An Expert System for the Automatic Assignment of 1H NMR Spectra of Small Molecules

Carlos Cobas[1], Michael Bernstein[1], Esther Vaz[1], Felipe Seoane[1], Maruxa Sordo[1], Santiago Domínguez[1], Manuel Pérez[1], Stanislav Sýkora[2]
[1]Mestrelab Research, S.L Feliciano Barrera 9B – Bajo, 15706 Santiago de Compostela, Spain; corresponding author carlos@mestrelab.com
[2]Extra Byte, Castano Primo, Italy

Introduction

We present an integrated new software solution aimed at automatic assignment of 1H NMR spectra of small molecules. It constitutes an expert system using the principles of fuzzy logic and probabilistic methods which first classifies all the resonances (peaks) in the spectrum and then proceeds to enumerate the most likely assignments of experimental multiplets to a presumed molecular formula and score on them. It uses **as inputs** the **experimental spectrum** (or possibly various kinds of spectra spectra), the **suggested molecular formula**, and the **predicted NMR parameters** (shifts and coupling constants) and, **as output**, it generates the **most likely assignment**.

The functionality is now available in MestReNova software (www.mestrelab.com)

The Auto Assignment Algorithm

The Auto Assignment Algorithm combines several software techniques we had developed in recent years as tools for expert tasks such as automatic detection and characterization of spectral peaks, automatic solvent detection, and automatic structure verification (or which the auto-assignment feature is, in its own term, a building block.

Real-life spectra always contain a number of artifacts such as noise, baseline distortions, relaxation induced and radiation-damping induced distortions of peak intensities, lineshape distortions due to magnetic field inhomogeneity, lineshape distortions due to unresolved weak long-range couplings, second-order interactions, peaks crowding causing peaks and multiplets to overlap, etc …

For these reasons it is impossible to construct any NMR-data evaluation wizard without an extensive usage of statistical methods, allowing for a degree of logical "fuzziness". In our case this is done by applying at every step, to the full depth of the algorithm, a proprietary scoring system approach.

The Auto Assignment global flowchart, shown in Fig.1, consists of the following constituent blocks:

