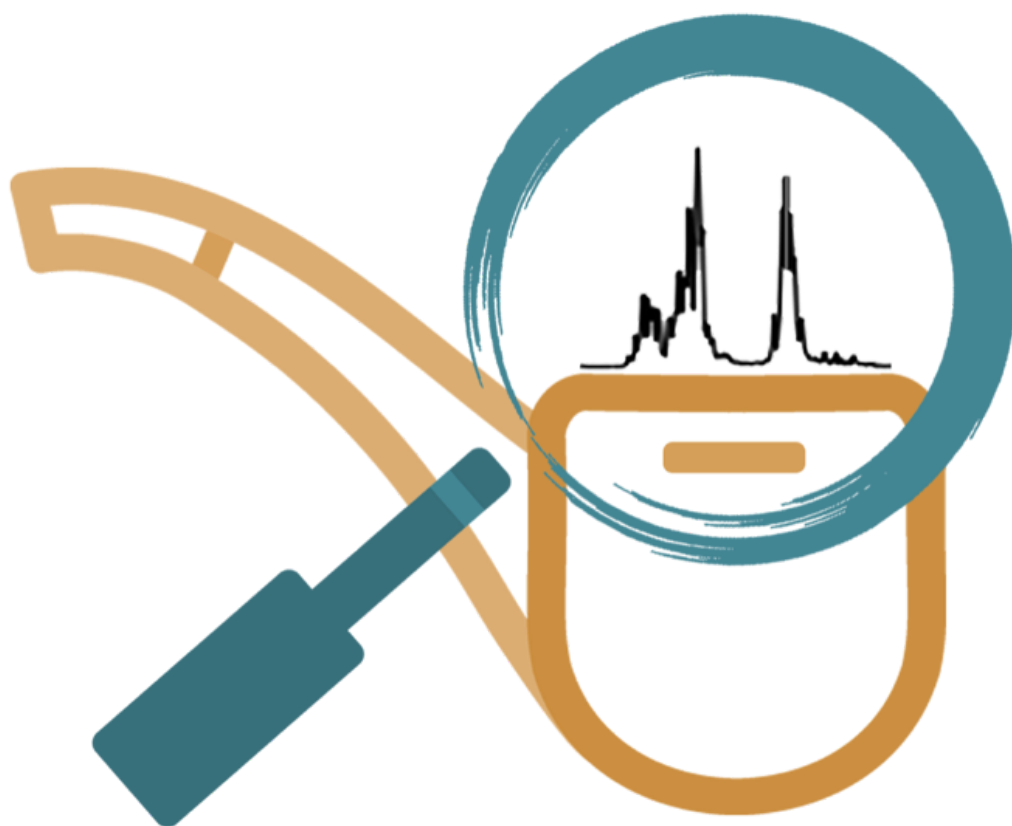


# **A Manual for structure elucidation using Sherlock**



**Written by** Michaela Fricke

The computer-assisted structure elucidation system (CASE-System) Sherlock which was written by Michael Wenk supports the structural analysis of compounds based on NMR data.

To process the NMR spectra and picking the signals the software NMRium of Zakodium Sàrl (<https://www.nmrium.org/>) is included into Sherlock and has been further developed by Michael Wenk to meet the requirements for the preparation of the spectra for the CASE-System. The computation of the possible molecular structures is performed by the software PyLSD. Sherlock can not only generate structures of molecules based on the picked signals and the settings, the so-called elucidation, but also carries out a spectra-spectra comparison with spectra from already known compounds, a process called dereplication. Based on the database NMRShiftDB and the generated NMR spectra of the compounds from the database COCONUT the dereplication will be performed.

The input data to use Sherlock for structure elucidation are 1D and 2D NMR spectra and the molecular formula. The knowledge in evaluation of NMR spectra is necessary as well.

## 1. Installation for Unix-based Systems

Before installing Sherlock, the software docker should be present on the computer because Sherlock uses Docker containers.

The program is divided into the two parts: the backend and the graphical interface, the frontend. The instructions for the installation can be found here:

<https://github.com/michaelwenk/sherlock> (backend)

<https://github.com/michaelwenk/sherlock-frontend> (frontend)

