accessed Nov 13, 2020).

(43) Lehmer, D. H. Combinatorial Problems with Digital Computers. *Proceedings of the Fourth Canadian Mathematical Congress*, 1957; pp 160−173.

(44) *Intelligent Decision Systems*; Holtzman, S., Ed.; Addison-Wesley: Reading, Massachusetts, USA, 1987.

(45) Sudalai, A.; Jagdale, A. R.; Paraskar, A. S. Cobalt(II) Chloride Hexahydrate-Diisopropylamine Catalyzed Mild and Chemoselective Reduction of Carboxylic Esters with Sodium Borohydride. *Synthesis* **2009**, *2009*, 660−664.

(46) Cárdenas, J.; Morales-Serna, J. A.; García-Ríos, E.; Bernal, J.; Paleo, E.; Gaviño, R. Reduction of Carboxylic Acids Using Esters of Benzotriazole as High-Reactivity Intermediates. *Synthesis* **2011**, *2011*, 1375−1382.

(47) Spectral Database for Organic Compounds. https://sdbs.db.aist.go.jp/sdbs/cgi-bin/cre_index.cgi (accessed Nov 13, 2020).