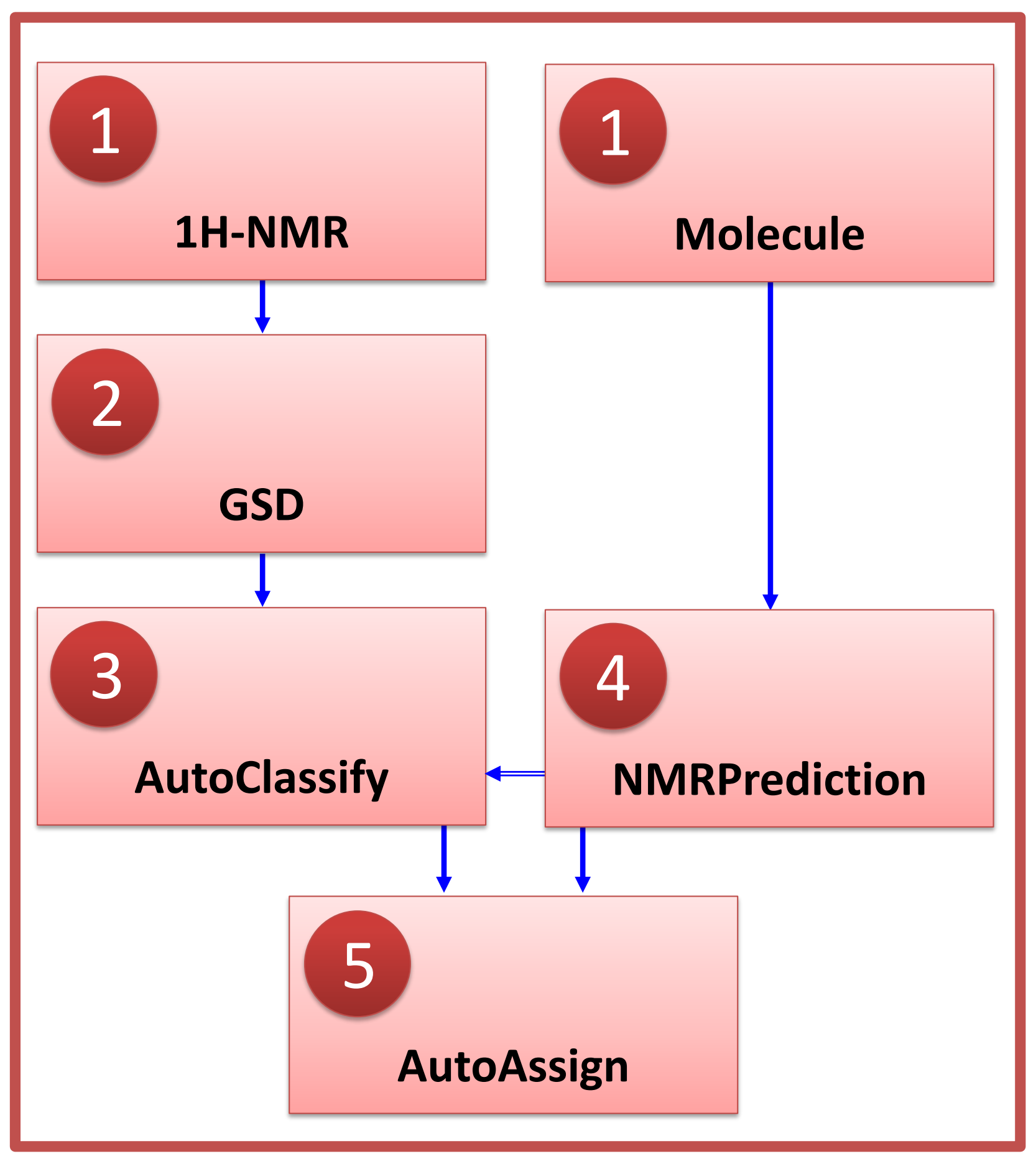