## 2. Use of Sherlock

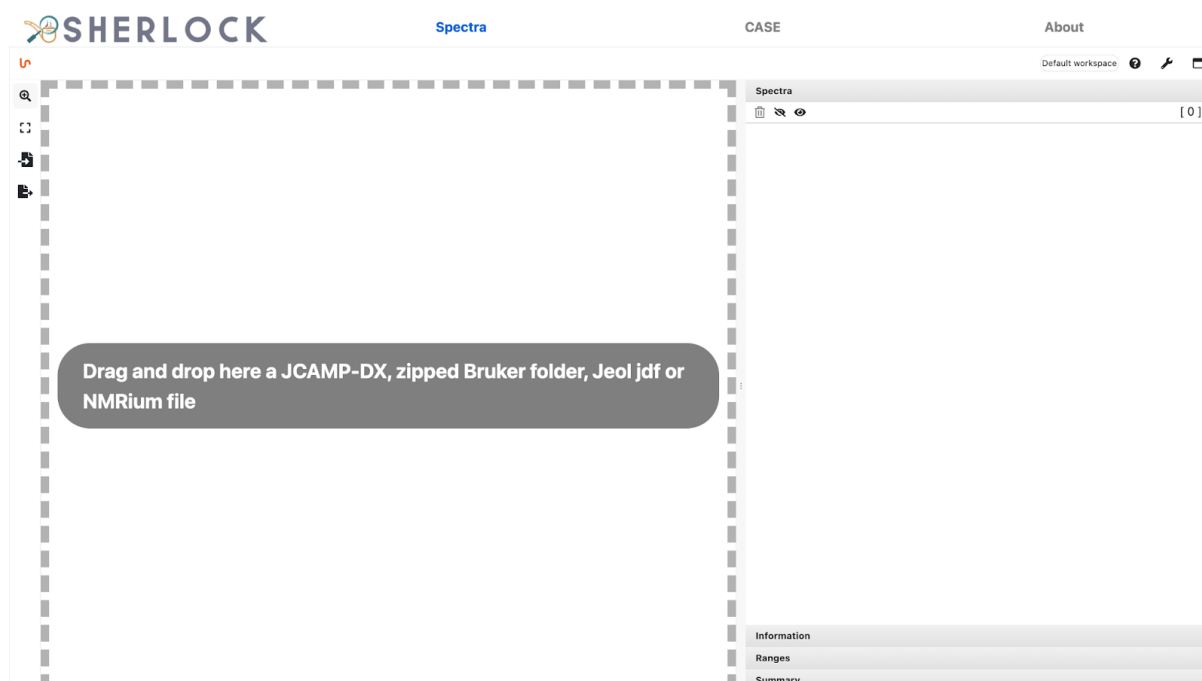


Figure 1 Overview of the graphical interface from Sherlock

Sherlock is a free and open web service. **Figure 1** shows an overview of the graphical interface. On the top you can see the three tabs: *Spectra*, *CASE* and *About*.

On the left side of the tab *Spectra* is the field for the NMR spectra view. On the left edge are the tools for the processing of the spectra. On the right side the user finds different panels which can be de- or reactivated if needed. Substantial panels for structure determination are the panel *Spectra* to select the a 1D or 2D spectrum to process, the two panels showing picked signals within ranges (1D) or zones (2D) and the panel *Summary* which shows an overview of correlations across all spectra and offers several settings for signals used in structure identification/elucidation. Further helpful panels are the panel *Information* containing a list of spectral and measurement parameters as well as the panel *Database* with information about common solvent signals and their characteristics (**Fig 2**).

Above the panels the user has some options to adjust the graphical interface according to individual requirements. Through the left button the user can choose between different settings depending on the task which will be processed. Under the 'wrench' button the workspace can be set more individually. The right button is to enter the full screen mode (**Fig 1**).

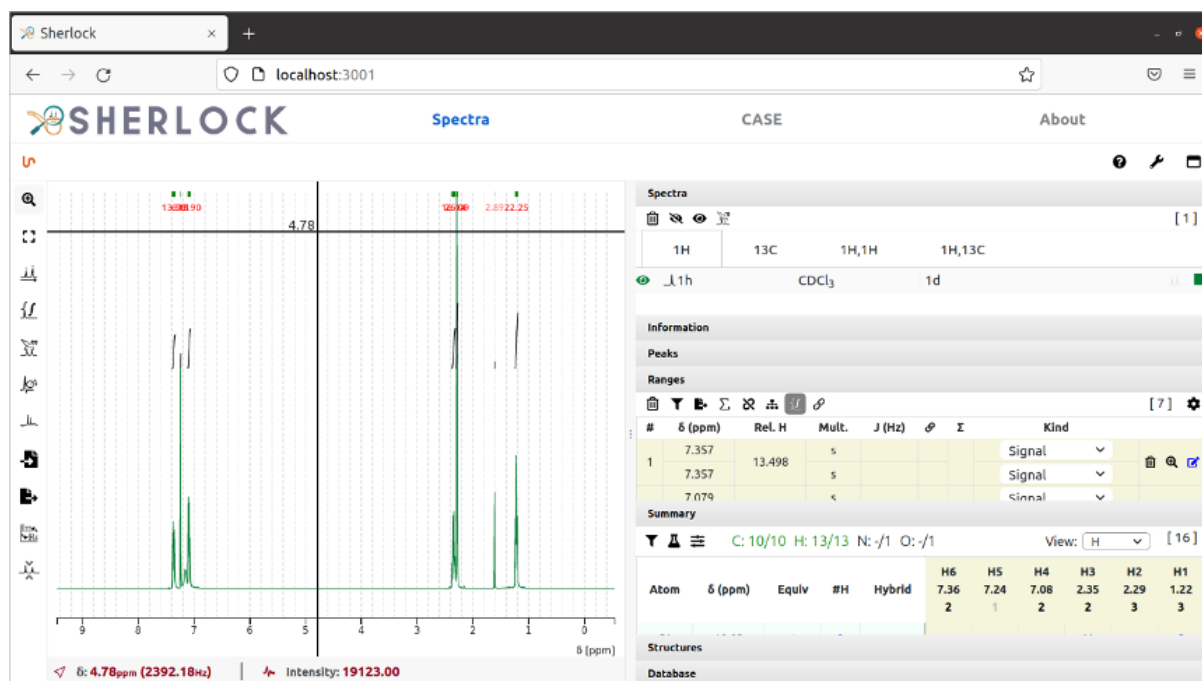


Figure 2 The tab *Spectra* with a NMR dataset and the open panels *Spectra*, *Ranges* and *Summary*

In the tab *CASE* on the left there is an overview of the heavy atoms with their shift value, equivalence, and multiplicity (**Fig 3**).

A click on the arrow above the table (Fig 3) enables an extension of the table with the signals of protons by default. The table view can be changed, e.g. to see proton-proton correlations via COSY.

On the right part choosing between the tabs *Dereplication*, *Elucidation* and *Retrieval* is possible. In the two first tabs the parameters can be set, and the processes can be started. In the tab *Retrieval* the results from previous elucidations are listed.

In the following the different functionalities will be described in more detail.

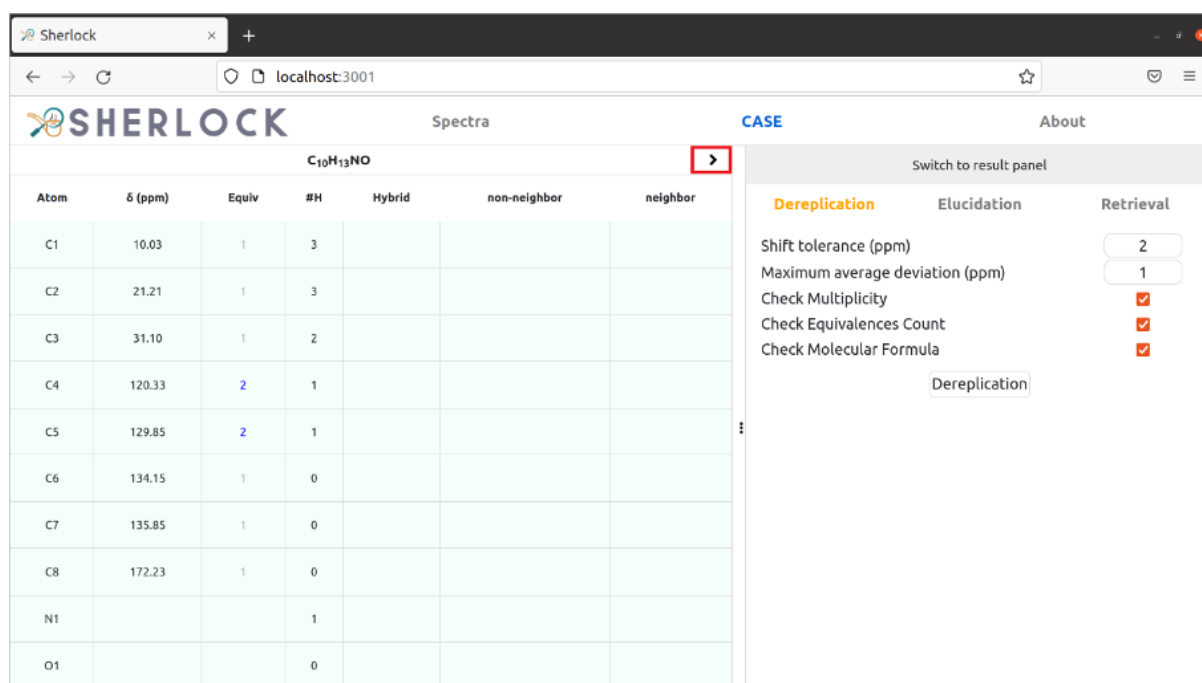


Figure 3 An overview of the tab CASE

## 2.1. Processing the NMR spectra

In the tab *Spectra* the processing of the NMR data will be done. A detailed instruction is found under the question mark button on the right side above the panels or under this link <https://docs.nmrium.org/>.

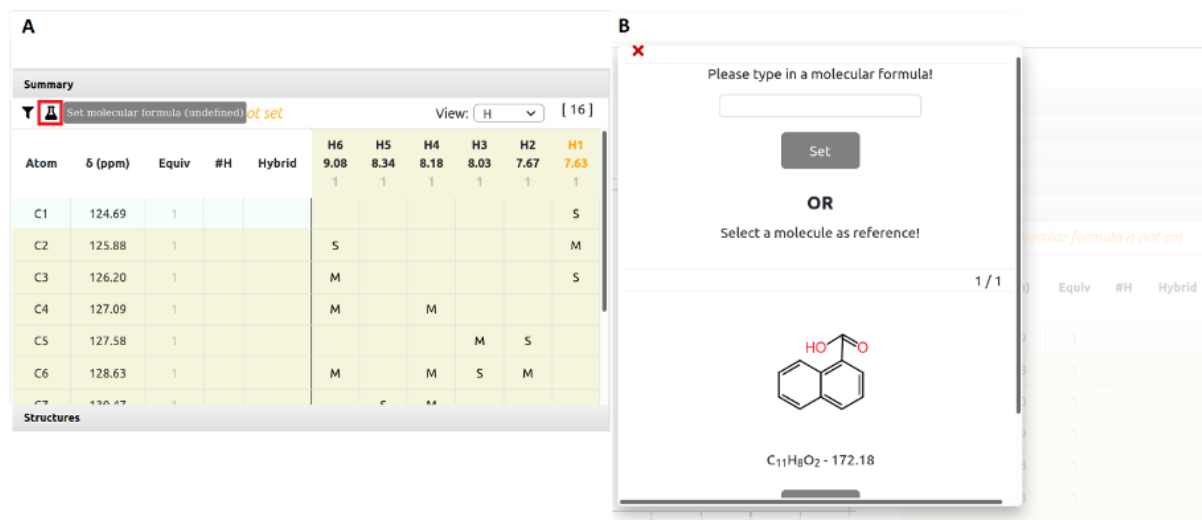


Figure 4 Open panel *Summary* and the open 'Set molecular formula' window

At first the molecular formula should be set in the summary table. Through the flask button above the table the molecular formula can be entered in the extra window, or it can be set via an expected structure (**Fig 4A&B**). Now the actual number of the atoms and the found number of atoms through the spectra should be shown on top of the summary table (**Fig 5**).