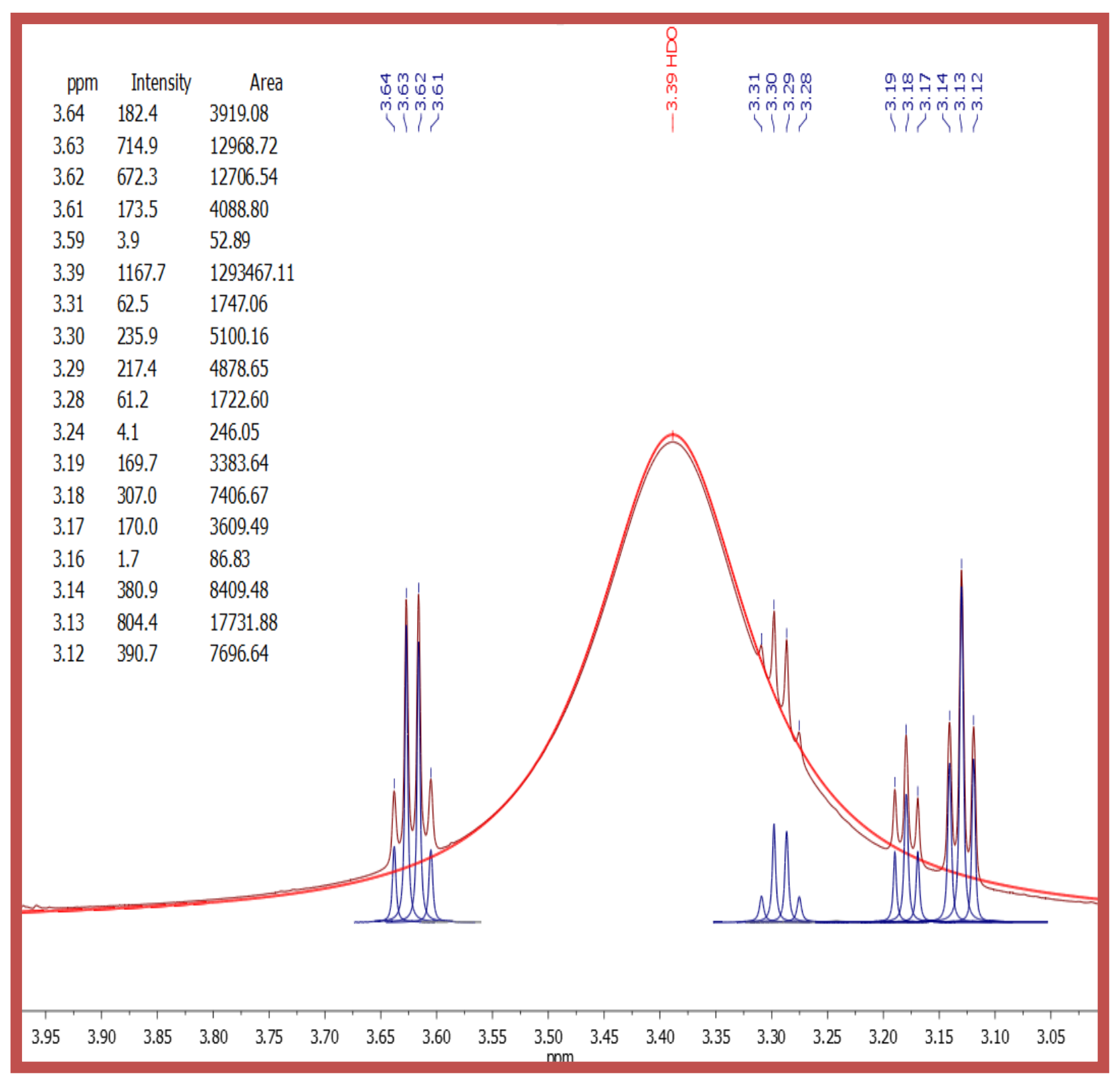