Summary

C: 6/9 H: 8/11 N: -/1 O: -/1

View: 

H

 [ 17 ]

Atom	$\delta$ (ppm)	Equiv	#H	Hybrid	H6 7.55 1	H5 7.34 1	H4 7.29 1	H3 7.12 1	H2 2.42 2	H1 1.27 3
C3	119.83	1			S/M	M		M		
C4	124.28	1			M			S		

Structures

Figure 5 Part of the panel *Summary*

The CASE system expects ranges and zones which contain the necessary signal information. Thus the tools 'Ranges Picking' in 1D and 'Zone Tool' in 2D have to be used. The preparation of the NMR data is done in the panel *Summary*. In the next section this is described in detail.

Figure 6 illustrates the procedure to set a new ranges sum for the list of ranges in the 1D proton NMR spectra. Panel A shows the open panels *Spectra* and *Ranges*. The *Ranges* panel shows a list of ranges with a red box highlighting the 'Change Ranges Sum' button. Panel B shows the 'Set new Ranges Sum' dialog box, which allows the user to enter a new value or select a molecule as a reference. The dialog box shows the chemical structure of C<sub>10</sub>H<sub>12</sub>NO and the current sum of 131. Panel C shows the open panels *Spectra* and *Ranges* with the updated ranges sum of 131.000.

Figure 6 The procedure to set a new ranges sum for the list of ranges in the 1D proton NMR spectra  
**A** The open panels *Spectra* with the selected 1H NMR spectra and *Ranges* with the red marked sum button  
**B** The open window to set a new ranges sum  
**C** The open panels *Spectra* and *Ranges* with the changed ranges sum

To support the summarization of the NMR data and the interpretation of the spectra the range sum of a 1D proton experiment (or other 1D experiments) can be changed to an already known one derived by molecular formula. This causes a re-calculation of the relative integration values based on the new set number.

The sum value can be adjusted in the panel *Ranges* of the 1D proton NMR spectra. On the top of the list of picked ranges is the sum sign which leads to the window to set a new range sum (Fig 6A-C).

## 2.2. Summary of the NMR data

After picking the ranges and zones in 1D and 2D NMR spectra an overview of all signals with their shift value, multiplicity, equivalence as well as their correlation to each other is seen in the panel *Summary*.

To match the conditions for the dereplication or elucidation the numbers of the needed carbon atoms (molecular formula) and the picked ones in the NMR data should be the same and colored green. That is necessary because the shift value of each carbon atom is needed for both procedures to work properly .

The chemical shift value and the multiplicity can be used together with the molecular formula to predict the possible hybridization states as well as set or prohibited neighbor atoms for each carbon atom. That structural information will be part of the input for the underlying structure generator.

Such prediction is based on a statistical analysis of the two NMR databases used in Sherlock to determine the hybridization, non-neighbor and neighbor atoms based on the shift value, multiplicity of a carbon atom and the molecular formula. For the spectra-spectra comparison (dereplication) the multiplicity and the equivalence are optional criteria for a specified search. That provides a certain flexibility during the spectral matching process.

If the number of carbon atoms is colored red in the summary table, then the number of picked signals, including equivalences, does not correspond to the actual number.

The numbers of the protons can be colored blue which means some proton signals are not correlated to one of the heavy atoms. In this case the causes could be that the multiplicity or the equivalence of the heavy atoms are not yet set correctly or a proton couldn't be assigned to a heavy atom yet.

Another possible color can be orange which has the meaning that one type of proton signal is correlated to two different heavy atoms at the same time. If two types of proton signal are in one picked range, then it is necessary to pick two ranges and split the signal afterwards. Thereby the program recognizes both signals individually and correlates one each to a heavy atom.

The number of listed proton signals in the table does not have to match the actual number from the molecular formula. This could lead to a higher variation of candidate structures during the elucidation which might cause a longer processing time.

To support the visualization of the picked signals to which structure element it belongs, the cell of the signal as well as the structure elements will be highlighted in orange if the user hovers over one of it (**Fig 11**).

**A**

Summary

View: [ H ] [ 16 ]

Atom	$\delta$ (ppm)	Equiv	#H	Hybrid	H6 7.36 2	H5 7.24 1	H4 7.08 2	H3 2.35 2	H2 2.29 3	H1 1.22 3
C4	120.33	2	1		S	M				
C5	129.85	2	1				S		M	
C6	134.15	1	0		M				M	
C7	135.85	1	0				M			
C8	172.23	1	0					M		M
N1										
O1		0								

add pseudo HSQC

Structures

Database

**B**

Summary

View: [ H ] [ 16 ]

Atom	$\delta$ (ppm)	Equiv	#H	Hybrid	H6 7.36 2	H5 7.24 1	H4 7.08 2	H3 2.35 2	H2 2.29 3	H1 1.22 3
C4	120.33	2	1		S	M				
C5	129.85	2	1				S		M	
C6	134.15	1	0		M				M	
C7	135.85	1	0				M			
C8	172.23	1	0					M		M
N1			1			S				
O1		0								

Structures

Database

**Figure 7** The panel *Summary* with the 'add pseudo HSQC' display and the added red marked signal in the summary table

If a correlation between a proton and another heavy atom than carbon is suspected by a not yet assigned signal, e.g. in the proton NMR spectrum, a pseudo HSQC signal can be added

and also removed by a right mouse click in the according cell in the summary table (**Fig 7A&B**).

An equivalence value has to be set manually if assumed that it is higher than the default value of one. The multiplicity can be entered or changed manually by a left mouse click. If there are picked signals within DEPT or multiplicity-edited HSQC data available, Sherlock pre-sets the multiplicity value for carbons automatically. Multiple values are possible in case of a positive signal in HSQC.

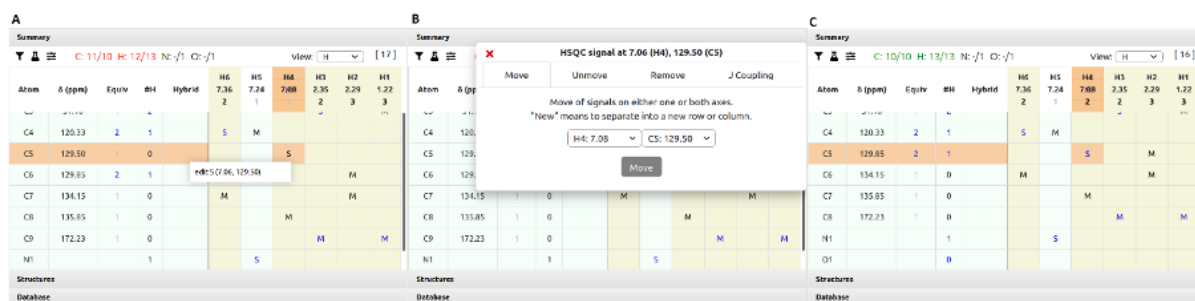


Figure 8 The procedure to move a signal to another carbon atom

Further possibility to assign a picked signal to another heavy atom or proton group in the summary table is to move the signal. The user only has to do a right click into the field of the affected signal (**Fig 8A**). In the window that appears the signal can be moved to another signal group, unremoved or removed (**Fig 8B**). If it is assigned to another signal it will move to the chosen atom signal in the table and change its color to blue to indicate the movement (**Fig 8C**).

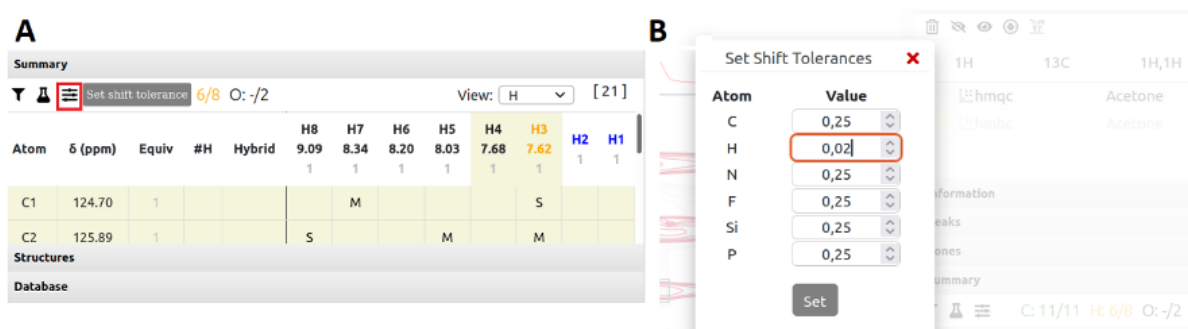


Figure 9 The panel *Summary* with the marked 'set shift tolerance' button and the open window

During the peak picking signals having similar chemical shifts within a certain search window will be grouped together. Thus the grouping method depends on a given tolerance value which spans such a window. If a preset tolerance value for a certain atom type is too large then and wrong signals fall into one group, the shift tolerance can be changed through the right button above the table (**Fig 9A**). In the open window the shift tolerance can be set for each individual atom in the list (**Fig 9B**). By this the user has the ability to control the signal



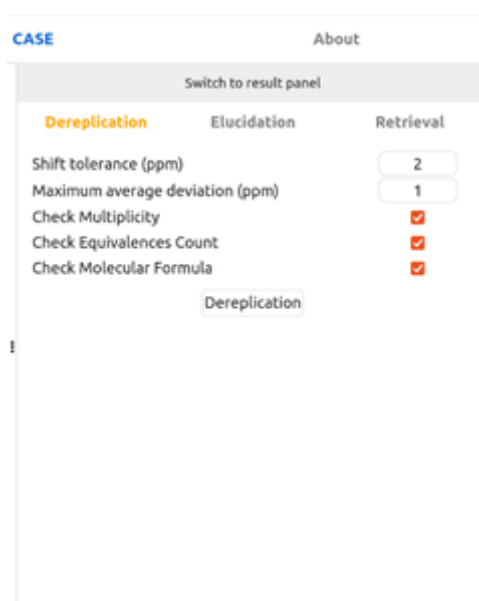
grouping in advance. This action will have effect on further picked signals only. Already built groups will stay untouched.