Fig.1 *Basic flowchart diagram of the new 1H-NMR Automatic Assignments algorithm.*
See the text for a description of its constituent blocks.

1 **Basic processing:** An NMR-FID is loaded, apodized, transformed, phased and baseline corrected, typically in a transparent, fully unattended way (the process, howwever, can be predefined by the software user).

In addition, a suggested molecular formula is loaded, using one of the popular formula-encoding formats.

2 **GSD:** The resulting frequency domain 1H spectrum is automatically deconvolved using the sophisticated **Global Spectrum Deconvolution** algorithm [1] in order to generate a reliable list of peaks and their parameters (position, height, width, kurtosis, area, etc), even in situations characterized by a strong peaks overlap (Fig.2).



ppm	Intensity	Area
3.64	182.4	3919.08
3.63	714.9	12968.72
3.62	672.3	12706.54
3.61	173.5	4088.80
3.59	3.9	52.89
3.39	1167.7	1293467.11
3.31	62.5	1747.06
3.30	235.9	5100.16
3.29	217.4	4878.65
3.28	61.2	1722.60
3.24	4.1	246.05
3.19	169.7	3383.64
3.18	307.0	7406.67
3.17	170.0	3609.49
3.16	1.7	86.83
3.14	380.9	8409.48
3.13	804.4	17731.88
3.12	390.7	7696.64

Fig.2 *Example of information about the spectral peaks extracted by GSD in the presence of a strong overlap*

3 **AutoClassify:** Using another sophisticated fuzzy-logic algorithm [2], each peak in the GSD list is classified according to whether it belongs to the compound or to the solvent, or whether it an impurity, an artefact, a ^{13}C satellite, etc (Fig.3). The algorithm even attempts to pinpoint possible labile peaks.