## 2.3. The dereplication

Before generating structures based on the given spectral information a lookup through spectral databases is a common step. Sherlock uses for this purpose the databases NMRShiftDB and COCONUT and does spectrum-spectrum comparisons between a  $^{13}\text{C}$  query spectrum formed by the correlation data and spectra in the spectral knowledge base.

The default setting for the dereplication is adjusted in Sherlock. The parameters are chemical shift tolerance and maximum average deviation. Both are given in ppm (parts per million) (**Fig 10**).

The chemical shift tolerance means the allowed deviation from every picked to the associated signals in the spectra of the databases. The default value is not allowed to be higher than 2 ppm.



The screenshot shows the 'CASE' tab in the Sherlock application. At the top, there are tabs for 'CASE' and 'About'. Below them is a 'Switch to result panel' button. The main area is divided into three sections: 'Dereplication' (highlighted in orange), 'Elucidation', and 'Retrieval'. Under 'Dereplication', there are two input fields: 'Shift tolerance (ppm)' with a value of 2, and 'Maximum average deviation (ppm)' with a value of 1. Below these are three checkboxes: 'Check Multiplicity' (checked), 'Check Equivalences Count' (checked), and 'Check Molecular Formula' (checked). At the bottom of the 'Dereplication' section is a button labeled 'Dereplication'.

**Figure 10** The tab CASE with the open panel Dereplication with the default setting

The second parameter is the maximum average deviation which stands for the mean value of differences of all matching signal-signal pairs between two spectra. The mean should be less than 1 ppm by default.

In Sherlock the user can decide whether the multiplicity, equivalences count and molecular formula should be checked during the dereplication process. This is optional to allow more flexibility while searching through the database entries.

The results of the dereplication are then shown in the right panel too (**Fig 11**). Above the found compounds it shows the number of results, the time which was needed for the spectra-spectra-comparison, a button for downloading the results in SDF with several meta information per molecule, settings for the image size, the number of shown results on one

page and the sorting of the results. The list of results can be sorted by average deviation (avgDev), the similarity coefficient Tanimoto, the measure Root-mean-square-deviation (rmsd) and the number of signal matches (hits) derived from spectrum-spectrum comparison (**Fig 11**). The default setting is average deviation.

In **Figure 11** an example of results from a dereplication is shown. Sherlock found two entries which have the same molecular structure. This can happen because the dereplication is based on two databases and if the compound is deposited in both, Sherlock shows both as individual results. At the bottom of every result the identification number of the compound is found which is also a link to the entry in the database.

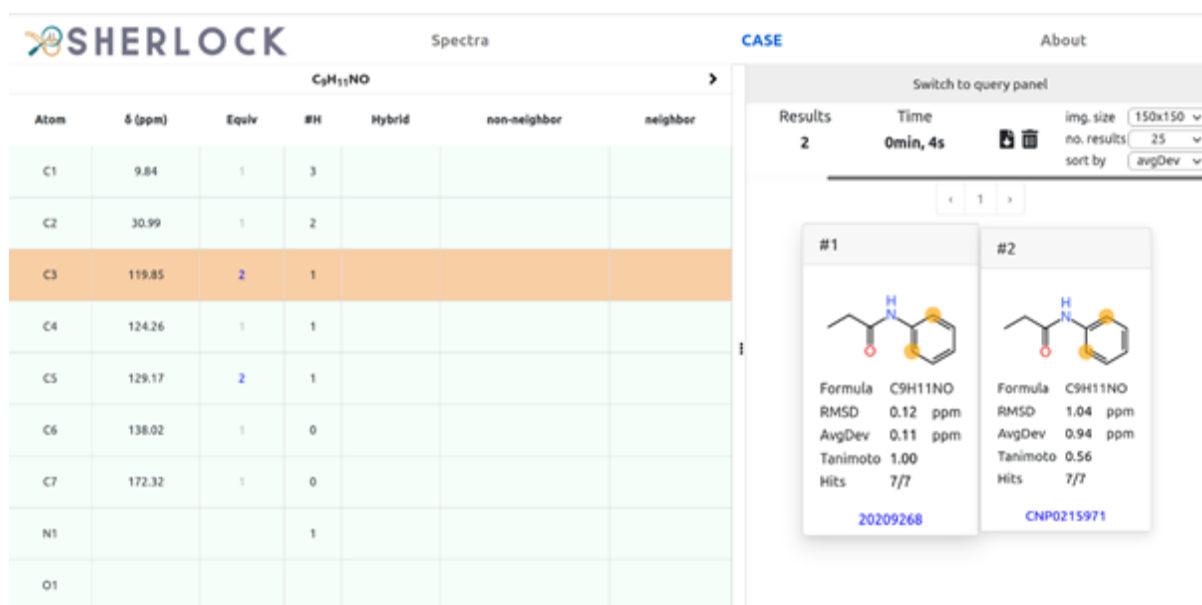


Figure 11 The results of a dereplication search in the tab CASE

Not only the identification number is displayed in every result entry but also the molecular structure, the molecular formula, rmsd, the average deviation, Tanimoto coefficient and the number of hits. In the molecular structure it can be shown which heavy atom corresponds to a signal from the NMR data which are shown in the left table. The other way around is also possible.