An important part of this process is also the recognition and of multiplets due to J-couplings and a detailed characterization of their many properties (this results in a multiplets list). Inter-multiplet coupling patterns are also detected and stored internally (the so-called *Edited JC Splittings List*) as another tool for the subsequent auto-assignment step.

While part of this process can be done without any knowledge of the molecule. This branch of the algorithm points towards automatic spectrum elucidation – a logical future project. In the case discussed here, however, the molecule is presumably known and part of the information about it is used in the AutoClassification process (indicated in the diagram by the arrow from block 4 to block 3).

All this, moreover, is done in an iterative way which confers to the algorithm a capability to 'backtrack' and repeat earlier steps with a newly gained or corrected information.

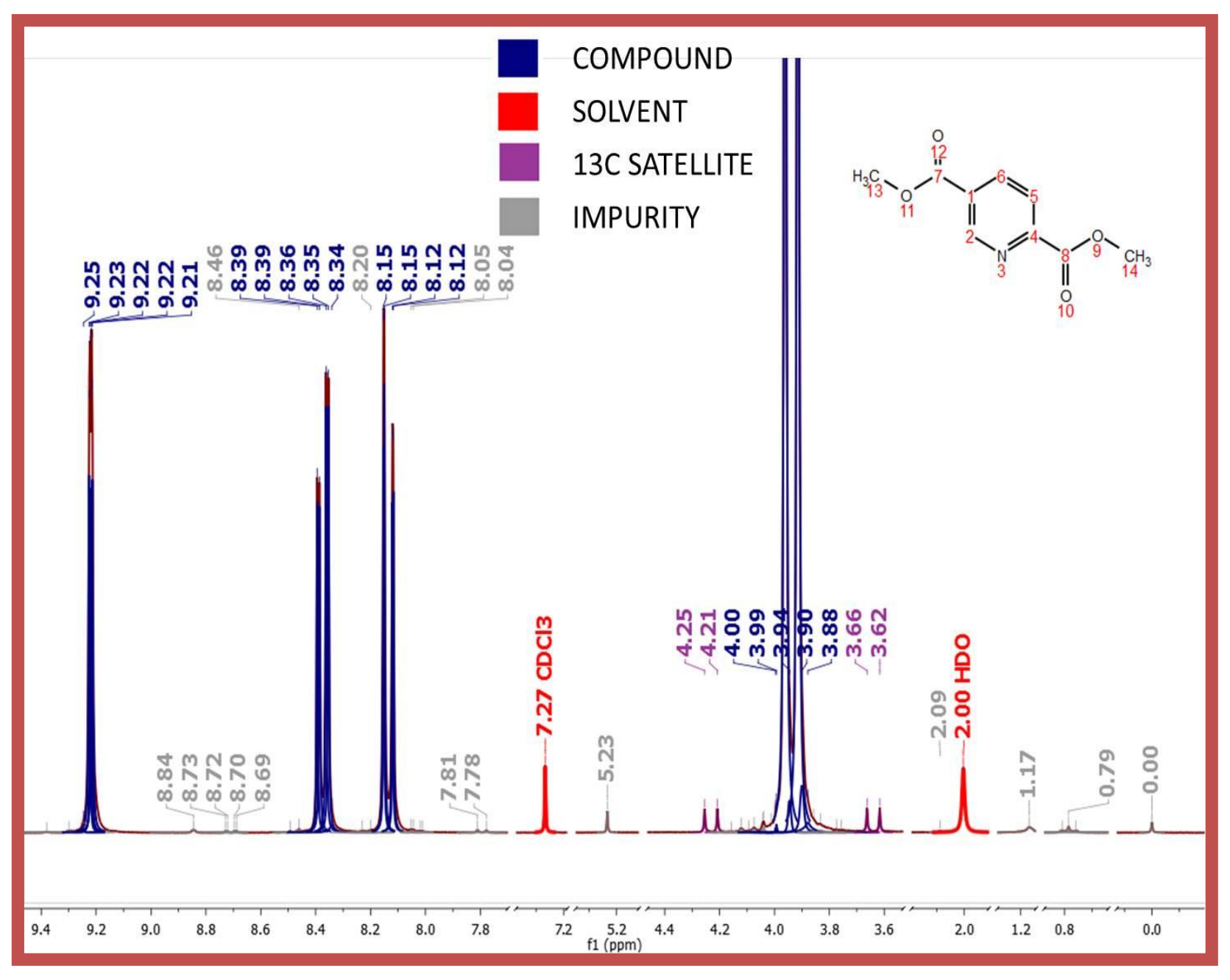


Fig. 3. *Illustration of the AutoClassiffy algorithm.*
Peaks are color coded according to their their type

… continued in right column

… continued from left column

4 **NMRPredict:** NMR spectral parameters (chemical shifts and scalar coupling constants) of the suggested molecule are predicted using three complementary approaches (3D conformer, substituent chemical shift and HOSE code database) which are then combined by means of the NMRPredict Best Algorithm [3]. Users can also add their own assignments to the HOSE code database to further refine the accuracy of the predictions.

5 **AutoAssign:** The final step of the algorithm consists in combining all the information collected so far. Basically, the wizard tries to find the best possible match between the experimental multiplets and the predicted multiplets, subject also to constraints dictated by NMR know-how. Mathematically, the number of possible assignments is staggering, but applying a prior enumeration filter passing only a limited number (about 100) of the most likely ones. In this way it becomes feasible to score each assignment against all available information and come up with the best one.

Since a poster is obviously totally inadequate for the purpose, the details of the algorithm will be presented elsewhere. However, we hope that the present description provides a sufficiently clear picture of its underlying concepts and its most important features.

Tests with artificial spectra

To validate the performance of the algorithm (as a proof of concept), the system was first tested with the five compounds of Fig. 4 whose NMR spectra were 'synthesized' using the Mnova intrinsic spectrum simulation facility and the 'theoretical' NMR spectral parameters generated by NMRPredict. The goal was to isolate common practical NMR issues such as peaks overlap, insufficient resolution, etc. Additionally, for each structure, new spectra where calculated in which the chemical shifts were randomly shifted using a normal distribution with SD = 0.25.

It was found that of the 69 potential assignments, all were correctly determined by the algorithm except for one labile proton in Quinine.

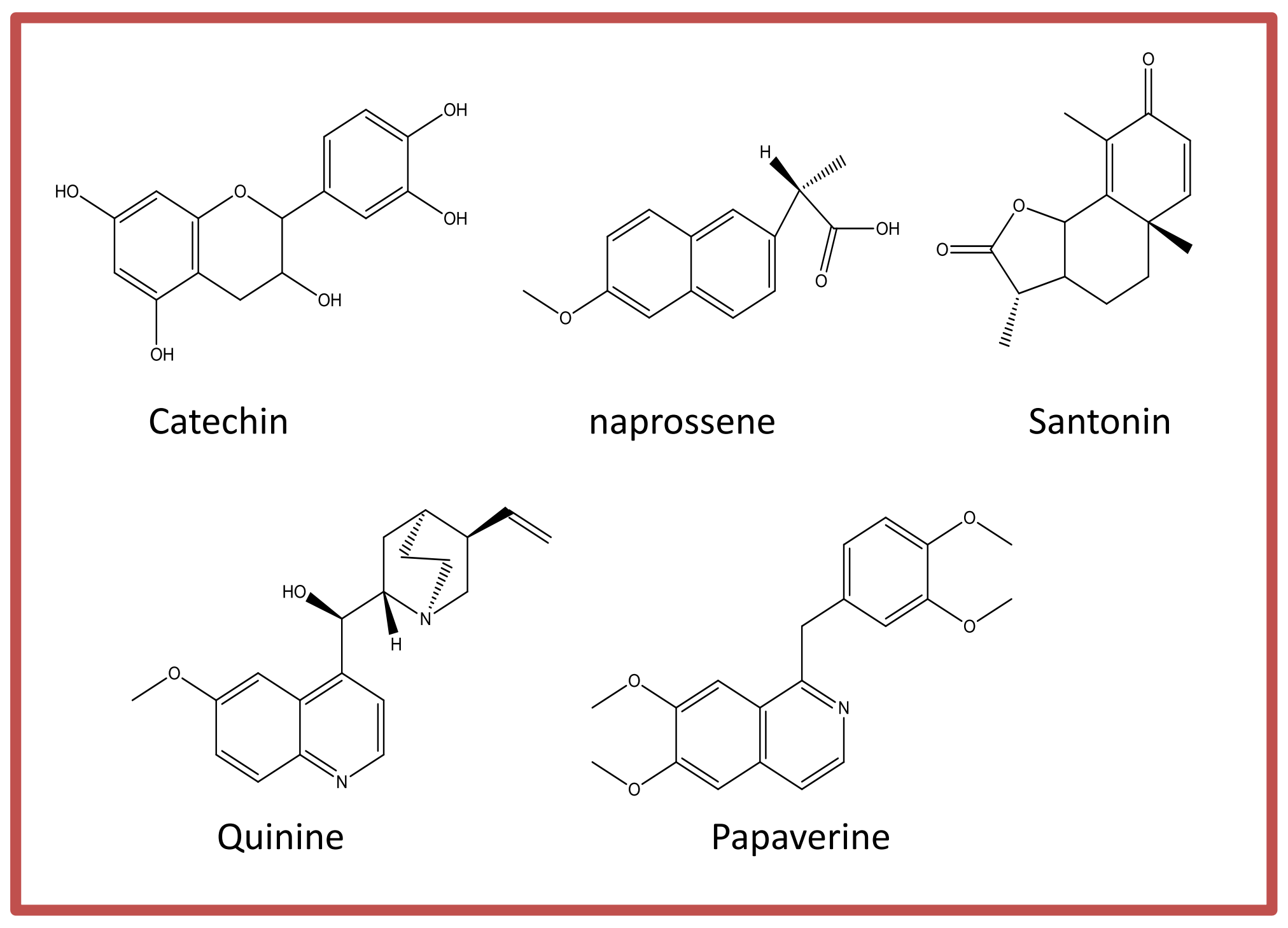


Fig. 4 *Compounds used to synthesize the spectra used in the validation test.*

Tests with experimental spectra

Further tests were conducted with a fully assigned 1H-NMR library consisting of **39** molecules and a total of **355** proton assignments.

In a first test, the performance of the algorithm was evaluated without including into NMRPredict HOSE code DB any of the assigned molecules used in this study. In a second test, some of the molecules where added to the DB in order to get more accurate chemical shift predictions, as well as smaller error bounds. The results obtained when the Auto-Assignment wizard was executed in a fully automatic mode are summarized in the following table:

User DB	Correct	Wrong	%
No	280	75	**73**
Yes	295	60	**80**

Conclusions

The performance of the auto-assignment wizard is quite impressive and usable, especially considering that while the its structure is now well set, its potential is still in development (we might say that it is still learning). Closer inspection to the results showed that any incorrect results are often due to the presence of several assignments with similar probabilities, a situation which will need to be handled.

In this work, only 1D 1H NMR spectra were used but the system is already armed to accept HSQC spectra. Results obtained with a combined 1H & HSQC approach will be covered in a future publication.

References:

[1] Carlos Cobas, Stanislav Sykora, *The Bumpy Road towards Automatic Global Spectral Deconvolution (GSD)*, 50th ENC Conference, Asilomar, CA (USA), March 29-April 4, 2009
[2] Spectroscopy Europe, 23(1), 25-30 (2011)
[3] Spectroscopy Europe, 20(1), 21-23 (2008)

MESTRELAB RESEARCH
Chemistry Software Solutions

Available also via
DOI 10.3247/sl4nmr12.001

Download all Mestelab Posters
with this QR Code or visit the page
www.mestrelab.com/publications