On top of the right panel the button 'Switch to query panel' is responsible to go back to the panels *Dereplication*, *Elucidation* and *Retrieval*.

## 2.4. Structure Elucidation

If the molecular structure is still unknown after the dereplication, the elucidation can be applied to generate possible molecular structures based on the picked signals.

At first Sherlock determines statistically possible hybridization, neighbor atoms and non-neighbor atoms for each carbon atom. Therefore, it compares the data with the knowledge gained from the databases. The lookup of the possible hybridization is based on

the molecular formula, the shift value, and the hybridization of the picked signal. The neighbor and non-neighbor detection includes the possible hybridization, molecular formula, the shift value and the multiplicity of the signal for the computation. Thus, the neighborhood detection can only be switched on if the hybridization detection has been activated too, because it depends on the assumed possible hybridization. Under the 'Connectivity Statistics Detection' part of the settings both detections can be adjusted or switched off. For the hybridization the 'Minimal occurrence of hybridization in DB' which means the hybridization has to occur with the minimal percentage at this shift value in the databases can be set. The default setting is one percentage (**Fig 12A**).

For the non-neighbor detection, the parameter 'Lower limit for non-neighbors detection' can be set. It defines the percentage of what minimal occurrence a bond between a carbon atom with its gained properties from the NMR data and another heavy atom type in the pre-built statistics should have. It is preset to one percentage (**Fig 12A**).

The 'Lower limit for set neighbors detection' parameter means the ratio of atomic bonds between a carbon atom and another heavy atom has to turn up with a minimal percentage of 95 %. It can also be changed manually (**Fig 12A**).

By double-clicking on a field of the columns *Hybrid*, *non-neighbor* and *neighbor* they can be changed individually (**Fig 12B**).

**Panel A: CASE - Elucidation Settings**

Switch to result panel: Dereplication | **Elucidation** | Retrieval

**Connectivity Statistics Detection:**

- Use hybridization detection: ☒
- Minimal occurrence of hybridization in DB (%): 1
- Use neighbor detection: ☒
- Lower limit for non-neighbors detection (%): 1
- Lower limit for set neighbors detection (%): 95

**Elimination of correlations:**

- Allow: ☐
- Number of eliminations: 1
- Maximal path length: 4

**Structure Generation:**

- Total time limit (min): 5
- Shift tolerance (ppm): 30
- Maximum average deviation (ppm): 5
- Allow combinatorics: ☒

**Further settings:**

- Allow hetero-hetero bonds: ☐
- Filter out 3-membered rings: ☐
- Filter out 4-membered rings: ☐

**Task name:**

**Panel B: SHERLOCK - Spectra**

Molecular formula: C<sub>6</sub>H<sub>11</sub>NO

Atom	δ (ppm)	Equiv	#H	Hybrid	non-neighbor	neighbor
C1	9.84	1	3	sp3	N*, O*	C*
C2	30.99	1	2	sp3	N*, O*	C*
C3	119.85	2	1	sp2	O*	C*
C4	124.26	1	1	sp2	O*	C*
C5	129.17	2	1	sp2	O*	C*
C6	138.02	1	0	sp2		C*
C7	172.32	1	0	sp2		C*, O*
N1			1			
O1						

**Panel C: CASE - Elucidation Settings (continued)**

Switch to result panel: Dereplication | **Elucidation** | Retrieval

**Elimination of correlations:**

- Allow: ☐
- Number of eliminations: 1
- Maximal path length: 4

**Structure Generation:**

- Total time limit (min): 5
- Shift tolerance (ppm): 30
- Maximum average deviation (ppm): 5
- Allow combinatorics: ☒

**Further settings:**

- Allow hetero-hetero bonds: ☐
- Filter out 3-membered rings: ☐
- Filter out 4-membered rings: ☐

**Task name:**

**Elucidation**

**Figure 12** The tab **CASE** with the open panel **Elucidation** and the results of the neighbor atom detection

**A** The setting for the neighbor detection 'Connectivity statistics Detection' **B** The results of the neighbor detection and the settings for the elucidation

Further settings for the elucidation are the 'Elimination of correlations', 'Structure Generation' and 'Spectra prediction'. With a tick the elimination of correlations can be allowed and the number of eliminations as well as the maximal path length can be set. The default setting is one elimination and a maximal path length of four. For the 'Structure Generation' the total time limit in minutes can be adjusted, combinatorics and hetero-hetero-bonds can be allowed, and 3-membered and 4-membered rings can be filtered out. Under the 'Spectra prediction' the shift tolerance and the maximum average deviation can be set. Both are given in ppm (**Fig 12B**) and used during spectral matching between the <sup>13</sup>C query spectrum and predicted <sup>13</sup>C spectrum of one candidate structure, similar to what is done in the dereplication process.

For a better overview and refinding in the list of all result records the user has the opportunity to enter a task name.

After the setting and the detection of different structural constraints the elucidation can be started.

The results of the elucidation are displayed like the ones of the dereplication. If the number of hits does not match, the result is colored red.

Additionally, a table containing the chemical shift predictions and assignments to atoms in each structure is available, if expanded via pressing the according button (Fig 13).

Figure 13 displays two panels (A and B) showing the results of an elucidation query in the 'CASE' tab. Both panels show a list of results with columns: Name, Results, Time, img. size, no. results, and sort by. The results are sorted by hits.

Panel A shows a single result entry with the following details:

- Name: 624605c948deee7b51376d1f
- Results: 5
- Time: 0min, 1s
- img. size: 200x200
- no. results: 25
- sort by: hits

The details section shows:

- Formula: C9H11NO
- AvgDev: 0.23 ppm
- RMSD: 0.32 ppm
- Tanimoto: 0.75
- Hits: 7/7

Panel B shows the same result entry, but with the 'Open prediction table' expanded, displaying a table of chemical shift predictions and assignments to atoms in each structure.

Shift	Dev	Sphere	Count	Range
9.71	0.14	6	4	0.57
30.25	0.74	6	3	1.56
119.80	0.05	6	2	0.00
119.80	0.05	6	2	0.00
124.35	0.09	6	404	3.60

Figure 13 A list of results from an elucidation and one result with the open prediction table in the tab CASE

## 2.5. The retrieval

In the panel *Retrieval* the results of the elucidation queries are listed with name, date, count, preview and actions. The possible actions are loading the results including correlation data again or deleting the entry. Above the list searching through the list by name is possible (Fig 14).

The screenshot shows the SHERLOCK web application interface. The top navigation bar includes 'SHERLOCK', 'Spectra', 'CASE', and 'About'. The 'CASE' tab is active, and the 'Retrieval' panel is selected. The left panel displays a table for the molecular formula  $C_9H_{11}NO$  with columns: Atom,  $\delta$  (ppm), Equiv, #H, Hybrid, non-neighbor, and neighbor. The right panel shows a search bar and a table of results. One result is visible with ID '22a73fa91538b1aa83f6925', date '2022-2-4 22:56', count '5', and a chemical structure of N-ethylbenzamide.

Atom	$\delta$ (ppm)	Equiv	#H	Hybrid	non-neighbor	neighbor
C1	9.84	1	3	sp3	N*, O*	C*
C2	30.99	1	2	sp3	N*, O*	C*
C3	119.85	2	1	sp2	O*	C*
C4	124.26	1	1	sp2	O*	C*
C5	129.17	2	1	sp2	O*	C*
C6	138.02	1	0	sp2		C*
C7	172.32	1	0	sp2		C*, O*
N1			1			
O1						

Figure 14 The panel *Retrieval* in the tab *CASE*

### 3. Troubleshooting

Below is the list of the most commonly-encountered issues, with some suggested causes and solutions.

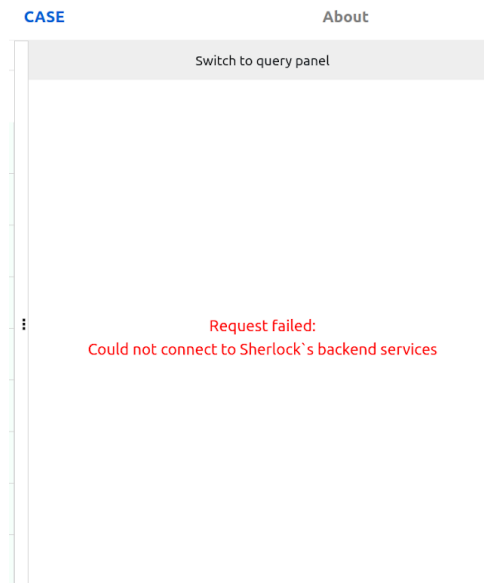
If an error message like in **Figure 15** pops up after requesting to the backend, there are two possible causes: the molecular formula is not set or the multiplicity of a carbon atom is not clearly set, which means exactly one value.

The screenshot shows the SHERLOCK web application interface. The 'CASE' tab is active, and the 'Switch to query panel' button is visible. The left panel displays a table for the molecular formula  $C_9H_{11}NO$  with columns: Atom,  $\delta$  (ppm), Equiv, #H, Hybrid, non-neighbor, and neighbor. The right panel shows an error message: 'Request failed: At least for one carbon the number of attached protons is missing!!!'. The error message is highlighted in red.

Atom	$\delta$ (ppm)	Equiv	#H	Hybrid	non-neighbor	neighbor
C1	9.84	1	1,3			
C2	30.93	1	2			
C3	77.34	1	1,3			
C4	119.82	1	1,3			
C5	124.22	1	1,3			
C6	129.15	1	1,3			
C7	138.01	1	0			
C8	172.34	1	0			
C9						
N1						
O1						

Figure 15 An error message after a request for a dereplication or an elucidation in the tab *CASE*

Another possible issue is a failed connection to the backend service (**Fig 16**).



**Figure 16** The query panel in the tab CASE

One reason might be that after starting the backend service it needs a little time to boot completely. It should work properly when retrying after one or two minutes.

In case one or multiple backend services do not work as expected, a restart of those or all services could be done followed by an examination if every service works accordingly.

If the problem can't be figured out, you can get in touch with us via creating an issue on GitHub under <https://github.com/michaelwenk/sherlock> or directly via email under [michael.wenk@uni-jena.de](mailto:michael.wenk@uni-jena.de). A detailed description of the problem and what steps have been done to reproduce the error should be attached.